Genes Genet. Syst. (1999) **74**, p. 117–127

Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method

Yi-Ju Li, Yoko Satta, and Naoyuki Takahata*

Department of Biosystems Science, The Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan

(Received 7 June 1999, accepted 3 September 1999)

The species divergence times and demographic histories of *Drosophila melanogaster* and its three sibling species, *D. mauritiana*, *D. simulans*, and *D. yakuba*, were investigated using a maximum likelihood (ML) method. Thirty-nine orthologous loci for these four species were retrieved from DDBJ/EMBL/GenBank database. Both autosomal and X-linked loci were used in this study. A significant degree of rate heterogeneity across loci was observed for each pair of species. Most loci have the GC content greater than 50% at the third codon position. The codon usage bias in *Drosophila* loci is considered to result in the high GC content and the heterogenous rates across loci. The chi-square, G, and Fisher's exact tests indicated that data sets with 11, 23, and 9 pairs of DNA sequences for the comparison of *D. melanogaster* with *D. mauritiana*, *D. simulans*, and *D. yakuba*, respectively, retain homogeneous rates across loci. We applied the ML method to these data sets to estimate the DNA sequence divergences before and after speciation of each species pair along with their standard deviations. Using 1.6×10^{-8} as the rate of nucleotide substitutions per silent site per year, our results indicate that the *D. melanogaster* lineage split from *D. yakuba* approximately 5.1 ± 0.8 million years ago (mya), *D. mauritiana* 2.7 ± 0.4 mya, and *D. simulans* 2.3 ± 0.3 mya. It implies that *D. melanogaster* became distinct from *D. mauritiana* and *D. simulans* at approximately the same time and from *D. yakuba* no earlier than 10 mya. The effective ancestral population size of *D. melanogaster* appears to be stable over evolutionary time. Assuming 10 generations per year for *Drosophila*, the effective population size in the ancestral lineage immediately prior to the time of species divergence is approximately 3×10^6, which is close to that estimated for the extant *D. melanogaster* population. The *D. melanogaster* did not encounter any obvious bottleneck during the past 10 million years.

INTRODUCTION

The *melanogaster* species subgroup of *Drosophila* consists of eight members. These eight species differ from one another in male genitalia, ecology, and polymorphism patterns in their populations (Lachaise *et al*., 1988). Although this subgroup has been studied extensively in various respects, the dates of the speciation events and the phylogenetic relationships are not fully established. An early phylogenetic study was based on polytene chromosome banding sequences (Lemeunier and Ashburner, 1976). Since then, various approaches, including biogeographical and geological evidence (Lemeunier *et al*., 1986), allozymes (Cariou, 1987), DNA-DNA hybridization techniques (Caccone *et al*., 1988; Powell *et al*., 1986), and estimations of nucleotide substitution rates at the DNA sequence level (Stephens and Nei, 1985; Moriyama, 1987; Sharp and Li, 1989), have been used to estimate the species divergence time in *Drosophila*. From these studies, two conclusions that are generally accepted for the *D. melanogaster* subgroup have been drawn: (1) *D. melanogaster* diverged from the stem lineage of *D. simulans* and *D. mauritiana*, and (2) the divergence between *D. melanogaster* and *D. simulans* occurred approximately 2 to 3 million years ago (mya). There is still uncertainty about the remaining branches of the *D. melanogaster* subgroup. For instance, the divergence time between *D. melanogaster* and *D. yakuba* has been estimated to be 13 to 17 mya (Beverley and Wilson, 1982; Bodmer and Ashburner, 1984), 5.1 mya (Cariou, 1987), 10

* Corresponding author.

mya (Lachaise *et al*.,1988), 7.2 mya (Sawyer and Hartl, 1992), and 6.1 mya (Russo *et al*., 1995). If we take the divergence time between *D. melanogaster* and *D. simulans* as 2 to 3 mya, the above estimates indicate that the possible divergence time between *D. melanogaster* and *D. yakuba* ranges from two- to eight-fold of that between *D. melanogaster* and *D. simulans*. Given this range, *D. yakuba* could be either a closely- or distantly-related outgroup of the *D. melanogaster* trio. It is, therefore, necessary to re-examine the divergence time between *D. melanogaster* and *D. yakuba.*

Ancestral polymorphism is receiving some attention in phylogenetic studies. Various studies have shown that the topology of molecular phylogenetic trees can differ from locus to locus (Nei, 1987). Gene trees differ from species trees because genes sampled from different species must have diverged prior to the speciation event. One source of discrepancies between gene trees and species trees is ancestral polymorphism. Therefore, to estimate the species divergence time accurately, the DNA sequence divergence accumulated before speciation should be excluded (Takahata *et al*., 1995; Takahata and Satta, 1997). In contrast, the conventional method simply takes the average of DNA sequence divergences across loci for a pair of species to estimate their divergence time. When the ancestral polymorphism is substantial, the species divergence time obtained by the constant conventional method is overestimated. Given the nucleotide substitution rate per site per year and the assumption of neutrality, the effective ancestral population size is the only parameter that affects ancestral polymorphism. A comparison of demographic histories between the ancestral and extant populations may clarify how the polymorphism has changed through the evolutionary history. Thus, the effective ancestral population size is an important parameter for phylogenetic as well as demographic studies.

In *Drosophila* and most other species, the effective population size has been investigated primarily for extant organisms (Fuerst *et al*., 1977; Zouros, 1979; Nei and Graur, 1983; Sawyer and Hartl, 1992; Hamblin and Aquadro, 1996). Few studies have addressed the effective ancestral population size. For instance, Wakely and Hey (1997) determined that the effective population size of the ancestor of *D. simulans* and *D. mauritiana* was intermediate between those of these descendants. For other species pairs even within the *D. melanogaster* subgroup the effective ancestral population sizes have not been investigated.

Here, we applied the maximum likelihood method (ML) of Takahata and colleagues (Takahata *et al*., 1995; Takahata and Satta, 1997) to three pairs of species: *D. melanogaster-D. mauritiana*, *D. melanogaster-D. simulans*, and *D. melanogaster -D. yakuba* for which reasonable amounts of DNA sequence data are available. The ML method separates DNA sequence divergences into two categories: before and after species divergence. The former can be used to estimate the effective ancestral population size and the latter the species divergence time. The silent sites and silent substitutions are used in our analysis, because they are considered to close to neutrality assumed in the ML method. Our purpose here is three-fold: to decipher a part of evolutionary history of the *D. melanogaster* subgroup by estimating the current and historical population parameters, to examine the power of the ML method, and to test the constancy of the silent substitution rate across loci.

