The Coalescent and Infinite-Site Model of a Small Multigene Family

Hideki Innan[1]

Department of Biological Science, University of Southern California, Los Angeles, California 90089-1340 and Human Genetics Center, School of Public Health, University of Texas Health Science Center, Houston, Texas 77030

Manuscript received August 11, 2002
Accepted for publication November 6, 2002

ABSTRACT

The infinite-site model of a small multigene family with two duplicated genes is studied. The expectations of the amounts of nucleotide variation within and between two genes and linkage disequilibrium are obtained, and a coalescent-based method for simulating patterns of polymorphism in a small multigene family is developed. The pattern of DNA variation is much more complicated than that in a single-copy gene, which can be simulated by the standard coalescent. Using the coalescent simulation of duplicated genes, the applicability of statistical tests of neutrality to multigene families is considered.

RECENT genomic data show that a substantial proportion of genes in the eukaryotic genome have been created by gene duplication, forming multigene families (Ohno 1970; Lynch and Conery 2000; Bailey *et al.* 2002). It is suggested that gene duplication plays an important role in genome evolution. To understand the evolutionary mechanism to generate and maintain multigene families, it is important to investigate the pattern of nucleotide polymorphism, in addition to phylogenetic and comparative genomic analysis.

The pattern of polymorphism in a multigene family is much more complicated than that in a single-copy gene, because duplicated genes do not likely evolve independently due to recurrent exchanges of genetic materials between genes (*i.e.*, concerted evolution of multigene families, reviewed in Arnheim 1983). Gene conversion is considered to be the most important mechanism for the concerted evolution of small multigene families. Consider a multigene family with two duplicated genes. Gene conversion transfers DNA segments between the two genes, so that it creates sites that are polymorphic in both genes. Therefore, to analyze DNA polymorphism, it is reasonable to make a parallel alignment table of the duplicated genes. An example is shown in Table 1: the alignment of two genes, I and II, for $n = 5$ chromosomes. There are seven polymorphic sites, which are classified into three types: (1) specific polymorphic sites, at which polymorphism is observed in either of the two genes; (2) shared polymorphic sites, at which polymorphism is shared by the two genes; and (3) fixed polymorphic sites, at which each gene has a different fixed nucleotide. The second type of polymorphic sites (shared polymorphic sites) could be evidence for gene conversion when mutation rate per site is low. A number of shared polymorphic sites are observed in multigene families (*e.g.*, Inomata *et al.* 1995; King 1998; Bettencourt and Feder 2002).

There are not many theories for analyzing this complicated pattern of DNA polymorphism in a multigene family. In the 1980s, Ohta (1981, 1982, 1983), Nagylaki (1984a,b), and others considered the identity coefficients between pairs of genes in a multigene family. Several authors applied the coalescent to multigene families (Griffiths and Watterson 1990; Hey 1991; Bahlo 1998), but their results are not directly related to the analysis of the pattern of DNA polymorphism. I have recently obtained the expectations of the amounts of DNA variation in a two-locus multigene family (Innan 2002), but their variances and distribution are unknown. In this article, a coalescent simulation method for a small multigene family is developed to investigate the pattern of nucleotide variation. The simulation is based on the infinite-site model (Kimura 1969), which assumes that the mutation rate is so small that each polymorphism is produced by a single mutation. That is, shared polymorphic sites can be created only by gene conversion, not by independent mutations at corresponding sites in both genes. With the simulation, the frequency distributions of the three types of polymorphic sites are investigated, and I consider the applicability of standard statistical tests of neutrality to multigene families.

INFINITE-SITE MODEL FOR A SMALL MULTIGENE FAMILY

In this section, my previous theoretical result based on a two-locus gene conversion model (Innan 2002) is reviewed, and then I consider its extension to the infinite-site model of a multigene family with two copies of genes. Consider two linked loci, I and II, in a random-mating population with N diploids. The two loci were

[1]*Address for correspondence:* Human Genetics Center, School of Public Health, University of Texas Health Science Center, 1200 Hermann Pressler, Houston, TX 77030. E-mail: hinnan@sph.uth.tmc.edu

created by a gene duplication event, which occurred a very long time ago so that the population is at equilibrium. At each site, consider two neutral alleles, *A* and *a*, and therefore there are four haplotypes, *A-A*, *A-a*, *a-A*, and *a-a* (the first letter represents the allele at locus I and the second one represents the allele at locus II). It is assumed that the symmetric mutation rate between two alleles is μ per locus per generation. The recombination rate between two loci is assumed to be *r* per generation. Intrachromosomal gene conversion occurs at the rate *c* per locus per generation; *e.g.*, *A-a* changes into *A-A* with probability *c* and into *a-a* with the same probability. In this section, interchromosomal gene conversion is not considered for mathematical simplicity (this assumption is relaxed in the DISCUSSION).

