*Genet. Res., Camb.* (1999), **73**, *pp.* 15–28. *With 3 figures.* Printed in the United Kingdom © 1999 Cambridge University Press 

# The effect of selection on the amounts of nucleotide variation within and between allelic classes

HIDEKI INNAN AND FUMIO TAJIMA*

*Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-Ku, Tokyo 113-0033, Japan*

(*Received 5 January 1998 and in revised form 24 June 1998*)

**Summary**

The effect of selection on the amounts of nucleotide variation within and between allelic classes was studied when two allelic classes exist in a population. Two selection models – the genic selection model and the overdominant selection model – were used. The average numbers of pairwise nucleotide differences within two allelic classes were investigated by computer simulation and the average number of pairwise differences between two allelic classes was obtained analytically. It was indicated that selection largely affects the amounts of variation within and between allelic classes. However, the sum of the average numbers of pairwise differences within two allelic classes is nearly constant and always close to $\theta(\theta = 4N\mu)$, even when selection is acting, where $N$ is the effective population size and $\mu$ is the mutation rate per sequence per generation. This result suggests that the sum of the average numbers of pairwise differences within two allelic classes can be used to estimate $\theta$. It may be useful for a region where selection may be acting. As examples, several gene regions of *Drosophila melanogaster* and a region of *Mus domesticus* were analysed. The effect of recombination on the sum of the average numbers of pairwise differences within two allelic classes was discussed.

## 1. Introduction

Two or more alleles can coexist in a population because of mutations, random genetic drift, natural selection and so on (Hubby & Lewontin, 1966; Lewontin & Hubby, 1966; Harris, 1966). In order to interpret the evolutionary history and maintenance mechanism for these alleles, the amounts of nucleotide variation within and between alleles have been investigated. In our previous study (Innan & Tajima, 1997), the expectations of the average number of pairwise nucleotide differences within and between two allelic classes were obtained following the theory of gene genealogy (Griffiths, 1980; Kingman, 1982; Hudson, 1983*a*; Tajima, 1983) under the neutral model (Kimura, 1968, 1983). The allelic class is defined as follows: When DNA sequences are sampled from a population and two nucleotides are segregating in a particular site, the sequences can be divided into two classes. Such classes are called allelic classes. For example, when A and T are segregating in a site, we have two allelic classes for this site: sequences with A belong to one allelic class and sequences with T belong to the other. Assume that we have $n$ sequences sampled from a random mating population with $N$ diploid individuals, and that there are two allelic classes, A1 and A2. We also assume that A1 allelic class consists of $i$ sequences and A2 consists of $n-i$ sequences. Denote this state by $A(i, n-i)$, and the expectations of the average number of pairwise nucleotide differences within A1 allelic class, within A2 allelic class and between two allelic classes by $K_1(i, n-i)$, $K_2(i, n-i)$ and $D(i, n-i)$, respectively. Note that the amount of nucleotide variation in the population can be measured by $\theta = 4N\mu$, where $\mu$ is the mutation rate per sequence per generation. Then, Innan & Tajima (1997) have shown that these three expected values under the neutral model are given by

$$K_1(i, n-i) = \frac{i}{n}\theta, \tag{1}$$

$$K_2(i, n-i) = \frac{n-i}{n}\theta, \tag{2}$$

$$D(i, n-i) = 2\left[S(n) - \frac{i}{n}S(i) - \frac{n-i}{n}S(n-i)\right] + \frac{n-2}{n}\theta, \tag{3}$$

* Corresponding author. Telephone: +81 3 3818 5398. Fax: +81 3 3818 5399. e-mail address: ftajima@biol.s.u-tokyo.ac.jp.

Table 1. *The sum of the average numbers of pairwise differences within A1 and A2 allelic classes under the genic selection model when* n = *10*

| | $K_1(i,n-i)$ | | $K_2(i,n-i)$ | | Sum | | |
|---|---|---|---|---|---|---|---|
| *i* | Mean | Variance | Mean | Variance | Mean | Variance | Frequency[a] |
| $Ns = 0{\cdot}1$ | | | | | | | |
| 2 | 0·207 | 0·335 | 0·789 | 0·607 | 0·996 | 0·870 | 0·164 (26730) |
| 3 | 0·314 | 0·369 | 0·676 | 0·544 | 0·990 | 0·805 | 0·127 (20664) |
| 4 | 0·418 | 0·418 | 0·583 | 0·502 | 1·001 | 0·797 | 0·116 (18929) |
| 5 | 0·523 | 0·496 | 0·474 | 0·423 | 0·996 | 0·792 | 0·116 (18851) |
| 6 | 0·616 | 0·529 | 0·383 | 0·384 | 0·998 | 0·798 | 0·127 (20581) |
| 7 | 0·718 | 0·574 | 0·277 | 0·314 | 0·995 | 0·792 | 0·149 (24166) |
| 8 | 0·819 | 0·628 | 0·191 | 0·308 | 1·009 | 0·865 | 0·201 (32742) |
| All | | | | | 0·999 | 0·822 | 1·000 (162663) |
| $Ns = 1$ | | | | | | | |
| 2 | 0·354 | 0·660 | 0·593 | 0·452 | 0·947 | 0·973 | 0·036 (2952) |
| 3 | 0·514 | 0·689 | 0·467 | 0·338 | 0·982 | 0·915 | 0·043 (3526) |
| 4 | 0·623 | 0·669 | 0·362 | 0·237 | 0·985 | 0·823 | 0·058 (4750) |
| 5 | 0·736 | 0·707 | 0·299 | 0·235 | 1·035 | 0·858 | 0·087 (7095) |
| 6 | 0·807 | 0·702 | 0·232 | 0·182 | 1·039 | 0·829 | 0·131 (10633) |
| 7 | 0·870 | 0·695 | 0·174 | 0·156 | 1·044 | 0·810 | 0·222 (18065) |
| 8 | 0·920 | 0·684 | 0·121 | 0·156 | 1·041 | 0·814 | 0·423 (34415) |
| All | | | | | 1·032 | 0·830 | 1·000 (81436) |
| $Ns = 10$[b] | | | | | | | |
| 5 | 1·138 | 0·854 | 0·059 | 0·022 | 1·197 | 0·870 | 0·002 (58) |
| 6 | 0·938 | 0·822 | 0·037 | 0·021 | 0·975 | 0·847 | 0·013 (501) |
| 7 | 1·007 | 0·767 | 0·030 | 0·020 | 1·038 | 0·785 | 0·106 (3971) |
| 8 | 0·998 | 0·729 | 0·023 | 0·024 | 1·021 | 0·753 | 0·879 (32899) |
| All | | | | | 1·022 | 0·757 | 1·000 (37440) |

[a] The relative frequency of $A(i,n-i)$ is shown with the observed number of cases in parentheses.
[b] When the observed number of cases is smaller than 50, the results are not presented.

where

$$S(n) = \sum_{k=1}^{n-1} \frac{1}{k}\theta.$$

These results indicate that

$$K_1(i,n-1) + K_2(i,n-i) = \theta. \quad (4)$$

Namely, the sum of the average numbers of pairwise differences within two allelic classes is equal to $\theta$ under the neutral model.

On the other hand, let us consider a locus where two allelic classes are maintained by strong overdominant selection. In such a locus, the frequencies of these two allelic classes are expected to be close to their equilibrium values, so that the average number of pairwise differences within each allelic class might be proportional to its equilibrium frequency. Consequently, the sum of $K_1(i,n-i)$ and $K_2(i,n-i)$ might be close to $\theta$. We suspected that this relationship may hold even when selection is weak. The first purpose of the present report is to evaluate the sum of $K_1(i,n-i)$ and $K_2(i,n-i)$ under two selection models: the overdominant selection model and the genic selection model. For this purpose, computer simulations were conducted and the average numbers of pairwise differences within A1 and A2 allelic classes were investigated. The results indicate that $K_1(i,n-i) + K_2(i,n-i) \approx \theta$ not only under the overdominant selection model but also under the genic selection model, and suggest that the sum of the amounts of variation within allelic classes can be an estimate of $\theta$ even in a region where natural selection is acting.

Contrary to the constancy of the sum of the average numbers of pairwise differences within two allelic classes, the average number of pairwise differences between two allelic classes might depend on the type and strength of natural selection. In this study, the average number of pairwise differences between A1 and A2 allelic classes is also investigated under the two selection models. Although the genealogical relationship under these models is very complex if selection is involved (Kaplan *et al.*, 1988; Neuhauser & Krone, 1997), we can obtain the expectation of the average numbers of pairwise differences between A1 and A2 allelic classes when the sample consists of *i* A1 sequences and $n-i$ A2 sequences. Our analytical result is different from those of Kaplan *et al.* (1988) and Neuhauser & Krone (1997), because we do not allow any recurrent mutations between two allelic

classes after the divergence of the two allelic classes, following the infinite site model (Kimura, 1969).