MATERIALS AND METHODS

We surveyed DDBJ/EMBL/GenBank database and retrieved DNA sequences for orthologous nuclear genes that are available in any species pair for *D. melanogaster*, *D. mauritiana*, *D. simulans*, and *D. yakuba*. Of these, *Adhr* and *Gpdh* in the category of *D. yakuba* are actually from *D. teissieri* (Table 1). They are included since *D. yakuba* and *D. teissieri* are considered to form a monophyletic group within the *D. melanogaster* subgroup (*e. g*. David and Capy, 1988). The loci in *D. teissieri* and *D. yakuba*, therefore, should have approximately the same amount of sequence divergences to other members of the *D. melanogaster* subgroup. The sequences of the histone 3 (*H3*) gene are from Dr. Matsuo (Tokushima University, Japan, data unpublished). When there are more than one sequence for a locus, we selected one at random from each species. The chromosome location and the accession number are listed in Table 1. The abbreviation of a locus designation is based on FlyBase. We used coding regions only for all sequences except *cecropin A2 (Cec-A2)*, *B (Cec-B)*, and *C (Cec-C)* for which we included introns. Genes with short lengths (less than 200 bp) or with unknown starting codon positions in the data file were excluded. All orthologous sequences were aligned first by ClustalW (Thompson *et al*., 1994) and then followed by manual improvement by eye. Because an accurate sequence alignment is a prerequisite, genes with uncertain alignments were discarded. In the final data set, we compiled 14, 31, and 16 pairs of DNA sequences for the comparison of *D. melanogaster* with *D. mauritiana*, *D. simulans*, and *D. yakuba*, respectively.

Maximum likelihood method. Under an assumed distribution of nucleotide substitutions per unit time, a probability model of the number of nucleotide substitutions at a locus can be obtained. Given the observed data, the probability model can be interpreted as a likelihood function of the parameters in the model. The following log likelihood equation (Takahata *et al*., 1995; Takahata and Satta, 1997) was derived under the assumption that nucleotide substitutions per unit time follow the Poisson distribution,

Table 1. A list of gene names and their accession numbers.

Chromosome	Gene	*D. melanogaster*	*D. mauritiana*	*D. simulans*	*D. yakuba* (*D. teissieri*)
X	*Cyp4D1*	AF016992		AF017005	
	Cyp4D2	X75955		AF017019	
	Pgd	M80598		U02288	
	Per	L07817	L07816	L07832	X61127
	V	M34147		U27204	
	Zw	L13880		L13875	U42750
	W	X51749		U64875	
	Nullo	X65444	U64710	U44733	U44732
	Ac	M17120		X62400	
II.	*Adh*	M17827	X63953	X57364	X57370
	Adhr	X98338			(X54118)
	Amy-d	L22734	D17729	D17733	D17737
	Amy-p	L22725	D17730	D17734	D17738
	Amyrel	AF022713	U96157	U96159	AF039561
	Gpdh	J04567			(U47809)
	Dipt	AF019020	AF019035		
	Dpp	U63857		U63854	
	Ref2p	X16993		U23930	
	Sala	X57474		M21227	
	Fbp2	S57693		AF045786	
	Pgi	U20573		L27552	L27685
	Mst26Aa	X70888	X70898	X70899	
III.	*GstD1*	X14233	M84581	M84577	M84580
	Hsp82	X03810		X03811	
	Sod	X17332		X15685	
	Tra	M17478		X66930	
	Est6	M33780	L10671	L10670	
	Act88f	M18826		M87274	
	Sry-alpha	X03121	U64715	U64718	U64719
	Cec-A2	AF018978			AB010798
	Cec-B	AF018994	AF019006		
	Cec-C	AF019007	AF019019		
	Hb	DMU17742			AJ0053576
	Mlc1	L37313	L49006	L49010	L49007
	Tpi	X57576		U60861	U60870
	Gld	M29298		U63324-5	
	Lsp1-gamma	AF016033		AF016034	
	H3	*	*	*	*

* Gift sequences from Dr. Y. Matsuo of Tokushima University, Japan.

$$L(x,y)=\sum_{i=1}^{m}\left[-n_i y-\ln(1+n_i x)+\ln\sum_{d=0}^{K_i}\frac{(n_i y)^d}{d!}\left(\frac{n_i x}{1+n_i x}\right)^{K_i-d}\right], \quad (1)$$

$$\hat{N}=\frac{\hat{x}}{4rg} \text{ and } \hat{t}=\frac{\hat{y}}{2r}. \quad (2)$$

where K_i and n_i are the number of silent nucleotide substitutions and silent sites at locus i, respectively, and m is the number of loci. Two parameters, DNA sequence divergence before speciation (x) and that after speciation (y), can be estimated by maximizing Equation 1. Their standard deviations were obtained from the inverse of the expected information matrix (Casella and Berger, 1990; Weir 1996). The estimated x and y are expected to equal $4Nrg$ and $2rt$, respectively, where g is the generation time and r is the rate of nucleotide substitutions per silent site per year. The estimated effective ancestral population size (N) and the divergence time (t) can be estimated by

Before applying this ML method, data should be examined for agreement with the assumption of rate homogeneity across loci. Genes with very different evolutionary rates will increase the variance of silent substitutions, which will lead to an overestimation of the effective ancestral population size and an underestimation of the species divergence time. It is noted that the ML method can be applied to each pair of species independently as long as the rate homogeneity across loci is held. Therefore, the locus excluded in one species pair is not necessary to be excluded in the other species pair. We explain how to test the rate heterogeneity in the next section.

For each pair of species, we first estimated the numbers of silent substitutions (K_i , $i = 1, ... m$) and silent sites (n_i, $i = 1, ... m$) at m loci. For the number of silent sites, a four-fold degenerate site was always counted as one, but this does not apply to other degenerate sites. We first computed the proportion of transitions at the four-fold degenerate sites for each pair of species. Each two-fold degenerate site was then considered so as to contribute the proportion of transitions to the number of silent sites. This rule was also applied to the three-fold degenerate site (the third codon position of isoleucine). This scheme contrasts with the Nei and Gojobori method (1986) in which they counted each two-fold and three-fold degenerate site as 1/3. Their method does not account for the rate difference between transitions and transversions. For the *cecropin* introns, all sites were considered to be silent. We used the Kimura's two-parameter method (Kimura 1980) to make corrections for multiple-hit substitutions and then obtained the total number of silent substitutions.