Let the frequencies of *A-A*, *A-a*, *a-A*, and *a-a* be x_1, x_2, x_3, and x_4 ($x_1 + x_2 + x_3 + x_4 = 1$), respectively. The amount of variation within a locus, h_w, is defined as heterozygosity within a particular locus [*i.e.*, $h_w = 2(x_1 + x_2)(x_3 + x_4)$ at locus I, $h_w = 2(x_1 + x_3)(x_2 + x_4)$ at locus II]. The expectation of h_w at equilibrium is given by a function of three parameters ($\theta = 4N\mu$, $C = 4Nc$, and $R = 4Nr$),

$$E(h_w) = 1 - 2\frac{\lambda}{\omega}, \tag{1}$$

where

$$\alpha = 2\theta + C, \quad \beta = 2 + 2\alpha + R,$$
$$\lambda = 4C^2 + \beta[2\theta C + 2\alpha(1 + \theta)],$$
$$\omega = 8C^2 + 4\beta[\alpha(1 + \alpha) - C^2]$$

(INNAN 2002) when $\theta \neq 0$ and $C \neq 0$. The amount of variation between two loci, h_b, is defined as the probability that two independent alleles sampled from different loci are different [*i.e.*, $h_b = (x_1 + x_2)(x_2 + x_4) + (x_1 + x_3)(x_3 + x_4)$]. The expectation of h_b is given by

$$E(h_b) = 1 + \frac{1 + \theta}{C} - \frac{2(1 + \alpha)\lambda}{C\omega}. \tag{2}$$

The expectation of linkage disequilibrium between two loci ($D = x_1x_4 - x_2x_3$) is given by

$$E(D) = \frac{C}{\beta}\left(1 - \frac{2\lambda}{\omega}\right). \tag{3}$$

Here, we consider h_w, h_b, and *D* in a small multigene family with two duplicated genes, I and II, each of which consists of *L* nucleotides. Assume that *n* chromosomes are randomly sampled from a population and both genes are sequenced for each chromosome. The amount of nucleotide variation within a gene is usually measured by the average number of pairwise differences, π_w. Denote the numbers of nucleotide differences between the *i*th and *j*th chromosomes in the first and second genes by $d_{11}(i, j)$ and $d_{22}(i, j)$, respectively. Then, π_w for genes I and II are given by

TABLE 1

Example of a parallel alignment table

Gene	Chromosome	1	2	3	4	5	6	7
I	1	A	T	G	T	C	C	A
I	2	A	T	G	C	C	C	A
I	3	C	T	G	C	C	A	A
I	4	C	G	A	T	C	C	A
I	5	A	G	G	T	C	C	A
II	1	A	T	G	C	C	C	G
II	2	C	G	G	C	C	C	G
II	3	C	G	G	C	G	C	G
II	4	A	G	G	C	C	C	G
II	5	A	T	G	T	G	C	G
Type of polymorphism		S	S	I	S	II	I	F

I, specific to gene I; II, specific to gene II; S, shared polymorphism; F, fixed polymorphism.

$$\pi_{w1} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{11}(i, j)$$
$$\pi_{w2} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{22}(i, j), \tag{4}$$

respectively. $\pi_{w1} = 2.6$ and $\pi_{w1} = 2.4$ in the example of Table 1. Let $d_{12}(i, j)$ be the number of nucleotide differences between gene I of the *i*th chromosome and gene II of the *j*th chromosome. The average of $d_{12}(i, j)$ represents the amount of variation between two genes. That is,

$$\pi_b = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} d_{12}(i, j). \tag{5}$$

Note that π_b is defined to correspond to h_b derived by INNAN (2002) so that π_b does not involve $d_{12}(i, j)$ when $i = j$ (*i.e.*, π_b does not consider the nucleotide differences between two genes on the same chromosome). This is because h_b is defined as the probability that two *independent* alleles sampled from different loci are different. In the data of Table 1, $\pi_b = 3.4$. Define D_{sum} as the sum of linkage disequilibria at all *L* sites. Let D_m be the linkage disequilibrium at the *m*th site, which is calculated as $D_m = (n_{AA}n_{aa} - n_{Aa}n_{aA})/[n(n-1)]$, where n_{xy} represents the number of chromosomes with nucleotides *x* and *y* at genes I and II, respectively. Then, D_{sum} is given by

$$D_{sum} = \sum_{m=1}^{L} D_m. \tag{6}$$

In the data of Table 1, since $D_1 = 0.05$, $D_2 = -0.05$, and $D_4 = 0.1$, the sum is $D_{sum} = 0.1$. Note that only shared polymorphic sites contribute linkage disequilibrium ($D = 0$ for the other types of polymorphic sites).