## 2. The average numbers of pairwise differences within A1 and A2 allelic classes under the selection models

In order to evaluate $K_1(i, n-i)$ and $K_2(i, n-i)$ under the selection models, we conducted computer simulations. For the simulation, we employ a simple two-allele model, where two alleles, A1 and A2, exist in a random mating population with $N$ diploid individuals. In the genic selection model, the fitnesses of genotypes are given as follows:

| A1A1 | A1A2 | A2A2 |
|---|---|---|
| $1+2s$ | $1+s$ | 1 |

In the overdominant selection model, their fitnesses are

| A1A1 | A1A2 | A2A2 |
|---|---|---|
| $1-s_1$ | 1 | $1-s_2$ |

Following these fitnesses of genotypes, the computer simulations are conducted. The simulations follow the infinite site model with no recombination (Kimura, 1969; Watterson, 1975). Assume that the selection is acting on a particular site that distinguishes two allelic classes, mutations on the other sites being selectively neutral. We assume that the population size, $N$, is 5000. According to each mode of selection presented above, the frequency of A1, $x$, is determined by the pseudosampling method (Kimura, 1980; Kimura

Table 2. *The sum of the average numbers of pairwise differences within A1 and A2 allelic classes under the symmetrical overdominant selection model when* n = *10*

| | $K_1(i, n-i)$ | | $K_2(i, n-i)$ | | Sum | | |
|---|---|---|---|---|---|---|---|
| $i$ | Mean | Variance | Mean | Variance | Mean | Variance | Frequency[a] |
| $Ns_1 = Ns_2 = 0{\cdot}1$ | | | | | | | |
| 2 | 0·205 | 0·347 | 0·792 | 0·607 | 0·997 | 0·882 | 0·180 (31448) |
| 3 | 0·298 | 0·341 | 0·704 | 0·564 | 1·002 | 0·807 | 0·140 (24483) |
| 4 | 0·393 | 0·382 | 0·605 | 0·526 | 0·998 | 0·791 | 0·123 (21427) |
| 5 | 0·502 | 0·465 | 0·502 | 0·465 | 1·004 | 0·801 | 0·119 (20732) |
| 6 | 0·601 | 0·519 | 0·393 | 0·392 | 0·994 | 0·799 | 0·122 (21227) |
| 7 | 0·699 | 0·552 | 0·306 | 0·357 | 1·005 | 0·807 | 0·137 (23874) |
| 8 | 0·788 | 0·598 | 0·204 | 0·335 | 0·992 | 0·853 | 0·179 (31175) |
| All | | | | | 0·999 | 0·825 | 1·000 (174366) |
| $Ns_1 = Ns_2 = 1$ | | | | | | | |
| 2 | 0·240 | 0·390 | 0·733 | 0·560 | 0·972 | 0·876 | 0·155 (43607) |
| 3 | 0·342 | 0·385 | 0·646 | 0·497 | 0·978 | 0·791 | 0·141 (39747) |
| 4 | 0·416 | 0·387 | 0·564 | 0·448 | 0·980 | 0·749 | 0·136 (38118) |
| 5 | 0·487 | 0·412 | 0·487 | 0·419 | 0·974 | 0·732 | 0·135 (38078) |
| 6 | 0·557 | 0·437 | 0·414 | 0·388 | 0·971 | 0·733 | 0·136 (38355) |
| 7 | 0·641 | 0·489 | 0·330 | 0·367 | 0·971 | 0·772 | 0·141 (39528) |
| 8 | 0·727 | 0·540 | 0·243 | 0·407 | 0·970 | 0·879 | 0·156 (43723) |
| All | | | | | 0·974 | 0·794 | 1·000 (281156) |
| $Ns_1 = Ns_2 = 10$ | | | | | | | |
| 2 | 0·450 | 0·675 | 0·494 | 0·277 | 0·944 | 0·945 | 0·074 (69166) |
| 3 | 0·457 | 0·433 | 0·489 | 0·290 | 0·946 | 0·717 | 0·134 (125079) |
| 4 | 0·467 | 0·354 | 0·480 | 0·299 | 0·947 | 0·648 | 0·187 (174546) |
| 5 | 0·476 | 0·321 | 0·477 | 0·325 | 0·953 | 0·640 | 0·210 (194828) |
| 6 | 0·483 | 0·301 | 0·467 | 0·356 | 0·949 | 0·652 | 0·187 (174912) |
| 7 | 0·487 | 0·287 | 0·461 | 0·436 | 0·948 | 0·720 | 0·134 (125456) |
| 8 | 0·494 | 0·280 | 0·447 | 0·664 | 0·941 | 0·931 | 0·074 (69047) |
| All | | | | | 0·948 | 0·709 | 1·000 (933034) |
| $Ns_1 = Ns_2 =$ infinity[b] | | | | | | | |
| 2 | 0·500 | 0·750 | 0·500 | 0·289 | 1·000 | 1·039 | 0·045 |
| 3 | 0·500 | 0·472 | 0·500 | 0·300 | 1·000 | 0·772 | 0·120 |
| 4 | 0·500 | 0·384 | 0·500 | 0·317 | 1·000 | 0·701 | 0·210 |
| 5 | 0·500 | 0·342 | 0·500 | 0·342 | 1·000 | 0·683 | 0·250 |
| 6 | 0·500 | 0·317 | 0·500 | 0·384 | 1·000 | 0·701 | 0·210 |
| 7 | 0·500 | 0·300 | 0·500 | 0·472 | 1·000 | 0·772 | 0·120 |
| 8 | 0·500 | 0·289 | 0·500 | 0·750 | 1·000 | 1·039 | 0·045 |
| All | | | | | 1·000 | 0·744 | 1·000 |

[a] The relative frequency of $A(i, n-i)$ is shown with the observed number of cases in parentheses.
[b] The theoretical expectations are shown. The variance is calculated according to equation (30) in Tajima (1983).

Table 3. *The sum of the average numbers of pairwise differences within A1 and A2 allelic classes under the non-symmetrical overdominant selection model when* n = *10*

| | $K_1(i, n-i)$ | | $K_2(i, n-i)$ | | Sum | | |
|---|---|---|---|---|---|---|---|
| $i$ | Mean | Variance | Mean | Variance | Mean | Variance | Frequency[a] |
| $Ns_1 = 0{\cdot}01$, $Ns_2 = 0{\cdot}09$ | | | | | | | |
| 2 | 0·204 | 0·335 | 0·796 | 0·603 | 1·000 | 0·866 | 0·174 (29385) |
| 3 | 0·307 | 0·365 | 0·691 | 0·558 | 0·998 | 0·818 | 0·135 (22843) |
| 4 | 0·406 | 0·411 | 0·591 | 0·496 | 0·998 | 0·784 | 0·120 (20264) |
| 5 | 0·506 | 0·458 | 0·488 | 0·448 | 0·994 | 0·787 | 0·117 (19844) |
| 6 | 0·616 | 0·523 | 0·380 | 0·373 | 0·996 | 0·786 | 0·123 (20774) |
| 7 | 0·704 | 0·570 | 0·288 | 0·324 | 0·992 | 0·803 | 0·143 (24217) |
| 8 | 0·808 | 0·615 | 0·195 | 0·316 | 1·003 | 0·865 | 0·188 (31951) |
| All | | | | | 0·998 | 0·821 | 1·000 (169278) |
| $Ns_1 = 0{\cdot}1$, $Ns_2 = 0{\cdot}9$ | | | | | | | |
| 2 | 0·281 | 0·517 | 0·696 | 0·526 | 0·977 | 0·950 | 0·100 (18450) |
| 3 | 0·394 | 0·488 | 0·594 | 0·449 | 0·987 | 0·825 | 0·099 (18293) |
| 4 | 0·488 | 0·501 | 0·486 | 0·376 | 0·974 | 0·775 | 0·106 (19606) |
| 5 | 0·589 | 0·539 | 0·407 | 0·342 | 0·996 | 0·781 | 0·121 (22366) |
| 6 | 0·671 | 0·579 | 0·335 | 0·303 | 1·006 | 0·794 | 0·142 (26216) |
| 7 | 0·746 | 0·587 | 0·253 | 0·256 | 0·999 | 0·769 | 0·180 (33444) |
| 8 | 0·830 | 0·631 | 0·175 | 0·257 | 1·005 | 0·837 | 0·252 (46788) |
| All | | | | | 0·995 | 0·815 | 1·000 (185163) |
| $Ns_1 = 1$, $NS_2 = 9$ | | | | | | | |
| 2 | 0·853 | 1·658 | 0·202 | 0·083 | 1·055 | 1·668 | 0·002 (265) |
| 3 | 0·870 | 1·096 | 0·230 | 0·119 | 1·100 | 1·186 | 0·008 (920) |
| 4 | 0·869 | 0·869 | 0·211 | 0·108 | 1·080 | 0·977 | 0·025 (2726) |
| 5 | 0·867 | 0·748 | 0·183 | 0·102 | 1·050 | 0·853 | 0·061 (6699) |
| 6 | 0·892 | 0·733 | 0·165 | 0·103 | 1·057 | 0·828 | 0·132 (14459) |
| 7 | 0·911 | 0·706 | 0·141 | 0·106 | 1·053 | 0·805 | 0·262 (28740) |
| 8 | 0·930 | 0·688 | 0·113 | 0·129 | 1·043 | 0·813 | 0·510 (55736) |
| All | | | | | 1·049 | 0·824 | 1·000 (109545) |
| $Ns_1$ = infinity, $Ns_2$ = infinity ($Ns_1 : Ns_2 = 1:9$)[b] | | | | | | | |
| 2 | 0·900 | 1·710 | 0·100 | 0·046 | 1·000 | 1·756 | 0·000 |
| 3 | 0·900 | 1·050 | 0·100 | 0·048 | 1·000 | 1·098 | 0·000 |
| 4 | 0·900 | 0·845 | 0·100 | 0·050 | 1·000 | 0·895 | 0·000 |
| 5 | 0·900 | 0·747 | 0·100 | 0·054 | 1·000 | 0·801 | 0·006 |
| 6 | 0·900 | 0·690 | 0·100 | 0·060 | 1·000 | 0·750 | 0·042 |
| 7 | 0·900 | 0·653 | 0·100 | 0·072 | 1·000 | 0·725 | 0·217 |
| 8 | 0·900 | 0·627 | 0·100 | 0·110 | 1·000 | 0·737 | 0·734 |
| All | | | | | 1·000 | 0·735 | 1·000 |