A set of data (K_i, n_i, $i = 1, ... , m$) that exhibited homogeneous rates across loci was applied to the ML method. Since X-linked loci contribute three-fourths of the effective population size for autosomal loci, we replaced x with $3x/4$ in Equation 1 when X-linked loci were analyzed.

Test of the rate heterogeneity across loci. Takahata and Satta (1997) used the variance-to-mean ratio (dispersion index) of the number of silent substitutions as a measure of rate homogeneity across loci in their study of primates. This ratio presents the degree of the mixture between the Poisson and geometric distributions for the number of silent substitutions. It is, however, not a test statistic that gives a threshold to reject the constant rate hypothesis. Without assuming any distribution for the number of silent substitutions, we employed the χ^2, G, and Fisher's exact tests to examine the homogeneous evolutionary rates across loci. The null hypothesis was defined as equal per-site proportion of observed differences across loci in each pair of species. Our purpose was to identify a set of loci that failed to negate this null hypothesis. Two explanations need to be addressed for this hypothesis testing. First, since the number of silent substitutions per site is a function of the proportion of observed differences under the multiple-hit correction model, our test can lead to the same result as testing the equal number of silent substitutions per site across loci. Second, based on our null hypothesis, it is possible to remove loci that may have very different coalescent times despite the evolutionary rate similar to that of other loci. Because of this, our hypothesis testing is conservative, but it is still suitable to identify a set of loci with homogeneous evolutionary rates. The details of these three tests are described in the Appendix. The Fisher's exact test provides more accurate results for small samples and the G test is closer to the approximated distribution than the χ^2 test because of its additive property (Weir, 1996). If the null hypothesis is rejected by one of the tests, it determines the rate heterogeneity across loci in this data set. A new set of loci is then formed by eliminating a locus that shows the highest value in the χ^2 test (the locus with the highest X_i^2 in Appendix). We repeated these three tests until all three tests yielded nonsignificant results. For the Fisher's exact test, it is computationally difficult to survey all possible configurations for a fixed total number of silent differences when we have more than two loci. Thus, we carried out a Markov chain procedure (Raymond and Rousset, 1995). The algorithm of this Markov chain approach is described in the Appendix. For the final set of loci, we computed its dispersion index by

$$\hat{S} = \frac{\sum K_i(K_i -)(\sum n_i)^2}{(\sum K_i)^2 \sum n_i^2}. \quad (3)$$

To investigate possible causes of the rate heterogeneity in the *D. melanogaster* subgroup, we compared the GC content and codon usage bias to the rate of silent substitutions at the intersepecific level. For coding regions, the GC content was computed at the third codon position, because most, if not all, nucleotide changes at the third codon position are synonymous. For introns, we counted the number of Gs and Cs in the sequence. For the codon usage bias, we calculated the effective number of codons (ENC) (Wright, 1990; Powell and Moriyama, 1997), which is analogous to the effective number of alleles (Crow and Kimura, 1970). The ENC ranges from 20 to 61 and is correlated negatively with the codon usage bias. When all codons are used equally, the ENC should reach 61 (Wright, 1990).

RESULTS

Overall, 39 orthologous loci were surveyed. Of these, 30 are autosomal loci, and 9 are X-linked loci (Table 1). The proportion of transitions at the four-fold degenerate sites for each species pair is approximately 0.5 (Table 2). Therefore, to obtain the number of silent sites (n_i), we counted each two- and three-fold degenerate site as 1/2. Given the number of silent sites (n_i), we estimated the number of silent substitutions (K_i) by Kimura's (1980) two-parameter model (Table 3). The numbers of nucleotide substitutions and silent sites for *amylase* (*Amy*) gene are a combination for the *amylase-distal* (*Amy-d*) and *amylase-proximal* (*Amy-p*) genes. For each locus, the comparison between *D. melanogaster* and *D. yakuba* always yielded a greater number of silent substitutions than did the other two comparisons.

The distribution of silent substitutions per site for each species pair is presented by a box plot (Figure 1). A rela-

Table 2. The numbers of transitions (Ts) and transversions (Tv) at the four-fold degenerate sites.

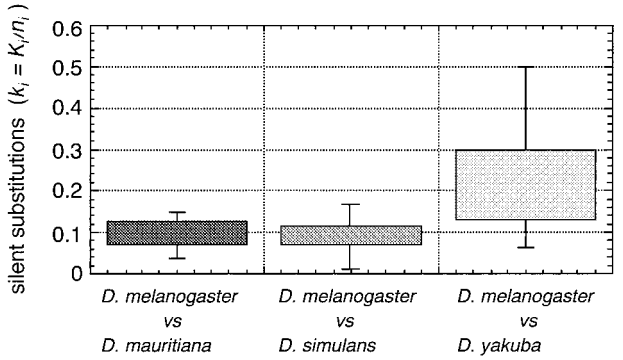
D. melanogaster vs.	No. of Loci	Ts	Tv	Total No. of sites	Ts/(Ts+Tv)*
D. mauritiana	14	116	94	2072	0.55
D. simulans	31	273	231	5810	0.54
D. yakuba	16	228	260	2882	0.46

* The proportion of transitions at the four-fold degenerate sites.

Table 3. The numbers of silent nucleotide substitutions (K_i) and silent site (n_i) for each locus in each pair of species (written as $K_i(n_i)$).