Equations 1–3 are applied to this two-gene model with *L* nucleotides. Since it is possible to consider that there are *L* two-locus models in the duplicated genes,

the expectations of three amounts of variation are given by

$$E(\pi_{w1})/L = E(\pi_{w2})/L = E(h_w), \quad E(\pi_b)/L = E(h_b),$$
$$E(D_{sum})/L = E(D). \tag{7}$$

When gene conversion occurs between a pair of DNA sequences, it should be considered that gene conversion involves a certain length of DNA tract, indicating that L nucleotide sites in the duplicated genes are not independent. However, these equations for the expectations hold without the assumption of independence among L sites. That is, the distribution of gene conversion tract does not affect the expectations if the gene conversion rate per site (C) is given. On the other hand, the variances of π_w, π_b, and D_{sum} are affected by the distribution of gene conversion tract.

Under the infinite-site model, the mutation rate is assumed to be so small that there are no multiple mutations at a single site (Kimura 1969). With this assumption, the expected amounts of variation are obtained from (7) by letting $L \to \infty$ with $L\theta = \Theta$. That is,

$$E(\pi_w) = \frac{2\Theta(2C + R + 2)}{4C + R + 2}, \tag{8}$$

$$E(\pi_b) = \frac{\Theta(4C^2 + 4C + 2CR + R + 2)}{C(4C + R + 2)}, \tag{9}$$

and

$$E(D_{sum}) = \frac{2\Theta C}{4C + R + 2}. \tag{10}$$

From (8–10), Θ, C, and R can be estimated by π_w, π_b, and D_{sum}:

$$\hat{\Theta} = \frac{\pi_w + 2D_{sum}}{2}, \tag{11}$$

$$\hat{C} = \frac{\pi_w - 2D_{sum}}{2(\pi_b - \pi_w)}, \tag{12}$$

and

$$\hat{R} = \frac{\pi_w^2 + 4D_{sum}^2 - 4\pi_b D_{sum}}{2(\pi_b - \pi_w)D_{sum}}. \tag{13}$$

With the example data of Table 1, Θ, C, and R are estimated to be 1.3, 1.1, and 22.2, given $\pi_w = 2.4$, $\pi_b = 3.4$, and $D_{sum} = 0.1$.

Equations 11–13 are also applied to data of three small multigene families in *Drosophila melanogaster*. As shown in Table 2, these equations work when $\pi_w < \pi_b$ and $D_{sum} > 0$. Equation 12 does not work well when $\pi_w > \pi_b$ because (8) and (9) indicate $E(\pi_w) \leq E(\pi_b)$ [$E(\pi_w) = E(\pi_b)$ when $C = \infty$]. Equation 13 also does not work well when $D_{sum} < 0$ because the theory predicts $E(D_{sum}) \geq 0$. See Innan (2002) for another method for estimating these population parameters.

TABLE 2

Application of Equations 11–13 to three multigene families of *D. melanogaster*

	Observation			Estimate		
	π_w	π_b	D_{sum}	Θ	C	R
Amylase[a]	20.40	22.04	2.72	12.92	4.55	22.99
Attacin[b]	8.93	31.41	−0.03	4.47	0.20	NA[c]
Hsp70[d]	6.41	6.38	1.69	4.90	NA[e]	NA

[a] Data of the distal and proximal amylase genes in the Kenyan sample ($n = 10$) from Araki *et al.* (2001).

[b] Data of the *AttacinA* and *-B* genes ($n = 11$) from Lazzaro and Clark (2001). The haplotype with a big deletion (2CPA 43) is excluded.

[c] An estimate of $R = \infty$ according to Innan (2002).

[d] Data of the *Hsp*70 *Aa* and *Ab* gene ($n = 11$) from Bettencourt and Feder (2002).

[e] An estimate of $C = \infty$ according to Innan (2002).