[a] The relative frequency of $A(i, n-i)$ is shown with the observed number of cases in parentheses.
[b] The theoretical expectations are shown. The variance is calculated according to equation (30) in Tajima (1983).

& Takahata, 1983). At the start of the simulation, $x = 1/2N$ is given. If A1 is extinct (i.e. $x$ becomes 0), a new mutant A1 is introduced and $x = 1/2N$ is given at the next generation. In the same way, if A1 is fixed (i.e. $x$ becomes 1), a new mutant A2 is introduced and $x = 1-1/2N$ is given. This procedure can save time until a new mutant allelic class appears. It is not problematic because we investigate $K_1(i, n-i)$ and $K_2(i, n-i)$ only when A1 and A2 are coexisting in the population. At every generation, $x$ is recorded. Every 1000 generations, $n$ sequences are sampled from the population. Among the $n$ sequences, the number of sequences belonging to A1 allelic class, $i$, is recorded. If $2 \leqslant i \leqslant n-2$, we calculate the average number of pairwise nucleotide differences within $i$ A1 sequences and that within $n-i$ A2 sequences as follows. We first consider the genealogical relationship among A1 allelic class. The length of time, $t_1(i)$, during which $i$ A1 sequences coalesce into $i-1$ sequences is obtained by simulating the coalescent process from present to past using the previously recorded frequency of A1, $x$. Two sequences between which coalescence occurs are randomly chosen. These procedures are continued until reaching the most recent common ancestor of $i$ A1 sequences. Thus we obtain $t_1(i)$, $t_1(i-1)$, $t_1(i-2)$, $\ldots, t_1(2)$ and construct the genealogy of $i$ A1 sequences. Using this genealogical relationship, the average number of pairwise differences within A1 allelic class is calculated. Note that we assume the number of mutations on a branch with length $t$

Table 4. *The sum of the average numbers of pairwise differences within A1 and A2 allelic classes under the selection models when* n = *50*

| | $K_1(i, n-i)$ | | $K_2(i, n-i)$ | | Sum | | |
|---|---|---|---|---|---|---|---|
| *i* | Mean | Variance | Mean | Variance | Mean | Variance | Frequency[a] |
| Genic selection model ($Ns = 1$) | | | | | | | |
| 5 | 0·147 | 0·136 | 0·786 | 0·505 | 0·932 | 0·584 | 0·0029 (3213) |
| 10 | 0·290 | 0·236 | 0·647 | 0·380 | 0·937 | 0·535 | 0·0025 (2754) |
| 15 | 0·432 | 0·328 | 0·523 | 0·325 | 0·955 | 0·554 | 0·0027 (3037) |
| 20 | 0·581 | 0·451 | 0·425 | 0·255 | 1·006 | 0·627 | 0·0037 (4139) |
| 25 | 0·685 | 0·479 | 0·341 | 0·179 | 1·027 | 0·592 | 0·0053 (5937) |
| 30 | 0·754 | 0·499 | 0·273 | 0·151 | 1·027 | 0·596 | 0·0085 (9548) |
| 35 | 0·834 | 0·529 | 0·209 | 0·103 | 1·042 | 0·597 | 0·0142 (16001) |
| 40 | 0·901 | 0·549 | 0·141 | 0·071 | 1·042 | 0·599 | 0·0281 (31561) |
| 45 | 0·956 | 0·574 | 0·077 | 0·044 | 1·032 | 0·611 | 0·0747 (83876) |
| All | | | | | 1·024 | 0·603 | 1·0000 (1123178) |
| Symmetrical overdominant selection model ($Ns_1 = Ns_2 = 1$) | | | | | | | |
| 5 | 0·115 | 0·081 | 0·863 | 0·510 | 0·978 | 0·563 | 0·0263 (12299) |
| 10 | 0·223 | 0·143 | 0·757 | 0·475 | 0·979 | 0·563 | 0·0190 (8860) |
| 15 | 0·323 | 0·197 | 0·656 | 0·396 | 0·978 | 0·525 | 0·0176 (8224) |
| 20 | 0·412 | 0·252 | 0·567 | 0·347 | 0·980 | 0·526 | 0·0173 (8069) |
| 25 | 0·484 | 0·303 | 0·497 | 0·286 | 0·981 | 0·519 | 0·0169 (7904) |
| 30 | 0·576 | 0·355 | 0·405 | 0·241 | 0·982 | 0·516 | 0·0172 (8040) |
| 35 | 0·654 | 0·404 | 0·324 | 0·193 | 0·978 | 0·538 | 0·0175 (8195) |
| 40 | 0·740 | 0·438 | 0·220 | 0·142 | 0·960 | 0·534 | 0·0190 (8874) |
| 45 | 0·864 | 0·534 | 0·122 | 0·092 | 0·986 | 0·595 | 0·0255 (11922) |
| All | | | | | 0·977 | 0·552 | 1·0000 (467422) |
| Non-symmetrical overdominant selection model ($Ns_1 = 0{\cdot}1$, $Ns_2 = 0{\cdot}9$) | | | | | | | |
| 5 | 0·122 | 0·096 | 0·832 | 0·489 | 0·953 | 0·542 | 0·0136 (4851) |
| 10 | 0·242 | 0·176 | 0·725 | 0·462 | 0·968 | 0·568 | 0·0104 (3715) |
| 15 | 0·360 | 0·240 | 0·597 | 0·368 | 0·957 | 0·527 | 0·0101 (3617) |
| 20 | 0·479 | 0·333 | 0·508 | 0·309 | 0·986 | 0·536 | 0·0109 (3879) |
| 25 | 0·580 | 0·420 | 0·434 | 0·255 | 1·014 | 0·579 | 0·0125 (4446) |
| 30 | 0·657 | 0·419 | 0·341 | 0·197 | 0·998 | 0·547 | 0·0147 (5253) |
| 35 | 0·742 | 0·490 | 0·262 | 0·143 | 1·003 | 0·582 | 0·0190 (6766) |
| 40 | 0·821 | 0·507 | 0·182 | 0·102 | 1·004 | 0·572 | 0·0265 (9450) |
| 45 | 0·913 | 0·544 | 0·096 | 0·062 | 1·009 | 0·591 | 0·0484 (17262) |
| All | | | | | 0·996 | 0·574 | 1·0000 (356769) |

Results for $i = \{5, 10, 15, 20, 25, 30, 35, 40, 45\}$ are shown. The average and the variance for all the cases are calculated for all of $i$ ($2 \leqslant i \leqslant 48$).

[a] The relative frequency of $A(i, n-i)$ is shown with the observed number of cases in parentheses.

follows the Poisson distribution with mean $t\mu$. In the same way, the average number of pairwise differences within A2 allelic class is obtained by constructing the genealogy of $n-i$ A2 sequences.

The results for $n = 10$ and $\theta = 1$ are summarized in Tables 1–3. The averages and the variances of $K_1(i, n-i)$ and $K_2(i, n-i)$ are shown with the relative frequency of the cases where the allelic state was $A(i, n-i)$ during each run of simulation. One million times of sampling were conducted for each run, except that ten million samplings were conducted for $Ns = 10$ under the genic selection model.