	Gene	*D. melanogaster* vs *D. simulans*	*D. melanogaster* vs *D. mauritiana*	*D. melanogaster* vs *D. yakuba*
Autosomal	*Amy*	60 (755)*	66 (755)*	80 (756)
	Amyrel	53 (419)*	49 (421)	92 (421)*
	Mlc1	3 (62)*	3 (62)*	4 (63)
	Adh	14 (211)*	11 (211)*	28 (208)*
	GstD1	10 (146)*	18 (146)*	19 (146)*
	Cec-A2			45 (211)*
	Cec-B	4 (109)		
	Cec-C	18 (120)		
	Dpp		19 (465)	
	Hsp82		13 (275)*	
	Act88f		15 (291)*	
	Pgi		26 (431)*	91 (430)*
	Hb			101 (555)*
	Gpdh		14 (200)*	31 (201)*
	Ref2p		33 (460)*	
	Sala		18 (108)	
	Fbp2		18 (155)*	
	Sod		12 (120)*	
	Tra		23 (140)	
	Est6	49 (408)*	50 (408)	
	Sry-alpha	47 (381)*	49 (383)	120 (379)
	Adhr			65 (200)
	Dipt	9 (78)*		
	Mst26Aa	32 (218)	35 (221)	
	Tpi		16 (195)*	29 (195)*
	Lsp1-gamma		66 (489)	
	H3	12 (111)*	12 (111)*	29 (111)*
	Gld		53 (545)*	
X-linked	*Ac*		16 (142)*	
	Cyp4d1		42 (396)*	
	Cyp4d2		42 (379)*	
	Pgd		50 (376)*	
	Nullo	15 (134)*	13 (135)*	67 (134)
	V		42 (293)*	
	W		62 (560)*	
	Per	56 (419)*	46 (420)*	125 (421)
	Zw		36 (389)*	51 (389)
Dispersion Index **		0.99	1.04	1.06

* The largest set of loci that show no rate heterogeneity.
** It is computed for the loci with asterisk.



	D. melanogaster vs D. mauritiana	D. melanogaster vs D. simulans	D. melanogaster vs D. yakuba
Mean	0.102	0.097	0.211
SD	0.036	0.038	0.109

Fig. 1. The distribution of silent substitutions for each pair of species. The box in a boxplot contains the middle half of the data and the whiskers extending from the box reach to the maximum and minimum of the data.

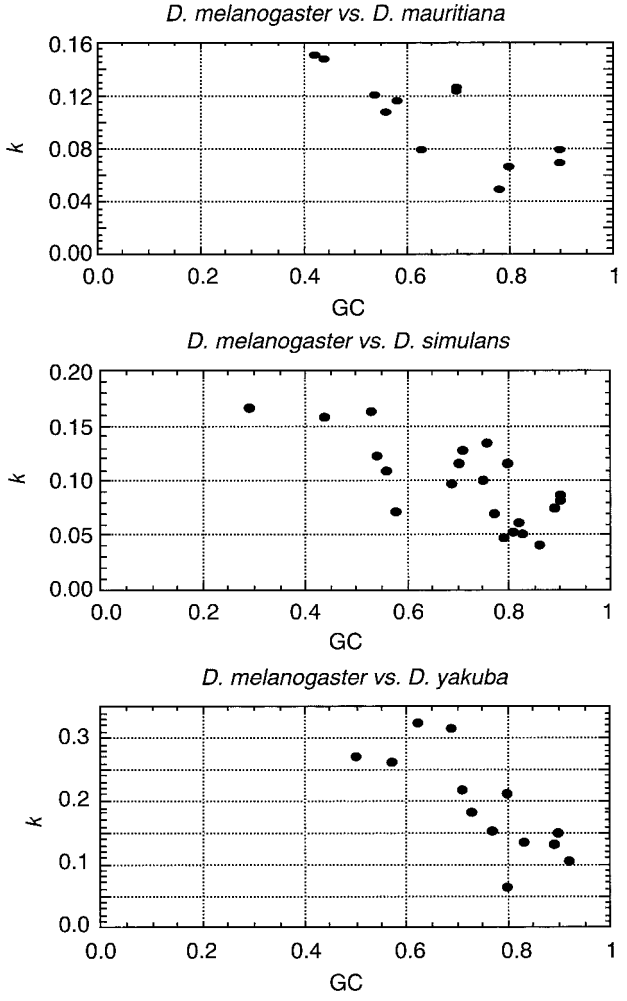


Fig. 2 The correlation between silent substitution rates (k) and the GC content of all autosomal loci in each pair of species. A negative correlation pattern is observed in each plot. The correlation coefficients are -0.68, -0.73, and -0.78 for the comparisons of *D. melanogaster* to *D. mauritiana*, *D. simulans*, and *D. yakuba*, respectively.

tively large standard deviation (SD) was observed for each pair of species. For instance, the standard deviation for 16 loci of *D. melanogaster* and *D. yakuba* was as high as 0.109. This indicates a wide range of per site silent substitutions in the *D. melanogaster* subgroup. Therefore, it is necessary to test the rate heterogeneity across loci. The final set of loci, which was not rejected by the three statistical tests we used, consists of 11, 23, and 9 pairs of DNA sequences for the comparison of *D. melanogaster* with *D. mauritiana*, *D. simulans*, and *D. yakuba*, respectively (Table 3). Their dispersion indexes obtained from Equation 3 are close to 1 (Table 3), which indicates that the distribution of the number of silent substitutions in the final set of loci is close to the Poisson distribution.

A negative correlation between the GC content and the number of silent substitutions per site from all autosomal loci was observed in all three species pairs (Figure 2). Their correlation coefficients ranged from -0.78 to -0.67. The lower the GC bias, the larger the number of silent substitutions per site. Furthermore, the relationship between ENC and the GC content was examined. Owing to the limited number of loci used for *D. melanogaster-D. mauritiana* and *D. melanogaster-D. yakuba*, only *D. melanogaster-D. simulans* shows a clear relationship between ENC and the GC content. The ENC peaks at approximately 50% GC content and decreases as the GC content moves away from 50% (Figure 3). The molecular mechanism that affects the relationships among codon usage bias, GC content, and rate of silent substitution in a gene is still not clear. We discuss more on this point in the next section.