COALESCENT SIMULATION OF A SMALL MULTIGENE FAMILY

To simulate patterns of polymorphism in a small multigene family with two duplicated genes, a standard coalescent model with recombination (Hudson 1983) is modified. Assume that the number of genes is constant at two for a very long time. For simplicity, it is also assumed that recombination occurs only between two genes, although intragenic recombination is easily incorporated (*e.g.*, see Nordborg 2001). Figure 1A illustrates an example of the ancestral recombination graph of a pair of duplicated genes for $n = 3$, which is generated backward in time. Following the standard two-locus coalescent (Hudson 1983), a pair of chromosomes coalesce with probability $1/2N$ per generation, and a chromosome splits into two by recombination with probability r per generation. Two modifications are needed to simulate the pattern of polymorphism in duplicated genes. First, genealogical information for lineages that are not ancestors of the sampled chromosomes is needed. Such lineages that are not needed in a standard coalescent simulation of a single-copy gene are represented by dashed lines in Figure 1A. Second, the coalescence and recombination process cannot stop when all sampled chromosomes reach their most recent common ancestor (MRCA). That is, the simulation should be continued until the MRCA of the two genes (see below).

On the way to generate the ancestral recombination graph, gene conversions are placed randomly (Figure 1A). Gene conversion occurs with probability c per site per generation whether lineages are ancestral to the sampled chromosomes (Figure 1A, solid lines) or not (Figure 1A, dashed lines). For each gene conversion event, the position and direction are determined. For convenience, the gene is represented by an interval of (0, 1), so that the position of a gene conversion tract