Table 1 shows the results for the genic selection model. Three values of selection intensity were used ($Ns = 0{\cdot}1$, 1 and 10). $K_1(i, n-i) + K_2(i, n-i)$ is close to 1 for any $i$ ($2 \leqslant i \leqslant 8$), although $K_1(i, n-i)$ increases and $K_2(i, n-i)$ decreases with increasing $Ns$. Note that, if $Ns$ is large (for example, $Ns = 10$), the frequency of the advantageous A1 allelic class is usually close to 1 and it is rare to obtain a small value of $i$. The averages of $K_1(i, n-1) + K_2(i, n-i)$ for all values of $i$ ($2 \leqslant i \leqslant 8$) are also close to 1 for all three values of selection intensity, although they tend to be a little larger than 1. The variances are 0·822, 0·830 and 0·757 when $Ns = 0{\cdot}1$, 1 and 10, respectively.

Table 2 shows the results when $Ns_1 = Ns_2 = 0{\cdot}1$, 1 and 10 under the symmetrical overdominant selection model. When $i < 5$, $K_1(i, n-i)$ increases and $K_2(i, n-i)$ decreases as $Ns$ increases, whereas $K_1(i, n-i)$ decreases and $K_2(i, n-i)$ increases with increasing $Ns$ when $i > 5$. For three values of selection intensity, $K_1(i, n-i) + K_2(i, n-i)$ is close to 1 for any value of $i$. The averages for all values of $i$ ($2 \leqslant i \leqslant 8$) are also close to 1, although the average shows some reduction

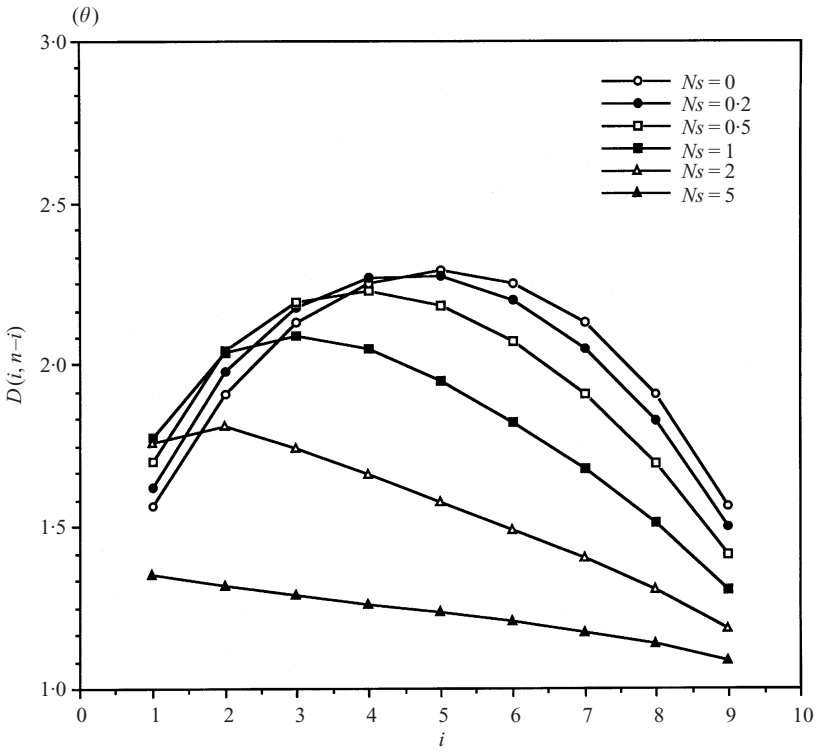


Fig. 1. The average number of pairwise differences between two allelic classes with sample size $n = 10$, under the genic selection model. The unit of the vertical axis is $\theta$.

as $Ns$ increases. The variances are 0·825, 0·794 and 0·709 when $Ns = 0{\cdot}1$, 1 and 10, respectively. As $Ns$ becomes larger, the average of $K_1(i,n-i)+K_2(i,n-i)$ is expected to approach 1 again since we expect that $K_1(i,n-i)$ and $K_2(i,n-i)$ approach 0·5. This is because we can consider that the population consists of two subpopulations with size $0{\cdot}5N$ when $Ns_1 = Ns_2 =$ infinity. The theoretical expectations and variances of $K_1(i,n-i)$ and $K_2(i,n-i)$ in this case are also shown in Table 1.

Table 3 shows the results for the non-symmetrical overdominant selection model, where $Ns_1 = 0{\cdot}01$ and $Ns_2 = 0{\cdot}09$, $Ns_1 = 0{\cdot}1$ and $Ns_2 = 0{\cdot}9$ and $Ns_1 = 1$ and $Ns_2 = 9$ are used. As $Ns_1$ and $Ns_2$ increase, $K_1(i,n-i)$ increases and $K_2(i,n-i)$ decreases. In all three cases, $K_1(i,n-i)+K_2(i,n-i)$ is close to 1 for any $i$. The averages for all values of $i$ ($2 \leqslant i \leqslant 8$) are also close to 1, although they are a little larger than 1 when $Ns_1 = 1$ and $Ns_2 = 9$. The variances are about 0·82 for three values of selection intensity. When $Ns$ is very large, we expect that $K_1(i,n-i)$ and $K_2(i,n-i)$ are close to 0·9 and 0·1, respectively. Table 3 also shows the theoretical expectations and variances of $K_1(i,n-i)$ and $K_2(i,n-i)$ when selection intensity is infinity.

The effect of sample size was also investigated. The results for $n = 50$ and $\theta = 1$ are shown in Table 4. The selection intensities used are as follows: $Ns = 1$ under the genic selection model, $Ns_1 = Ns_2 = 1$ under the symmetrical overdominant selection model and $Ns_1 =$ 0·1 and $Ns_2 = 0{\cdot}9$ under the non-symmetrical overdominant selection model. Under all three selection models, $K_1(i,n-i)+K_2(i,n-i)$ is close to 1 for any $i$. The variances of the sum of $K_1(i,n-i)$ and $K_2(i,n-i)$ are reduced in comparison with the results for $n = 10$. From these results, it can be concluded that $K_1(i, n-i)+K_2(i,n-i)$ is close to $\theta$ regardless of $i$ for a wide range of selection intensity under the genic selection model and under the overdominant selection model.

## 3. The average number of pairwise differences between A1 and A2 allelic classes

The effect of selection on the amount of nucleotide variation between two allelic classes are investigated. The expectation of the average number of pairwise differences between A1 and A2 allelic classes in $A(i, n-i)$, $D(i,n-i)$, is obtained analytically, and the derivations are presented in the Appendix. In this section, only the numerical results are shown.

From (A 7), $D(i,n-i)$ were numerically calculated when $n = 10$, and plotted in Figs. 1–3. Fig. 1 shows the expectation of the average number of pairwise differences between two allelic classes under the genic selection model. Although, under neutrality ($Ns = 0$), $D(i,n-i)$ distributes symmetrically with the highest peak when $i = 5$, the peak of the distribution of $D(i, n-i)$ moves to the left as $Ns$ increases. With strong selection, a considerable reduction in $D(i,n-i)$ is

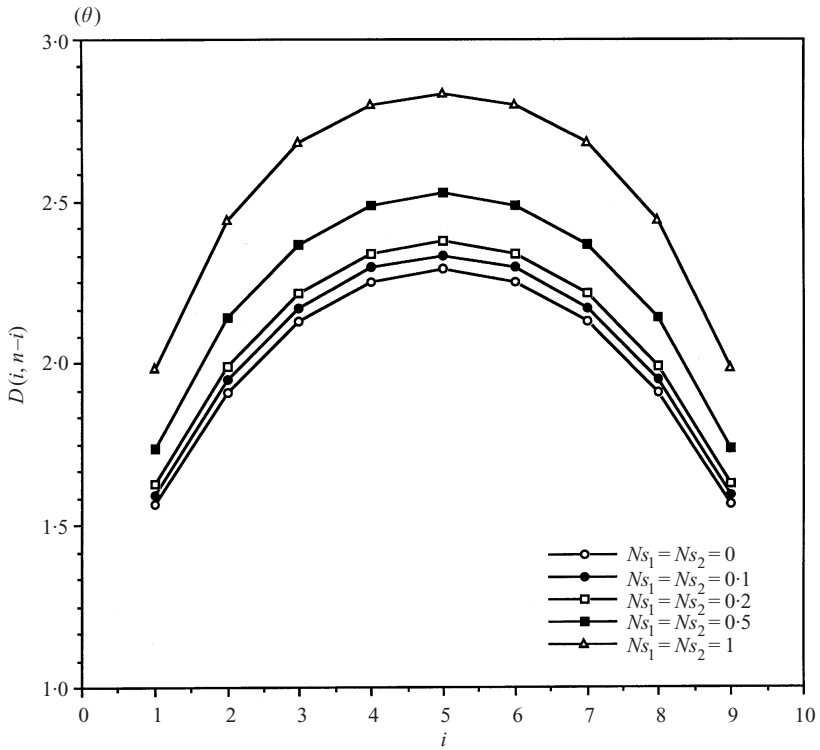


Fig. 2. The average number of pairwise differences between two allelic classes with sample size $n = 10$, under the symmetrical overdominant selection model. The unit of the vertical axis is $\theta$.