The maximum likelihood estimates of x and y for each pair of species and their standard deviations are given in Table 4. We also include the overall mean nucleotide diversity of 24 loci in *D. melanogaster* from Moriyama and Powell (1996), which can be interpreted as the estimate of

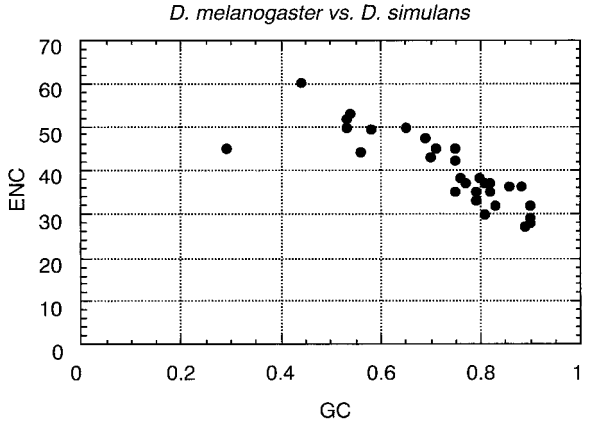


Fig. 3. The correlation between the effective number of codons (ENC) and the GC content from all loci between *D. melanogaster* and *D. simulans* in Table 1.

x in the extant *D. melanogaster* population. Assuming a constant rate of nucleotide substitution during evolution, we can further infer the species divergence time and the effective ancestral population size from these maximum likelihood estimates (Equation 2). To do so, we must know the evolutionary rate of silent substitutions in the *D. melanogaster* subgroup. Owing to the lack of fossil records, it is generally difficult to calibrate the evolutionary rate in insects. However, various rates of silent substitutions per site per year have been suggested previously for *Drosophila*, and they are within the two-fold range from 1×10^{-8} to 2×10^{-8} (Moriyama, 1987; Caccone *et al*., 1988; Sharp and Li, 1989; Russo *et al*., 1995). If the divergence time between *D. melanogaster* and *D. simulans* is 2 to 3 mya as suggested by Lemeunier *et al*. (1986), the evolutionary rate of silent substitutions is between 1.2×10^{-8} and 1.8×10^{-8} per site per year based on our ML estimate of y between these two species. Here, we used an intermediate rate 1.6×10^{-8} (Sharp and Li, 1989) to estimate the species divergence time and the effective ancestral population sizes. Both *D. mauritiana* and *D. simulans* show similar species divergence times from *D. melanogaster* (2.7 ± 0.4 mya and 2.3 ± 0.3 mya, respectively). The divergence time of *D. yakuba* from *D. melanogaster* is estimated to be 5.1 ± 0.8 mya, which is about twice that between *D. simulans* and *D. melanogaster*.

The effective ancestral population size (N) reflects the demographic history between t/g and $t/g+2N$ generations ago. Table 4 shows Ng to be 3.1×10^5, 3.9×10^5, and 3.3×10^5 for the ancestral lineage of *D. mauritiana*, *D. simulans*, and *D. yakuba*, respectively, always compared with *D. melanogaster*. Given the same evolutionary rate (1.6×10^{-8}), the value of Ng in extant *D. melanogaster* is estimated as 2.0×10^5.

In Figure 4, the 90% confidence intervals of x and y are indicated by the innermost line in the contour plots of the log likelihood function from Equation 1. The rest of contour lines are rather arbitrary. Owing to the limited number of samples, only the plot for *D. melanogaster-D. simulans* shows reasonable confidence intervals for x and y, while the other two plots show rather large confidence intervals. The standard deviations of x and y also explain the discrepancies among the widths of their confidence intervals from all species pairs (Table 4). For instance, the standard deviations of x and y for *D. melanogaster-D. simulans* were estimated as 0.012 and 0.009, respectively, which are the smallest estimates among three species pairs. It is consistent to the narrow confidence intervals of x and y observed in this species pair. We also observed

Table 4. Summary of maximum likelihood estimates.

D. melanogaster vs.	x (SD(x))[1]	y (SD(y))[2]	$x/0.013$	$y/0.072$	Divergence time (SD)[3]	Ng
D. melanogaster	0.013		1			2.0×10^5
D. mauritiana	0.02 (0.015)	0.085 (0.014)	1.5	1.2	2.7 (0.44)	3.1×10^5
D. simulans	0.025 (0.012)	0.072 (0.009)	1.6	1	2.3 (0.28)	3.9×10^5
D. yakuba	0.021 (0.024)	0.164 (0.025)	1.6	2.3	5.1 (0.78)	3.3×10^5

1 ML estimates of the DNA divergence before speciation (x) and its standard deviation (SD(x)).
2 ML estimates of the DNA divergence after speciation (y) and its standard deviation (SD(y)).
3 The divergence times and their standard deviations are in units of million years.

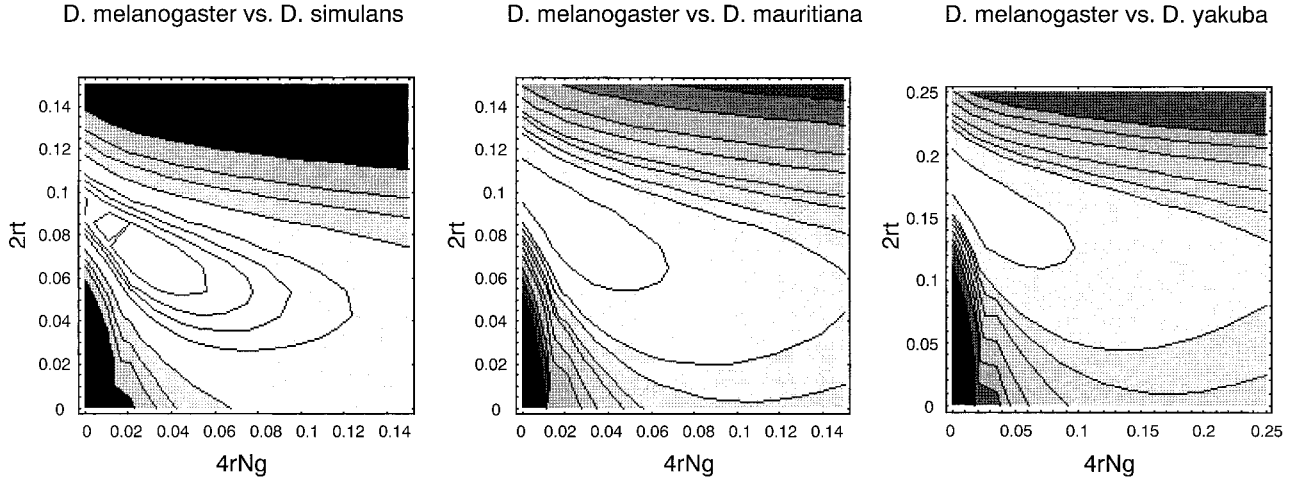


Fig. 4. Contour plots of the log likelihood function of $x = 4Nrg$ (abscissa) and $y = 2rt$ (ordinate) in Equation 1. The 90% confidence interval is depicted by the innermost contour line. It can reflect the 90% confidence intervals of x and y. Log likelihood values of all other contour lines are arbitrary.

a pattern of which the confidence intervals of x from all three species pairs overlap with each other and the confidence interval of y for *D. melanogaster-D. simulans* is within that for *D. melanogaster-D. mauritiana*. Further, the upper and lower limits of y between *D. melanogaster* and *D. yakuba* are roughly twice larger than those between *D. melanogaster* and *D. simulans*, which is congruent with the relationship of the ML estimates of y between these two comparisons.