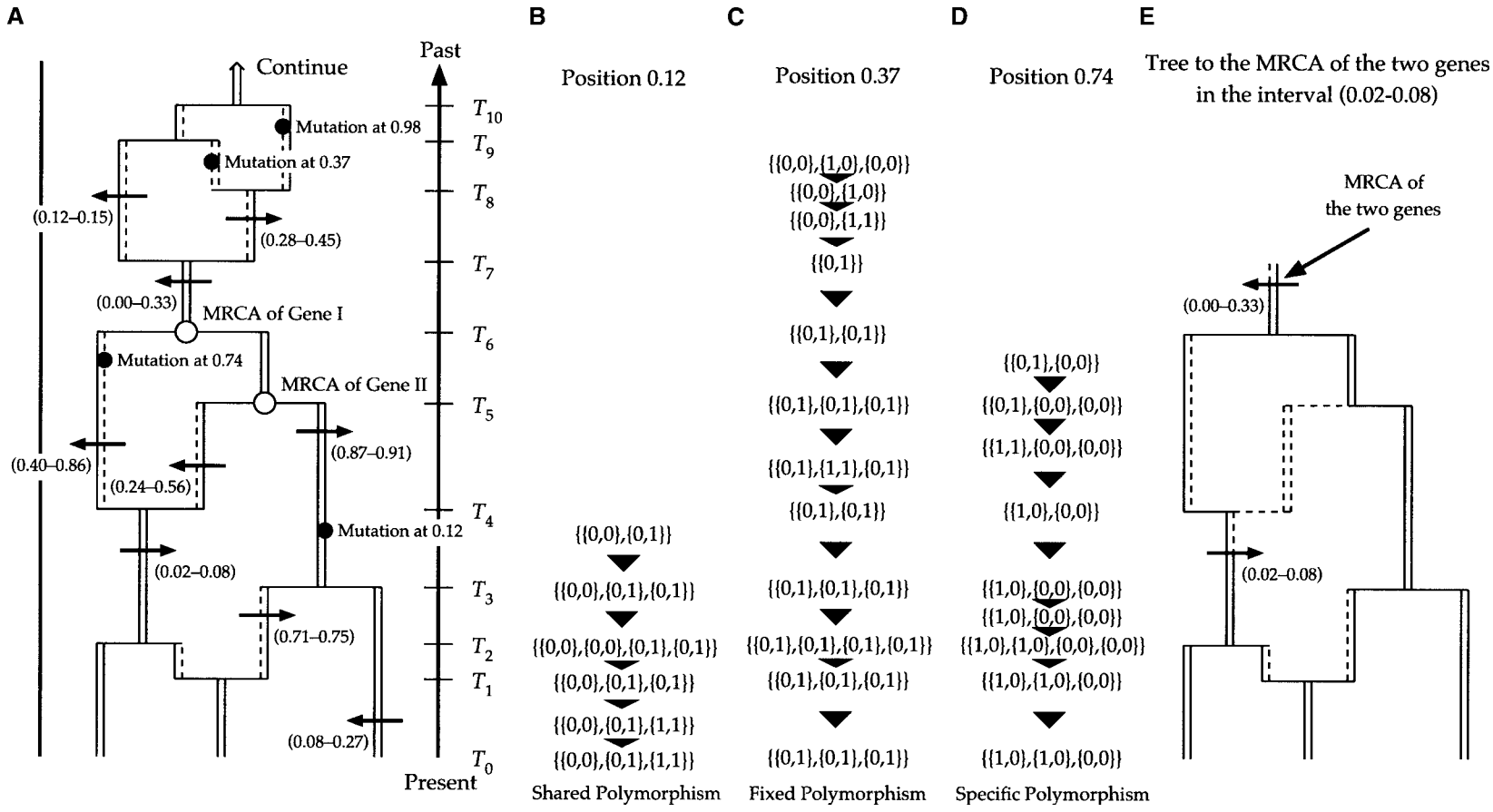

FIGURE 1.—(A) An example of the ancestral recombination graph for two duplicated genes, I and II. The lineages are represented by double lines. For a given pair of lines, the left shows gene I and the right shows gene II. The two open circles represent the MRCAs for the two genes. (B–D) Changes of allelic states. The histories of the mutations at positions 0.12, 0.37, and 0.74 are traced forward in time along the ancestral recombination graph. (E) Tree to the MRCA of the two genes in the interval (0.02–0.08). The solid lines represent the ancestral lineages of the sampled genes in the interval.

is given by an interval between 0 and 1. For example, the gene conversion between T_0 and T_1 in Figure 1A occurs between positions 0.08 and 0.27. Since the direction of this gene conversion is from II to I, the gene conversion changes allelic state {1, 0} to {0, 0} and {0, 1} to {1, 1} (see Figure 1, B–D). Note that the allelic state for a pair of lineages is represented by two numbers in brackets. The presence and absence of mutation are represented by 1 and 0, respectively. The first number is for gene I and the second one is for gene II. A gene conversion of the other direction changes {1, 0} to {1, 1} and {0, 1} to {0, 0}. Gene conversions do not change the allelic states {0, 0} or {1, 1}. The length of gene conversion tract might follow a certain function. Wiuf and Hein (2000) used a geometric distribution for homologous gene conversion, that is, gene conversion between copies of the same locus (gene conversion considered here is nonhomologous).

This two-gene coalescent simulation should be continued until the MRCA of the two genes (*i.e.*, the MRCA of all the $2n$ lineages) is reached. The MRCA of the two genes requires coalescence between the two genes, which occurs by gene conversion because gene conversion transfers the DNA segment from one gene to the other. Figure 1E shows the tree for the interval (0.02–0.08), which is used to explain the definition of the MRCA of the two genes. On the tree, a gene conversion event occurs between T_3 and T_4 and transfers the DNA segment between 0.02 and 0.08 of gene I to gene II. This event can be considered as a coalescent event between the two genes. That is, going backward in time, the right lineage merges into the left one. Treating gene conversion in this way, we can find the MRCA of the two genes when the $2n$ lineages coalesce into one lineage. On the tree in Figure 1E, it occurs with the gene conversion event between T_6 and T_7. The coalescent simulation can be stopped when all segments in the interval (0, 1) reach the MRCAs of the two genes.

Given an ancestral recombination graph with gene conversion, mutations are randomly distributed on lineages following the Poisson process (Figure 1A). Mutations occur at any position on the graph with equal probability density (μ per site per generation) whether lineages are ancestral to the sampled chromosomes or not. For each mutation, the position in the gene is also determined. The positions are random numbers between 0 and 1. In Figure 1A, there are four mutations: at position 0.12 of gene II, at position 0.37 of gene I, at position 0.74 of gene II, and at position 0.98 of gene I. The allelic state of the lineage on which mutation occurs is given by 1. For example, when the mutation at position 0.12 occurs in gene II between T_3 and T_4, the allelic state of the site for the two genes is given by {0, 1} (Figure 1B).