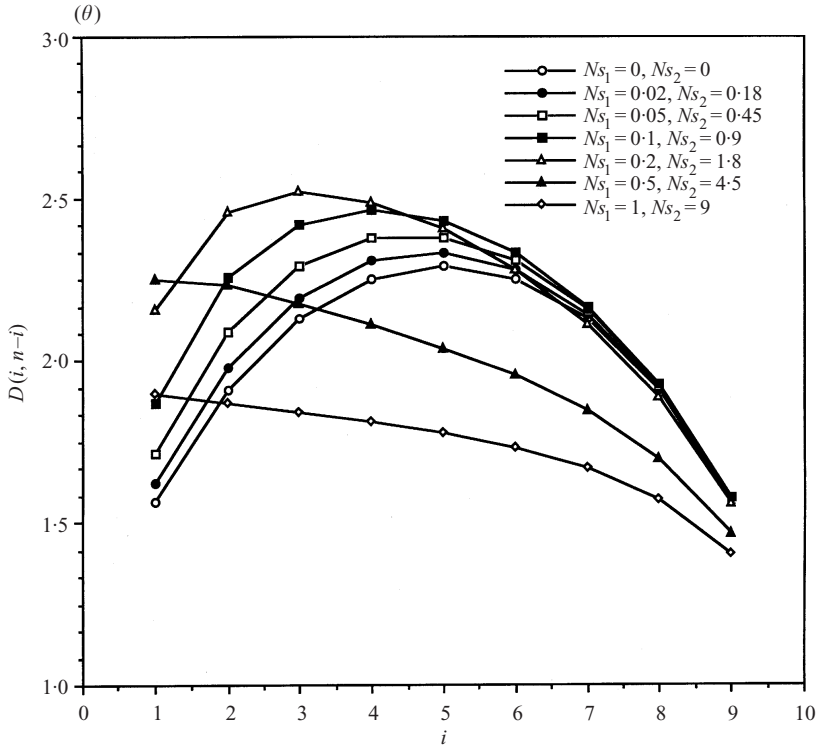


Fig. 3. The average number of pairwise differences between two allelic classes with sample size $n = 10$, under the non-symmetrical overdominant selection model. The unit of the vertical axis is $\theta$.

observed, and $D(i, n-i)$ appears like a linear function of $i$. Fig. 2 shows the plots of $D(i, n-i)$ under the symmetrical overdominant selection model. The shape of the distribution is symmetrical and similar to that under the neutral model ($Ns_1 = Ns_2 = 0$). The peak of each distribution is always in the centre ($i = 5$). D($i$,

Table 5. *Numerical examples for* $D(i, n-i)$ *under the overdominant selection model*

| | $Ns_1 = Ns_2$ | | | | |
|---|---|---|---|---|---|
| $i$ | 2 | 3 | 5 | 7 | 10 |
| 1, 9 | $2{\cdot}807\theta$ | $4{\cdot}453\theta$ | $14{\cdot}787\theta$ | $62{\cdot}399\theta$ | $693{\cdot}813\theta$ |
| 2, 8 | $3{\cdot}391\theta$ | $5{\cdot}162\theta$ | $15{\cdot}696\theta$ | $63{\cdot}410\theta$ | $694{\cdot}849\theta$ |
| 3, 7 | $3{\cdot}642\theta$ | $5{\cdot}419\theta$ | $15{\cdot}948\theta$ | $63{\cdot}647\theta$ | $695{\cdot}059\theta$ |
| 4, 6 | $3{\cdot}754\theta$ | $5{\cdot}524\theta$ | $16{\cdot}040\theta$ | $63{\cdot}728\theta$ | $695{\cdot}126\theta$ |
| 5 | $3{\cdot}787\theta$ | $5{\cdot}554\theta$ | $16{\cdot}065\theta$ | $63{\cdot}749\theta$ | $695{\cdot}143\theta$ |

$n$–$i$) increases as the selection intensity increases. For strong symmetrical overdominant selection ($Ns_1 = Ns_2 > 1$), the numerical examples of $D(i, n-i)$ are presented in Table 5. When $Ns_1 = Ns_2 = 10$, $D(i, n-i)$ is approximately $700\theta$. In other words, the mean coalescent time of two sequences sampled from different allelic classes is approximately $1400N$ generations. Fig. 3 shows the distributions of $D(i, n-i)$ under the non-symmetrical overdominant selection model. The peak moves to the left as $Ns_1$ and $Ns_2$ increase. Although the figure is similar to that under the genic selection model, the peak of distribution becomes high as $Ns$ increases when $Ns_1 \leqslant 0{\cdot}2$ under the non-symmetrical overdominant selection model (Fig. 3), whereas the peak is the highest when $Ns = 0$ under the genic selection model (Fig. 1).

## 4. Discussion

The effect of selection on the amounts of nucleotide variation within and between allelic classes was investigated. It was indicated that selection affects the average number of pairwise differences between allelic classes as shown in Figs. 1–3. The average number of pairwise differences within allelic class is also affected by selection (Tables 1–4). However, the sum of the average numbers of pairwise differences within two allelic classes is always close to $\theta$. Namely,

$$K_1(i, n-i) + K_2(i, n-i) \approx \theta \quad (5)$$

holds for any $i$ ($2 \leqslant i \leqslant n-2$) under the two selection models with a wide range of selection intensity. This means that selection has almost no effect on the sum of $K_1(i, n-i)$ and $K_2(i, n-i)$. It is also suggested that $K_1(i, n-i) + K_2(i, n-i)$ may be useful for estimating $\theta$ whether there is selection or not.

It is known that the expectation of the average number of pairwise nucleotide differences among a sample of sequences, K, is $\theta$ under the neutral model, and K is often used for the estimation of $\theta$. The variance is an important measure to know the reliability of the estimator. To test the reliability of $K_1(i, n-i) + K_2(i, n-i)$ as an estimator of $\theta$, the variance of $K_1(i, n-i) + K_2(i, n-i)$ was investigated under the neutral model and compared with the variance of K, which was theoretically obtained according to equation (30) in Tajima (1983). The results of simulations are shown in Table 6. When $n = 10$ the variance of $K_1(i, n-i) + K_2(i, n-i)$ is larger than that of K, while the variance of $K_1(i, n-i) + K_2(i, n-i)$ is smaller when $n \geqslant 20$ and $\theta \geqslant 10$. However, the difference in variance between $K_1(i, n-i) + K_2(i, n-i)$ and K is quite small, indicating that $K_1(i, n-i) + K_2(i, n-i)$ is useful for estimating $\theta$ with a similar level of reliability to K. $K_1(i, n-i) + K_2(i, n-i)$ is a little more reliable when $n$ and $\theta$ are large. When selection is acting, $K_1(i, n-i) + K_2(i, n-i)$ can give a more accurate estimate for $\theta$ than can K, since $K \neq \theta$. As shown in Tables 1–3, the variance of $K_1(i, n-i) + K_2(i, n-i)$ under the selection models is smaller than that under the neutral model (0·827), although a slightly larger variance is observed when $Ns = 1$ under the genic selection model (0·830). It is suggested that $K_1(i, n-i) + K_2(i, n-i)$ can be an estimator of $\theta$ whether selection is acting or not.

Table 6. *The average and variance of* $K_1(i, n-i) + K_2(i, n-i)$ *under the neutral model*

| | $\theta = 1$ | | | $\theta = 10$ | | | $\theta = 100$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Average | Variance | Number of cases[a] | Average | Variance | Number of cases[a] | Average | Variance | Number of cases[a] |
| $n = 10$ | | | | | | | | | |
| $K_1(i, n-i) + K_2(i, n-i)$ | 1·003 | 0·827 | 165962 | 10·00 | 36·09 | 163948 | 100·2 | 3117·3 | 164912 |
| K | 1·000 | 0·686 | | 10·00 | 31·98 | | 100·0 | 2830·9 | |
| $n = 20$ | | | | | | | | | |
| $K_1(i, n-i) + K_2(i, n-i)$ | 0·998 | 0·650 | 240640 | 9·99 | 28·09 | 240852 | 100·0 | 2441·9 | 240817 |
| K | 1·000 | 0·616 | | 10·00 | 28·42 | | 100·0 | 2510·5 | |
| $n = 50$ | | | | | | | | | |
| $K_1(i, n-i) + K_2(i, n-i)$ | 1·000 | 0·581 | 330150 | 10·00 | 25·13 | 332270 | 99·9 | 2174·9 | 330394 |
| K | 1·000 | 0·579 | | 10·00 | 26·63 | | 100·0 | 2350·3 | |

The average and the variance of $K_1(i, n-i) + K_2(i, n-i)$ when $2 \leqslant i \leqslant n-2$ are shown.
[a] Number of cases analysed in a run of simulation.