DISCUSSION

In recent years, ancestral polymorphism has been found to be important for phylogenetic studies, particularly among closely related species. Several mathematical models (Takahata, 1986; Takahata *et al*., 1995; Takahata and Satta, 1997; Yang, 1997; Wakely and Hey, 1997) allow us to estimate parameters in the current and ancestral populations. It should be noted that ancestral polymorphism places the time of gene divergence earlier than does the species divergence time. Thus, the ancestral polymorphism should contribute part of the observed sequence divergence. In our data for the *D. melanogaster* subgroup, the extent of ancestral polymorphism is not extensive (about 2%), but it still has some effects on the estimation of the speciation time. For example, the average number of silent substitutions per site in the same set of loci from *D. melanogaste*r and *D. simulans* (Table 4) is 0.091. If we ignore the ancestral polymorphism, the speciation time would have been estimated as 2.8 mya for the same evolutionary rate of 1.6×10^{-8} substitutions per silent site per year. This is slightly greater than our estimate of 2.3 mya based on the ML method, although the difference is not significant. When ancestral polymorphism is substantial, this ML method and other proposed methods may estimate very different species divergence times.

The rates of silent substitutions in *Drosophila* were reported to vary among genes (Sharp and Li, 1989; Moriyama and Gojobori, 1992). Why the rate of silent substitutions differs from locus to locus is less clear. It may be due to differences in genomic regions, codon usage bias, GC content and so on (Li 1997). Our analysis shows a negative correlation between the GC content and the number of silent substitutions per site. The *D. melanogaster* subgroup genes examined tend to have the GC content greater than 50% at the third codon position. From the relationship between ENC and the GC content, the codon usage bias is correlated positively with the GC content of these genes. These results support those of earlier reports (Shields *et al*., 1988; Sharp and Li, 1989; Powell and Moriyama, 1997). Moriyama and Powell (1997) showed that the *Adh* gene has a high GC content. They suggested that T to C changes are expected to predominate over other transition changes. If A to G or T to C changes occur more frequently than the reverse in *Drosophila*, GC-rich genes should remain stable while GC-poor genes should undergo more nucleotide substitutions (*Sala*, which has the 29% GC content, shows the highest silent substitution rate). Furthermore, relative tRNA abundance was also suggested to be related to synonymous codon preference in *Drosophila* genes (Sharp and Lloyd, 1993; Akashi 1995; Moriyama and Powell, 1997). In prokaryotes, codons recognized by abundant tRNAs are used more frequently than those recognized by less abundant tRNAs (Ikemura, 1981; Osawa, 1995). When G-ending and C-ending codons are favored by abundant tRNAs, the A to G and T to C transitions should occur more frequently than the reverse. This selective constraint on tRNA availability may explain the high GC content in *Drosophila* genes, leading to the rate heterogeneity across loci in these data.

In this study, we employed three testing methods to search for a set of loci that supports the assumption of a constant silent substitution rate across loci in the ML method. An alternative way is to consider the rate variation among loci in the ML method. This allows the original set of data to be used fully. Yang (1997) introduced the Gamma distribution for the evolutionary rate into this ML method. With his modification, one more parameter needs to be considered, which may affect the accuracy of parameter estimations. Furthermore, the estimation of ancestral population size is sensitive to the shape parameter (α) of the Gamma distribution. It is uncertain what value of the shape parameter is appropriate for use with the *D. melanogaster* subgroup. The other question that can be raised for this modification is the robustness of the Gamma distribution for the evolutionary rates among loci in *Drosophila*. Further investigation is necessary.

From our ML estimates, we can summarize the speciation time estimates (t) as follows. First, *D. mauritiana* and *D. simulans* may have diverged from *D. melanogaster* at approximately the same time. The speciation event between *D. simulans* and *D. mauritiana* was reported to have occurred ~770,000 years ago (Wakely and Hey, 1997). Our results support earlier reports that suggested *D. melanogaster* split off from *D. mauritiana* and *D. simulans* before speciation of the latter two (*e. g*. Cariou, 1987). Second, the divergence time between *D. melanogaster* and *D. yakuba* is about two-fold older than that between *D. melanogaster* and *D. simulans*. Our estimate for *D. yakuba* is, however, restricted by a rather small sample size. The upper limit of the confidence interval of t in *D. melanogaster-D. yakuba* is approximately 10 mya. However, this estimation may well be influenced by the fact that we could use only nine loci for this species pair. The confidence interval should narrow when the sample size increases. Thus, the upper limit of t could be less than 10 mya. In contrast to some earlier suggestions (Beverley and Wilson, 1982; Bodmer and Ashburner,

1984), we believe that *D. yakuba* did not diverge from *D. melanogaster* more than 10 mya. It should be within the two-fold divergence time between *D. melanogaster* and *D. simulans*.