The histories of the mutations in Figure 1A are traced forward in time in Figure 1, B–D, where allelic states are shown along the ancestral recombination graph (Figure 1A). Let us follow the mutation at position 0.12. Since the mutation occurs in gene II on the right pair of lineages between T_3 and T_4, the allelic states of the two pairs of lineages are given by {{0, 0}, {0, 1}} (the order of allelic states follows Figure 1A). At T_3 the right pair of lineages are duplicated (coalescent event), and the states for the three pairs of lineages are given by {{0, 0}, {0, 1}, {0, 1}}. Another duplication of the left pair of lineages at T_2 results in {{0, 0}, {0, 0}, {0, 1}, {0, 1}}, and a recombination event with the two middle pairs of lineages at T_1 makes {{0, 0}, {0, 1}, {0, 1}}. Between T_0 and T_1, a gene conversion event on the right pair of lineages results in {{0, 0}, {0, 1}, {1, 1}} at the bottom of the graph. Therefore, the mutation at 0.12 appears as a shared polymorphic site. In a similar way, the mutations at 0.37 and 0.74 are traced and appear as fixed and specific polymorphisms, respectively (Figure 1, C and D). Note the mutation at 0.98 is not observed because it is lost by the recombination event at T_8.

Following this process, patterns of DNA polymorphism are simulated and frequency spectra of three types of polymorphisms are investigated. For each parameter set, the expected frequency spectrum is obtained from 10,000 replications. The length of gene conversion tract is assumed to be so small that any gene conversion segment does not include more than one mutation. This assumption does not affect the expected spectrum as long as the gene conversion rate per site is constant as mentioned in the previous section. It is demonstrated that the averages of π_w, π_b, and D_{sum} in the simulations are in excellent agreement with the theoretical expectations obtained by (8–10).

Figure 2A shows the spectra of derived alleles (nucleotides) for a low gene conversion rate ($C = 0.2$). It is shown that a large proportion of polymorphic sites are fixed sites. Specific polymorphic sites are more frequent than shared polymorphic sites, and the shapes of spectra of these two types of polymorphic sites are ∪ shapes that are skewed toward the left (rare classes). The effect of recombination on the spectrum is relatively small. When $C = 1$ (Figure 2B), shared polymorphic sites are more frequent than specific ones, and fixed ones are very rare. The spectra of specific and shared sites are both ∟ shapes, and the former is more skewed than the latter. When gene conversion rate is high ($C = 5$), almost no fixed polymorphic sites are observed, and most polymorphic sites are shared sites (Figure 2C). Figure 3A shows the observed spectra in the distal and proximal *Amy* genes of *D. melanogaster*. They are similar to the expected spectrum obtained from a simulation with 10,000 replications given the estimated values of $\Theta = 12.92$, $C = 4.55$, and $R = 22.99$ (see Table 2).

APPLICABILITY OF TESTS OF NEUTRALITY

As demonstrated in this article, the pattern of polymorphism in a multigene family is much more complicated than that in a single-copy gene. Therefore, statistical tests of neutrality based on the standard coalescent

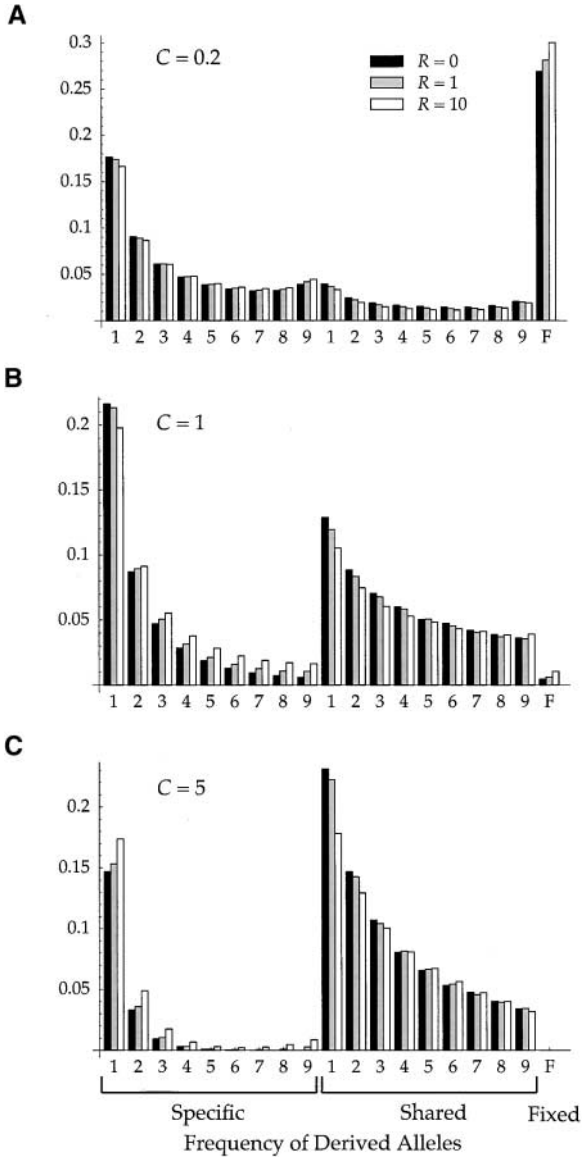

FIGURE 2.—Expected spectra of three types of polymorphic sites in a small multigene family. Simulations were carried out with $n = 10$ and $\Theta = 10$, although Θ does not affect the expected spectrum.