Table 7. *Analysis for ND5 gene region of* Drosophila melanogaster

| Position | Polymorphism[a] | $\hat{K}_1(i,n-i)$ | $\hat{K}_2(i,n-i)$ | Sum | $\hat{D}(i,n-i)$ |
|---|---|---|---|---|---|
| 240 | A(32)/G(27) | 0·558 | 2·154 | 2·712 | 2·281 |
| 813 | T(51)/C(8) | 1·540 | 1·250 | 2·790 | 3·581 |
| 840 | A(57)/G(2) | 2·193 | 2·000 | 4·193 | 2·211 |
| 1053 | G(52)/A(7) | 1·645 | 0·000 | 1·645 | 3·635 |
| 1122 | A(36)/G(23) | 1·033 | 1·676 | 2·709 | 2·373 |
| 1239 | G(57)/A(2) | 2·242 | 0·000 | 2·242 | 1·544 |
| 1442 | T(52)/C(7) | 1·645 | 0·000 | 1·645 | 3·635 |
| | $\hat{K}_1+\hat{K}_2$ | | | 2·562 | |
| | $\hat{K}$ | | | 2·261 | |
| | Ratio[b] | | | 1·133 | |

[a] Two segregating nucleotides are presented with the number of sequences in parentheses. The allelic class with the first nucleotide corresponds to A1 and the second to A2. Accordingly, the number in the first parentheses is $i$ and that in the second parentheses is $n-i$.
[b] The ratio of $\hat{K}_1+\hat{K}_2$ to $\hat{K}$.

When we have a sample of $n$ sequences with $m$ non-unique segregating sites, it is possible to obtain $K_1(i, n-i)+K_2(i,n-i)$ for each of $m$ sites. Note that a non-unique segregating site represents the site at which polymorphism is not unique (singleton) for the sample, so that $2 \leqslant i \leqslant n-2$. The unique segregating sites were excluded from this analysis because $K_1(1,n-1)$ or $K_2(n-1,1)$ cannot be obtained if $i=1$ or $i=n-1$, respectively. Denote the average of $m$ values of $K_1(i, n-i)+K_2(i,n-i)$ by $K_1+K_2$. We expect that $K_1+K_2$ should be equal to $\theta$. On the other hand, the expectation of K is $\theta$ under the neutral model. Therefore, when there is no selection, the ratio of $K_1+K_2$ to K is expected to be

$$(K_1+K_2)/K \approx 1. \tag{6}$$

As examples, the nucleotide polymorphism data in the mitochondrial gene regions ND5 of *Drosophila melanogaster* (Rand & Kann, 1996) and ND3 of *Mus domesticus* (Nachman *et al*., 1996) were analysed. Rand & Kann (1996) published 59 nucleotide sequences with 1515 bp, where 21 segregating sites are detected and $\hat{K}$ is 2·261. Note that the hat represents the estimated value. Among 21 segregating sites, seven exhibit non-unique polymorphism. For these non-unique segregating sites, we obtained the sum of the average numbers of pairwise differences within two allelic classes (Table 7). $\hat{K}_1(i,n-i)+\hat{K}_2(i,n-i)$ ranges from 1·645 to 4·193, and $\hat{K}_1+\hat{K}_2$ is 2·562. This value is consistent with $\hat{K}$ (2·261), and the ratio, $(\hat{K}_1+\hat{K}_2)/\hat{K}$, is 1·133. Nachman *et al*. (1996) obtained 56 nucleotide sequences with about 450 bp in ND3 of *Mus domesticus*. In these sequences, there are 27 segregating sites, of which 21 are non-unique. As shown in Table 8, the observed values of $\hat{K}_1(i, n-i)+\hat{K}_2(i,n-i)$ ranges from 2·927 to 4·286, and the average $(\hat{K}_1+\hat{K}_2)$ is 3·477, which is very close to $\hat{K}$ (= 3·328). These results may indicate that $K_1+K_2$ can be used to estimate $\theta$ as well as K.

The present study is based on the infinite site model with no recombination. Equations (5) and (6) hold under this condition. However, it is known that intragenic recombination occurs frequently in the nuclear region, and that the effect of recombination on the amount and pattern of nucleotide polymorphism may be large. Here, we consider the effect of recombination. As mentioned in our previous study (see Discussion in Innan & Tajima, 1997), if recombination occurs between two allelic classes, the amount of variation between two allelic classes decreases and the amounts of variation within both allelic classes increase. Now, let us consider the free recombination model. Under this model, since all the segregating sites are independent, it is apparent that both $K_1(i,n-i)$ and $K_2(i,n-i)$ are $\theta$, so that $K_1(i, n-i)+K_2(i,n-i)=2\theta$. Therefore, in the nuclear region where recombination occurs at a moderate rate, we expect

$$\theta < K_1(i,n-i)+K_2(i,n-i) < 2\theta, \tag{7}$$

and

$$1 < (K_1+K_2)/K < 2. \tag{8}$$

Note that K is expected to be $\theta$ even with recombination (Hudson, 1983*b*). Table 9 shows the results of analysis for the nucleotide polymorphism data in seven nuclear regions of *D. melanogaster*. In these regions, $(\hat{K}_1+\hat{K}_2)/\hat{K}$ ranges from 1·273 to 1·709 as expected from (8). In the mitochondrial gene regions (Tables 7, 8), $(\hat{K}_1+\hat{K}_2)/\hat{K}$ is smaller than those in all the seven nuclear regions in Table 9. It is suggested that the effect of recombination on the amounts of nucleotide variation is large in the nuclear regions.

Wesley & Eanes (1994) and Hasson & Eanes (1996)

Table 8. *Analysis for ND3 gene region of* Mus domesticus

| Position | Polymorphism[a] | $\hat{K}_1(i, n-i)$ | $\hat{K}_2(i, n-i)$ | Sum | $\hat{D}(i, n-i)$ |
|---|---|---|---|---|---|
| 9443 | A(52)/G(4) | 3·344 | 0·500 | 3·844 | 2·308 |
| 9461 | C(54)/T(2) | 3·292 | 0·000 | 3·292 | 2·833 |
| 9478 | T(54)/C(2) | 3·219 | 0·000 | 3·219 | 3·796 |
| 9479 | C(54)/T(2) | 3·219 | 0·000 | 3·219 | 3·796 |
| 9488 | T(53)/C(3) | 3·203 | 0·000 | 3·203 | 3·472 |
| 9497 | A(54)/G(2) | 3·364 | 0·000 | 3·364 | 1·870 |
| 9504 | A(49)/G(7) | 3·010 | 1·143 | 4·153 | 3·551 |
| 9513 | C(54)/A(2) | 3·231 | 0·000 | 3·231 | 3·648 |
| 9528 | C(52)/T(4) | 2·927 | 0·000 | 2·927 | 4·981 |
| 9530 | T(52)/C(4) | 3·189 | 0·667 | 3·856 | 3·288 |
| 9539 | A(48)/T(8) | 2·522 | 0·571 | 3·093 | 4·896 |
| 9578 | T(52)/C(4) | 3·189 | 0·667 | 3·856 | 3·288 |
| 9605 | A(49)/T(7) | 3·010 | 1·143 | 4·153 | 3·551 |
| 9624 | T(54)/C(2) | 3·292 | 0·000 | 3·292 | 2·833 |
| 9635 | T(43)/C(13) | 3·497 | 0·821 | 4·318 | 2·404 |
| 9645 | T(54)/C(2) | 3·286 | 1·000 | 4·286 | 2·907 |
| 9647 | A(53)/G(3) | 3·203 | 0·000 | 3·203 | 3·472 |
| 9692 | T(48)/C(8) | 2·522 | 0·571 | 3·093 | 4·896 |
| 9721 | T(48)/C(8) | 2·522 | 0·571 | 3·093 | 4·896 |
| 9738 | G(48)/A(8) | 2·522 | 0·571 | 3·093 | 4·896 |
| 9818 | A(54)/T(2) | 3·219 | 0·000 | 3·219 | 3·796 |
| | $\hat{K}_1 + \hat{K}_2$ | | | 3·477 | |
| | $\hat{K}$ | | | 3·328 | |
| | Ratio[b] | | | 1·045 | |

[a] Two segregating nucleotides are presented with the number of sequences in parentheses. The allelic class with the first nucleotide corresponds to A1 and the second to A2. Accordingly, the number in the first parentheses is $i$ and that in the second parentheses is $n-i$.
[b] The ratio of $\hat{K}_1 + \hat{K}_2$ to $\hat{K}$.