Whether or not the effective population size remains stable through the evolutionary time is one of interests in the study of molecular evolution. This question can be answered by comparing the effective population sizes for both ancestral and extant populations. In our analysis, the effective population sizes of all three ancestral lineages of *D. melanogaster* were found to be similar. As we described earlier, Table 4 only presents the results of *Ng*. If we assume 10 generations per year ($g = 0.1$) for the *D. melanogaster* subgroup (Sawyer and Hartl, 1992), the effective population size (N) for each ancestral lineage is ten-fold larger than the values given in Table 4. Comparing three ancestral population sizes of *D. melanogaster* to its extant population size, the bottleneck effect may not have operated during the evolutionary time of *D. melanogaster*. This scenario is different from that of primates, in which the effective ancestral population size of human was at least ten-fold larger than that of the extant population (Takahata and Satta, 1997). Li and Sadler (1991) showed that the nucleotide diversity in humans is of one order of magnitude lower than the diversity in *Drosophila* populations. From this observation and taking into account different evolutionary rates and generation times, we can compute the ratio of the effective population in the extant human to that in the extant *Drosophila*. Assuming that the nucleotide substitution rate per site per year is 1×10^{-9} and the generation time is 25 years in the human, the ratio of the human to *Drosophila* in effective population size for the extant population is less than 1%. The effective population sizes were estimated as 10^4 and 10^5 for the extant human and their ancestral population, respectively (Takahata and Satta, 1997). Thus, our finding that effective population sizes in the extant human is of approximately two order of magnitude lower than that in the extant *D. melanogaster* is consistent with the above ratio. Moreover, our results indicate that the ratio of the human to *D. melanogaster* in the effective ancestral population size is approximately 10%.

On the other hand, *D. simulans* shows different pattern of the demographic history from *D. melanogaster*. Akashi (1995) addressed that *D. simulans* has an about three- to sixfold smaller effective population size than *D. melanogaster*. Combining his finding to our results, *D. simulans* may encounter significant reduction in population size. Therefore, the stability of population size may not be held in the *D. simulans* lineage.

Obviously, the above estimates of species divergence time (t) and effective ancestral population size (N) depend on the evolutionary rate we used. It is true especially when we discuss the absolute values of t and N. However, the estimated DNA sequence divergences before and after speciation of each species pair still offers us the same relative t and N ratios between species pairs (Table 4). For instance, the ratio of *D. melanogaster-D. yakuba* to *D. melanogaster-D. simulans* in the estimated DNA sequence divergence after speciation is approximately 2 (Table 4). It implies that the twice older divergence time of *D. melanogaster-D. yakuba* than that of *D. melanogaster-D. simulans* holds true irrespective of the absolute evolutionary rates. As mentioned earlier, various evolutionary rates proposed for *Drosophila* are within a rather small range. Thus, our estimates using the intermediate rate of 1.6×10^{-8} should not be profoundly affected when the real rate becomes available. In fact, the factor that may alter our results most significantly is the number of available orthologous loci. Thus, it is important to gather more DNA sequences for this type of analysis. Because the number of DNA sequences continues to increase in DNA banks, this practice will become feasible in the near future.

We appreciate the comments from Drs. Werner Mayer, Peter Waddell, and Ziheng Yang on the preparation of this manuscript. We also thank Dr. Yoshinori Matsuo for giving us his unpublished H3 sequences. This is contribution no. 9 from the Department of Biosystems Science, Graduate University for Advanced Studies.

REFERENCES

Akashi, H. (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139**: 1067–1076.

Casella, G. and Berger, R. L. (1990) *Statistical Inference*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Crow, J. and Kimura, M. (1970) *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Beverley, S. M. and Wilson, A. C. (1982) Molecular evolution in *Drosophila* and higher Diptera. I. Micro-complement fixation studies of a larval hemolymph protein. J. Mol. Evol. **18**, 251–264.

Bodmer, M. and Ashburner, M. (1984) Conservation and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. Nature **309**, 425–430.

Caccone, A., Amato, A. D., and Powell, J. R. (1988) Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. Genetics **118**, 671–683.

Cariou, M. L. (1987) Biochemical phylogeny of the eight species in the *Drosophila melanogaster* subgroup, including *D. sechellia* and *D. orena*. Genet. Res. Camb. **50**, 181–185.

David, J. R, and Capy, P. (1988) Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4 (4),** 106–111.

Fisher, R. A. (1935) The logic of inductive inference. J. Roy. Stat. Soc. **98**, 39–54.

Fuerst, P. A., Chakraborty, R., and Nei, M. (1977) Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics **86**, 455–483.

Hamblin, M. T., and Aquadro, C. F. (1996) High nucleotide sequences variation in a region of low recombination in *Droso-*

phila simulans is consistent with the background selection model. Mol. Biol. Evol. **13**, 1133–1140.
Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translation system. J. Mol. Biol. **151**, 389–409.
Lachaise D., Cariou, M. L., David, J. R., Lemeunier, F., Tsacas, L., and Ashburner, M. (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22**, 159–225.
Lemeunier, F. and Ashburner, M. (1976) Relationships within the *melanogaster* species subgroup of the genus *Drosophila* (*Sophophora*). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc. R. Soc. Lond. B. Biol. Sci. **193 (1112)**, 275–294.
Lemeunier F., David, J. R., Tsacas, L., and Ashburner, M. (1986) The *melanogaster* species group. Vol 3e, pp. 147–256 in *The Genetics and Biology of Drosophila*, edited by Ashburner, M. , Thompson, J. N. Jr., and Carson, H. L. Academic Press, New York.
Li, W.-H. and Sadler, L. A., (1991) Low nucleotide diversity in man. Genetics **129,** 513–523.
Li, W.-H. (1997) *Molecular Evolution*, Sinauer, Massachusetts.
Kimura, M., (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**, 111–120.
Moriyama, E. (1987) Higher rates of nucleotide substitution in *Drosophila* than in mammals. Jpn. J. Genet. **62**, 139–147.
Moriyama, E. N. and Gojobori, T. (1992) Rates of synonymous substitution and base composition of nuclear genes in *Drosophila*. Genetics **130**, 855–864.
Moriyama, E. N. and Powell, J. R. (1996) Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13**, 261–277.
Moriyama, E. N. and Powell, J. R. (1997) Synonymous substitution rates in *Drosophila*: Mitochondrial versus nuclear genes. J. Mol. Evol. **45**, 378–391.
Nei, M. and Graur, D. (1983) Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. **17**, 73–113.
Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**, 418–426.
Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York.
Osawa, S. (1995) *Evolution of the Genetic Code*. Oxford, New York.
Powell, J. R., Caccone, A., Amato, G. D., and Yoon, C. (1986) Rates of nucleotide substitution in *Drosophila* mitochondrial DNA and nuclear DNA are similar. Proc. Natl. Acad. Sci. USA **83**, 9090–9093.
Powell, J. R. and Moriyama, E. N. (1997) Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. USA **94**, 7784–7790.
Raymond, M. and Rousset, F. (1995) An exact test for population differentiation. Evolution **49**, 1280–1283.
Russo, C. A., Takezaki, N. and Nei, M. (1995) Molecular phylogeny and divergence time of *Drosophilid* species. Mol. Biol. Evol. **12**, 391–404.
Sawyer, S. A. and Hartl, D. L. (1992) Population genetics of polymorphism and divergence. Genetics **132**, 1161–1176.
Sharp, P. M. and Li, W.-H. (1989) On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Evol. **28**, 398–402.
Sharp, P. M. and Lloyd, A. T. (1993) Codon usage, pp. 378–397 in *An Atlas of Drosophila Genes*: *Sequences and Molecular Features*, edited by G. Maroni. Oxford Univ. Press, New York.
Shields, D. C., Sharp, P. M., Higgins, D. G., and Wright, F. (1988) "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Mol. Biol. Evol. **5**, 704–716.
Sokal, R. R. and Rohlf, F. J. (1981) *Biometry*. 2nd edition, W. H. Freeman and Company, New York.
Stephens, J. C. and Nei, M. (1985) Phylogenetic analysis of polymorphic DNA sequences at the Adh locus in *Drosophila melanogaster* and its sibling species. J. Mol. Evol. **22**, 289–300.
Takahata, N. and Satta, Y. (1997) Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc. Natl. Acad. Sci. USA **94**, 4811–4815.
Takahata, N., Satta, Y. and Klein, J. (1995) Divergence time and population size in the lineage leading to modern humans. Theor. Pop. Biol. **48**, 198–221.
Takahata, N., (1986) An attempt to estimate the effective population size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet. Res. **48**, 187–190.
Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W, Nucleic Acids Res. **22**, 4673–4680.
Wakely, J. and Hey, J. (1997) Estimating ancestral population parameters. Genetics **145**, 847–855.
Weir, B. (1996) *Genetic Data Analysis*. Sinauer, Massachusetts.
Wright, F. (1990) The 'effective number of codons' used in a gene. Gene **87**, 23–29.
Yang, Z. (1997) On the estimation of ancestral population sizes of modern humans. Genet. Res. Camb. **69**, 111–116.
Zouros, E. (1979) Mutation rates, population size and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics **92**, 623–646.