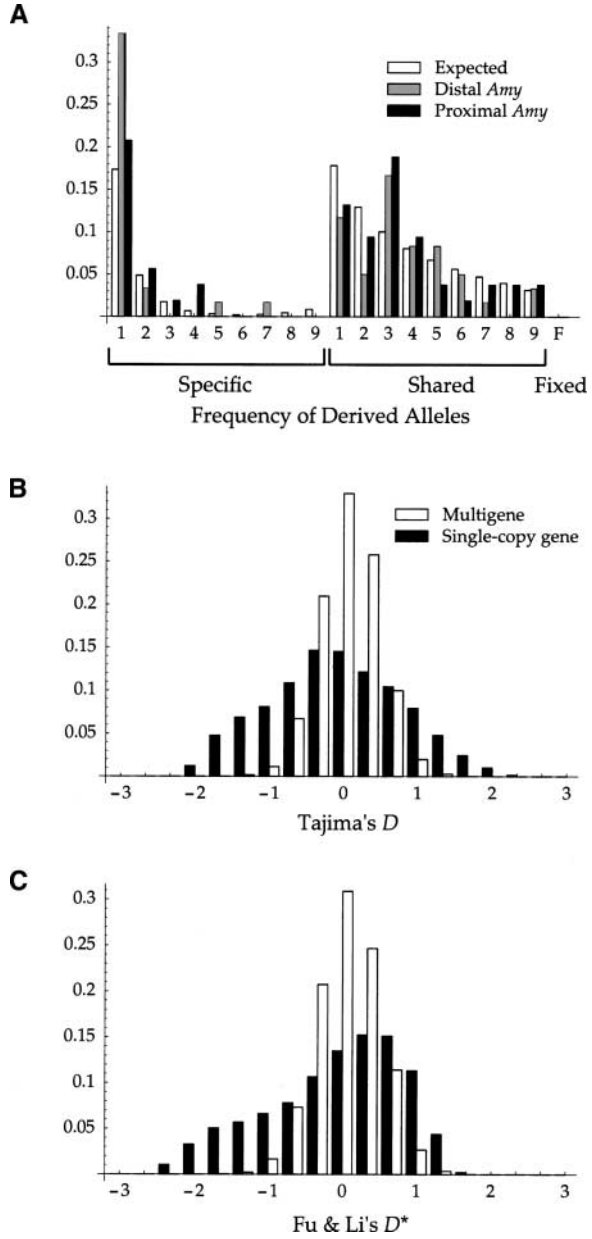

FIGURE 3.—(A) Observed spectra of three types of polymorphic sites in the distal and proximal Amylase genes with the expectations when $\Theta = 12.92$, $C = 4.55$, and $R = 22.99$. (B) Distribution of Tajima's D in a single-copy gene and in a gene of a two-copy multigene family. (C) Distribution of Fu and Li's D^* in a single-copy gene and in a gene of a two-copy multigene family.

theory for a single-copy gene may not be appropriate for genes in multigene families. TAJIMA's (1989) D and FU and LI's (1993) D^* tests are among these. Consider the distal and proximal *Amy* genes in *D. melanogaster* as examples. If the two genes are treated as two independent single-copy genes, the test statistics can be calculated for each gene. Tajima's D and Fu and Li's D^* are -0.13 and -0.38 in the distal gene and 0.10 and 0.09 in the proximal gene, respectively. However, the distributions of the test statistics for multigenes are different from those for single-copy genes. In Figure 3B, the distribution of Tajima's D in a single-copy gene is compared with that for a gene in a small multigene family with $\Theta = 12.92$, $C = 4.55$, and $R = 22.99$. The variance of the latter is much smaller than that of the former, indicating it is very unlikely to observe significant Tajima's D in a small multigene family if the confidence interval is determined by the distribution in a single-copy gene. A similar result is obtained for Fu and Li's D^* (Figure 3C). The results are consistent with the observed Tajima's D and Fu and Li's D^* values, which are quite close to zero. Hudson, Kreitman, and Aguadé's test (HUDSON *et al.* 1987) also cannot be used for multigene families, because the expected amount of variation within species in a duplicated gene is more than expected in a single-copy gene (see Equation 8).