Table 9. *Analysis for seven nuclear regions in* Drosophila melanogaster

| Region | $n$ | $\hat{K}_1 + \hat{K}_2$ | $\hat{K}$ | Ratio[a] | Reference |
|---|---|---|---|---|---|
| Adh | 11 | 20·049 | 15·745 | 1·273 | Kreitman (1983) |
| Mlc1 | 16 | 9·893 | 6·558 | 1·509 | Clark *et al.* (1996) |
| Mst26A | 10 | 20·706 | 13·156 | 1·574 | Aguadé *et al.* (1992) |
| Hsp83 | 13 | 4·370 | 3·500 | 1·249 | Wesley & Eanes (1994) |
| Breakpoint AB | 16 | 12·955 | 9·462 | 1·369 | Hasson & Eanes (1996) |
| Est6 | 16 | 21·382 | 12·508 | 1·709 | Hasson & Eanes (1996) |
| Breakpoint CD | 13 | 6·900 | 5·231 | 1·319 | Wesley & Eanes (1994) |

[a] The ratio of $\hat{K}_1 + \hat{K}_2$ to $\hat{K}$.

investigated the nucleotide polymorphisms in four regions: both breakpoint regions of the inversion *In(3L)Payne* (breakpoint AB and CD), Hsp83 and Est-6, on the third chromosome of *D. melanogaster*. Hsp83 is located outside and near the distal breakpoint of *In(3L)Payne*, breakpoint AB is a sequence encompassing the distal breakpoint of *In(3L)Payne*, Est-6 is located between the two breakpoints of *In(3L)Payne*, and breakpoint CD is a sequence encompassing the proximal breakpoint of *In(3L)Payne*. It is expected that the recombination between different chromosome arrangements is considerably restricted in a region near the breakpoint, although recombination can occur within the same chromosome arrangement. Hasson & Eanes (1996) reported that genetic exchange between chromosome arrangements was not observed in three regions – Hsp83, breakpoints AB and CD – whereas several genetic changes between arrangements were observed in Est-6. It may be suggested that recombination is more strongly restricted in Hsp83 and breakpoint AB and CD than in Est-6. As shown in Table 9, $(\hat{K}_1 + \hat{K}_2)/\hat{K}$ is 1·709 in Est-6, which is larger than in the other three regions (1·249–1·369). This result is consistent with the expectation from the difference in the recombination rate among the four regions.

Table 10. *Analysis for four regions associated with* In(3L)payne *of* Drosophila melanogaster

| Region | $\hat{K}_{std}$ [a] | $\hat{K}_{inv}$ [b] | Sum | $\hat{K}$ | Ratio[c] |
|---|---|---|---|---|---|
| Hsp83 | 3·111 (9) | 0·476 (7) | 3·587 | 3·500 | 1·025 |
| Breakpoint AB | 10·067 (6) | 0·857 (7) | 10·924 | 9·462 | 1·155 |
| Est6 | 11·167 (9) | 11·429 (7) | 22·596 | 12·508 | 1·807 |
| Breakpoint CD | 5·267 (6) | 0·000 (7) | 5·267 | 5·231 | 1·007 |

[a] The average number of pairwise differences within the standard chromosome. The number of samples is shown in parentheses.
[b] The average number of pairwise differences within the inversion chromosome with the number of samples in parentheses.
[c] The ratio of the sum of $\hat{K}_{std}$ and $\hat{K}_{inv}$ to $\hat{K}$.

These four regions were reanalysed in Table 10, where one allelic class is defined as the standard chromosome and the other is defined as the inversion *In(3L)Payne*. $K_{std}$ represents the average number of pairwise differences within the standard chromosome and $K_{inv}$ represents that within the inversion chromosome. In this case, since only the recombination rate between two allelic classes (chromosome arrangements) can affect the sum of the amounts of nucleotide variation within two allelic classes, it is expected that the difference in $(K_{std}+K_{inv})/K$ due to the recombination rate appears more clearly than the difference in $(K_1+K_2)/K$ in Table 9. In Est-6 $(\hat{K}_{std}+\hat{K}_{inv})/\hat{K}$ is 1·807, whereas it ranges from 1·007 to 1·155 in the other three regions. As expected, $(\hat{K}_{std}+\hat{K}_{inv})/\hat{K}$ in Est-6 is larger than in the other three regions and the difference is larger than that in Table 9. The average of $(\hat{K}_{std}+\hat{K}_{inv})/\hat{K}$ in the other three regions is 1·062, which is consistent with $(\hat{K}_1+\hat{K}_2)/\hat{K}=1{\cdot}133$ in the mitochondrial gene region ND5 (Table 7), where recombination is very rare. The sum of the average numbers of pairwise differences within two allelic classes may be positively related to the recombination rate in nuclear regions.

Chromosome regions involving inversions have been studied in population genetics and non-neutral patterns of polymorphism were reported (Dobzhansky, 1937, 1970). There is a possibility that natural selection is acting on *In(3L)Payne*. If so, selection may affect on the amounts of nucleotide variation within the standard chromosome, within the inversion and between them, especially in a region with restricted recombination rate between two chromosome types. Three regions (Hsp83, breakpoint AB and CD) correspond to such regions. If *In(3L)Payne* is maintained by balancing selection, we expect $K_{std}+K_{inv}\approx\theta$ and $K>\theta$ because of a long coalescent time between the two chromosome types, so that $(K_{std}+K_{inv})/K<1{\cdot}0$ is expected. As shown in Table 10, $(\hat{K}_{std}+\hat{K}_{inv})/\hat{K}$ is a little larger than 1·0, indicating that K is not larger than $\theta$. This is not consistent with the hypothesis that *In(3L)Payne* is maintained for a long time by strong balancing selection, but is rather consistent with the neutral theory, as already suggested by Hasson & Earns (1996).

Our results demonstrate that $K_1(i,n-i)+K_2(i,n-i)\approx\theta$ holds even under the selection models, suggesting that we can estimate $\theta$ by $K_1(i,n-i)+K_2(i,n-i)$ in a region with selection and without recombination. This result is notable because K gives a biased estimate of $\theta$ if selection is acting. If there is strong overdominant selection, K may be larger than $\theta$ because of a very long coalescent time between two allelic classes (see Fig. 3 and Table 5). For example, $K\approx 1{\cdot}80\theta$ in the case of $n=10$, $i=5$, $Ns_1=Ns_2=1$, and $K\approx 386\theta$ if $Ns_1=Ns_2=10$. In such cases, $\theta$ estimated from K results in a considerable overestimation. On the other hand, if we can identify the selected nucleotide site, $K_1(i,n-i)+K_2(i,n-i)$ at the selected site is useful to estimate $\theta$. The variance of $K_1(i,n-i)+K_2(i,n-i)$ is similar to that of K under the neutral model. It decreases if strong overdominant selection is acting. Note that it is necessary to detect the selected site because our model assumes that selection acts at only one particular site and that mutations in the other sites are neutral. To detect the selected site, the average number of pairwise differences between two allelic classes can be used, because it is largely affected by selection. It can be concluded that $K_1(i,n-i)+K_2(i,n-i)$ gives a good estimate for $\theta$ rather than does K in a region where strong selection is acting and there is no recombination. However, the effect of recombination on $K_1(i,n-i)+K_2(i,n-i)$ is large, although recombination does not affect the expectation of K. $K_1(i,n-i)+K_2(i,n-i)$ is sensitive to recombination and greatly exceeds K in a region with a high recombination rate. This result may suggest that the bias in $K_1(i,n-i)+K_2(i,n-i)$ due to recombination may be larger than the bias in K due to selection if selection is weak. We can conclude that $K_1(i,n-i)+K_2(i,n-i)$ can be a good estimator of $\theta$ in some cases.

Our analytical result (see Appendix) is different from that of Kaplan *et al.* (1988), because of different assumptions. In their study, it is assumed that the frequency of the allelic class is constant at $x_0$, where $x_0$ is a deterministic equilibrium frequency of the allelic

class in the selection model. The coalescent event between two allelic classes is dependent on the recurrent mutations between two allelic classes. Accordingly, the coalescent time between two allelic classes is given as a function of mutation rate and $x_0$. In the present study, we assume that there is a particular nucleotide site that distinguishes two allelic classes. Since we follow the infinite site model, there is only one mutation at this site. Therefore, the mutation rates at this site are zero, since one mutation has already taken place. The formula for the coalescent time between two allelic classes obtained in this study does not involve the mutation rate. Also this formula is not a function of $x_0$, because we consider the equilibrium distribution of the frequency of the allelic class ($x$). It is more realistic because $i$ (number of A1 allelic class) depends on the frequency of this allelic class ($x$), and $x$ is usually unknown.

## Appendix

To derive the average number of pairwise differences between A1 and A2 allelic classes, we first consider the probability that A1 allelic class is the mutant allelic class, given the frequency of A1 allelic class. Denote this probability by $P_1(x)$, where $x$ is the freqency of A1 allelic class. Watterson (1977) demonstrated that $P_1(x)$ is the same as the probability of extinction of an allele when its frequency is $x$, and that $P_1(x)$ is given by

$$P_1(x) = \frac{\int_x^1 G(y)\,dy}{\int_0^1 G(y)\,dy}, \tag{A 1}$$

where

$$G(y) = \exp\{-4Nsy\} \tag{A 2}$$

under the genic selection model and

$$G(y) = \exp\{-2Ns_1 y^2 - 2Ns_2(1-y)^2\} \tag{A 3}$$

under the overdominant selection model, respectively (Kimura, 1962).