APPENDIX

We denote the number of observed differences as d_i and the number of silent sites as n_i at locus i. To test the rate homogeneity across loci, we hypothesize the null hypothesis as H_0: $p_1 = p_2 = \ldots = p_m = p$, where $p_i = d_i / n_i$ for locus i. Three test statistics are described in the following. A computer program written in ANSI C is available upon request.

1. **χ^2 test** : The data structure here can be written as a simple $m \times 2$ contingency table. The expected number of nucleotide differences is computed by

$$\bar{d}_i = E(d_i) = n_i \times \frac{d}{n},$$

where d is the sum of d_i. We can compute the following statistic which is approximated to χ^2 distribution with $m-1$ degree of freedom,

$$X^2 = \sum_{i=1}^{m} X_i^2 = \sum_{i=1}^{n} \frac{(d_i - \bar{d}_i)^2}{\bar{d}_i^2} \sim \chi^2_{m-1}.$$

If X^2 is greater than χ^2_{m-1} at 5% level, the null hypothesis is rejected.

2. **G test** : For any one of loci, the probability of having d_i silent nucleotide differences actually follows the binomial distribution. Under the assumption of locus inde-

pendence, the total likelihood over loci becomes

$$L = \prod_{i=1}^{m} \binom{n_i}{d_i} p_i^{d_i} (1-p_i)^{n_i - d_i}.$$

Referring to the principle of G-test from Sokal and Rohlf (1981), the G statistics is simplified as

$$G = 2\ln\frac{L}{L_0} = 2\sum_{i=1}^{m}\left[d_i \ln\frac{d_i}{\bar{d}_i} + (n_i - d_i)\ln\frac{n_i - d_i}{n_i - \bar{d}_i}\right] \sim \chi^2_{m-1},$$

where L_0 is the likelihood function under the null hypothesis. If G is greater than χ^2_{m-1} at 5% level, the null hypothesis is rejected.

3. **Fisher exact test**: Based on our data structure, the exact value of type I error probability (p-value) is the proportion of the tables which have the same or less probabilities than the observed table under the condition of the same total nucleotide differences (Fisher, 1935). The null hypothesis is rejected when the p-value is less than the significant level α. The conditional probability of each table is derived as

$$P(d_1, d_2, \cdots, d_m \mid d) = \frac{d!(n-d)!\prod_{i=1}^{m} n_i!}{n!\prod_{i=1}^{m} d_i!(n_i - d_i)!}.$$

It is difficult to survey all possible tables under the same total nucleotide differences (d) for multiple loci. A Markov chain procedure was proposed to test the population differentiation for a $R \times C$ contingency table (Raymond and Rousset, 1995), where R and C are integer numbers. We applied their algorithm to our data ($m \times 2$ table). Let's denote the number in each cell as $N_{i,j}$, where $N_{i1} = d_i$, $N_{i2} = n_i - d_i$ and $i = 1, \ldots, m$. The algorithm is as follows.

(a) Set variables $\rho = 0$ and $T = 0$.

(b) Draw random numbers to select two cells in the table on different rows and columns (cell $i1, j1$, and $i2, j2$).

(c) If at least one of cells is zero, go to step (b).

(d) The new state of Markov chain is represented by a new table where

$N_{i1,j1} = N_{i1,j1} - 1$
$N_{i1,j2} = N_{i1,j2} + 1$
$N_{i2,j1} = N_{i2,j1} + 1$
$N_{i2,j2} = N_{i2,j2} - 1$

(e) If the ratio ($R = N_{i1,j1}N_{i2,j2}/(N_{i2,j1} + 1)(N_{i1,j2} + 1)$) of conditional probability of two tables (the old one vs. the new one) is equal or larger than 1, the chain moves to the new state. If it is less than one, the probability to move to new state is R. If the new state is reached, $\rho = \rho + ln$ (R).

(f) If ρ is equal or less than 0, $T=T+1$. T is the number of times that the Markov chain has encountered the tables with a lower or equal probability than the observed one.

(g) Repeat K times from (b). For example, set K = 50,000.

(h) The p-value is calculated as T/K.

We also consider a burning time of 1000 repeats before recording T as a usual Markov chain Monte Carlo (MCMC) procedure.