Second, we consider the age of A1 when A1 is mutant. Let $M_1(x)$ be the mean age of A1 allelic class when A1 is the mutant allelic class with frequency $x$. From equation (14) in Watterson (1977) (see also Maruyama, 1974; Li, 1975), $M_1(x)$ is given by

$$M_1(x) = 4N\int_0^1 G(y)\,dy\left\{\int_0^1 \frac{P_1(y)[1-P_1(y)]}{y(1-y)G(y)}dy - \int_x^1 \frac{P_1(y)[1-P_1(y)/P_1(x)]}{y(1-y)G(y)}dy\right\}, \tag{A 4}$$

which is equivalent to the mean extinction time of an allele with frequency $x$ (Kimura & Ohta, 1969).

Let $P_2(x)$ be the probability that A2 allelic class is the mutant allelic class and $M_2(x)$ be the mean age of A2 when A2 is mutant, given that the frequency of A1 is $x$. Apparently, $P_2(x) = 1-P_1(x)$. $M_2(x)$ can be given by sbstituting $s$ by $-s$ and $x$ by $1-x$ in (A 4) under the genic selection model. On the other hand, by exchanging $s_1$ and $s_2$ and substituting $x$ by $1-x$, $M_2(x)$ can be obtained from (A 4) under the overdominant selection model.

Next, we consider the mean age of the mutant allelic class. Denote by $T(x)$ the mean age of the mutant allelic class when the frequency of A1 is $x$. Then, since either A1 or A2 can be mutant, $T(x)$ is given as the mean of $M_1(x)$ and $M_2(x)$ weighted by $P_1(x)$ and $P_2(x)$, respectively. Namely, we have

$$T(x) = P_1(x)M_1(x) + P_2(x)M_2(x). \tag{A 5}$$

Let $T(i, n-i)$ be the mean age of the mutant allelic class in $A(i, n-i)$. $T(i, n-i)$ can be obtained as the average of $T(x)$ weighted by $F(x\,|\,i, n-i)$, the distribution of $x$ in $A(i, n-i)$. Namely,

$$T(i, n-i) = \int_0^1 F(x\,|\,i, n-i)\,T(x)\,dx. \tag{A 6}$$

We have $F(x\,|\,i, n-i)$ from the combination of Wright's allelic frequency distribution in the equilibrium population (Wright, 1931, 1937) and Ewens' sampling distribution (Ewens, 1972). In the genic selection model, the fitnesses of genotypes A1A1, A1A2 and A2A2 are given by $1+2s$, $1+s$ and 1, respectively. In equilibrium, the probability distribution of $x$ is given by

$$\Phi(x) = C\frac{\exp\{4Nsx\}}{x(1-x)}, \tag{A 7a}$$

where C is constant (Wright, 1931, 1937). In this formula, the mutation rates between A1 and A2 are zero. This is because we follow the infinite site model, where only one mutation is allowed at a nucleotide site. Since A1 and A2 allelic classes exist, the mutation has already taken place. Therefore, the mutation rates are zero in this case. In the same way, we have the probability distribution of $x$ in the overdominant selection model, where the fitnesses of genotypes A1A1, A1A2 and A2A2 are given by $1-s_1$, 1 and $1-s_2$, respectively. Namely,

$$\Phi(x) = C\frac{\exp\{-2Ns_1x^2 - 2Ns_2(1-x)^2\}}{x(1-x)}. \tag{A 7b}$$

Using $\Phi(x)$, we have $F(x\,|\,i, n-i)$, the conditional probability distribution of $x$ in $A(i, n-i)$, based on Ewens' sampling theory (Ewens, 1972). In the genic selection model,

$$F(x\,|\,i, n-i) = \frac{\binom{n}{i}x^i(1-x)^{n-i}\Phi(x)}{\int_0^1 \binom{n}{i}y^i(1-y)^{n-i}\Phi(y)\,dy} = \frac{x^{i-1}(1-x)^{n-i-1}\exp\{4Nsx\}}{\int_0^1 y^{i-1}(1-y)^{n-i-1}\exp\{4Nsy\}\,dy}, \tag{A 8a}$$

and, in the overdominant selection model,

$$\mathrm{F}(x\,|\,i,n-i)=\frac{x^{i-1}(1-x)^{n-i-1}\exp\{-2Ns_1x^2-2Ns_2(1-x)^2\}}{\int_0^1 y^{i-1}(1-y)^{n-i-1}\exp\{-2Ns_1y^2-2Ns_2(1-y)^2\}\,\mathrm{d}y}. \quad (\mathrm{A}\,8b)$$

It should be noted that (A 8*b*) is also applicable to one of the minority-advantage types of frequency-dependent selection model where the fitnesses of A1A1, A1A2 and A2A2 are given by $\{1-s_1x\}^2$, $\{1-s_1x\}\{1-s_2(1-x)\}$ and $\{1-s_2(1-x)\}^2$, respectively (Takahata & Nei, 1990; Denniston & Crow, 1990).

Finally, we have $\mathrm{D}(i,n-i)$, the expectation of the average number of pairwise differences between A1 and A2 allelic classes. Since the mean coalescent time between two sequences sampled from different allelic classes is $2N+\mathrm{T}(i,n-i)$, $\mathrm{D}(i,n-i)$ is given as

$$\mathrm{D}(i,n-i)=2\mu[2N+\mathrm{T}(i,n-i)]=[1+\mathrm{T}(i,n-i)/2N]\,\theta. \quad (\mathrm{A}\,9)$$

The authors thank two anonymous reviewers for their comments and suggestions. This work was supported in part by a grant-in-aid from the Ministry of Education, Science, Sports and Culture of Japan.

## References


Aguadé, M., Miyashita, N. & Langley, C. H. (1992). Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. *Genetics* **132**, 755–770.

Clark, A. G., Leicht, B. G. & Muse, S. V. (1996). Length variation and secondary structure of introns in the *Mlc1* gene in six species of *Drosophila*. *Molecular Biology and Evolution* **13**, 471–482.

Denniston, C. & Crow, J. F. (1990). Alternative fitness models with the same allele frequency dynamics. *Genetics* **125**, 201–205.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. New York: Columbia University Press.

Dobzhansky, T. (1970). *Genetics of the Evolutionary Process*. New York: Columbia University Press.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theoretical Population Biology* **17**, 37–50.

Harris, H. (1966). Enzyme polymorphisms in man. *Proceedings of the Royal Society of London, Series B* **164**, 298–310.

Hasson, E. & Eanes, W. F. (1996). Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. *Genetics* **144**, 1565–1575.

Hubby, J. L. & Lewontin, R. C. (1966). A molecular approach to the study of the genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* **54**, 577–594.

Hudson, R. R. (1983*a*). Testing the coalescent-rate neutral allele model with protein sequence data. *Evolution* **37**, 203–217.

Hudson, R. R. (1983*b*). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Innan, H. & Tajima, F. (1997). The amounts of nucleotide variation within and between allelic classes, and the reconstruction of common ancestral sequence. *Genetics* **147**, 1431–1444.

Kaplan, N. L., Darden, T. & Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics* **120**, 819–829.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903.

Kimura, M. (1980). Average time until fixation of a mutant allele in a finite population under continued pressure: studies by analytical, numerical, and pseudo-sampling methods. *Proceedings of the National Academy of Sciences of the USA* **77**, 522–526.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kimura, M. & Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**, 763–771.

Kimura, M. & Takahata, N. (1983). Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. *Proceedings of the National Academy of Sciences of the USA* **80**, 1048–1052.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27–43.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.

Lewontin, R. C. & Hubby, J. L. (1966). A molecular approach to the study of the genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**, 595–609.

Li, W.-H. (1975). The first arrival time and mean age of a deleterious mutant gene in a finite population. *American Journal of Human Genetics* **27**, 274–286.

Maruyama, T. (1974). The age of an allele in a finite population. *Genetical Research* **23**, 137–143.

Nachman, M. W., Boyer, S. N., Searle, J. B. & Aquadro, C. F. (1996). Mitochondrial DNA variation and the evolution of Robertsonian chromosomal races of house mice, *Mus domesticus*. *Genetics* **136**, 1105–1120.

Neuhauser, C. & Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.

Rand, D. M. & Kann, L. M. (1996). Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and human. *Molecular Biology and Evolution* **13**, 735–748.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Takahata, N. & Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967–978.

Watterson, W. A. (1975). On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology* **7**, 256–276.

Watterson, W. A. (1977). Reversibility and the age of an allele. II. Two-allele models, with selection and mutation. *Theoretical Population Biology* **12**, 179–196.

Wesley, C. S. & Eanes, W. F. (1994). Isolation and analysis of the breakpoint sequences of chromosome inversion *In*(*3L*)*Payne* in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the USA* **91**, 3132–3136.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1937). The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences of the USA* **23**, 307–320.