On the other hand, there is no problem in applying

TABLE 3

Summary of polymorphic sites in three multigene families of *D. melanogaster*

	Nonsynonymous	Synonymous	Total
Amylase			
Specific to distal	7	18	25
Specific to proximal	4	13	17
Shared	7	30	37
Fixed	0	0	0
Attacin			
Specific to *A*	11	36	47
Specific to *B*	2	5	7
Shared	0	10	10
Fixed	6	11	17
Hsp70			
Specific to *Aa*	5	15	20
Specific to *Ab*	5	5	10
Shared	1	11	12
Fixed	0	0	0

model-independent tests of neutrality. McDonald and Kreitman (1991) developed a simple statistic test based on a comparison of the ratio of the number of replacement substitutions to the number of synonymous substitutions. They compared the ratio between polymorphic sites and fixed sites between species. This kind of test can be used for multigene families. For example, the ratio can be compared among the three types of polymorphic sites in multigene families defined in Table 1. Table 3 summarizes the numbers of replacement and synonymous polymorphic sites in three multigene families in *D. melanogaster*. No pair of comparisons is significant at the 5% level by Fisher's exact test, although the ratio of replacement sites to synonymous sites in the shared class tends to be smaller than that in the other classes.

DISCUSSION

The pattern of nucleotide polymorphism in a multigene family is much more complicated than that in a single-copy gene because of exchanges of genetic materials between members of a family. In this article, the amounts and pattern of nucleotide polymorphism are studied under the infinite-site model. The expectations of three amounts of DNA variation (π_w, π_b, and D_{sum}) are obtained analytically, and a coalescent method for simulating patterns of nucleotide polymorphism is developed. From the simulation the frequency spectra of three types of polymorphic sites are investigated.

The simulations demonstrate that statistical tests that are based on the standard theory for a single-copy gene may not be appropriate to use for genes in multigene families (*e.g.*, Tajima's *D*; Fu and Li's *D**; and Hudson, Kreitman, and Aguadé's tests). New statistical tests should be developed for multigene families with the coalescent simulation described in this article. On the other hand, model-independent tests (*e.g.*, McDonald and Kreitman's test) can be used without any problem (see Table 3).

The coalescent simulation developed in this article can be easily extended to a model of a multigene family with more than two genes as long as the number of genes is constant. Patterns of polymorphism in such multigene families could be more complicated because the gene conversion rates among members may vary. An example is seen in the *hsp*70 multigene family (Bettencourt and Feder 2002), which consists of five genes, *hsp*70Aa, *hsp*70Ab, *hsp*70Ba, *hsp*70Bb, and *hsp*70Bc. Gene conversion might be frequent between *hsp*70Aa and *hsp*70Ab, between *hsp*70Ba and *hsp*70Bb, and between *hsp*70Bb and *hsp*70Bc, while the gene conversion rates between the other pairs may be quite low. There are too few data of multigene families to understand the mechanism that determines the gene conversion rate.

Interchromosomal gene conversion, which is ignored for mathematical convenience, can be easily incorporated in the simulation, because an interchromosomal gene conversion event can be considered as intragenic gene conversion and recombination events that occur at the same time. That is, going backward in time, immediately after placing an intragenic gene conversion event, a new pair of lineages is introduced in the ancestral recombination graph. It is not clearly understood how often interchromosomal gene conversion occurs in comparison with intrachromosomal gene conversion.

The author thanks H. Araki, J. Hey, M. Nordborg, and N. Rosenberg for comments and discussions, and the two anonymous reviewers for helpful suggestions. The C-program used in this study is available on request by the author.

LITERATURE CITED

Araki, H., N. Inomata and T. Yamazaki, 2001 Molecular evolution of duplicated amylase gene regions in *Drosophila melanogaster*: evidence of positive selection in the coding regions and selective constraints in the *cis*-regulatory regions. Genetics **157:** 667–677.

Arnheim, N., 1983 Concerted evolution of multigene families, pp. 38–61 in *Evolution of Genes and Proteins*, edited by M. Nei and R. K. Koehn. Sinauer, Sunderland, MA.

Bahlo, M., 1998 Segregating sites in a gene conversion model with mutation. Theor. Popul. Biol. **54:** 243–256.

Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte *et al.*, 2002 Recent segmental duplications in the human genome. Science **297:** 1003–1007.

Bettencourt, B. R., and M. E. Feder, 2002 Rapid concerted evolution via gene conversion at the Drosophila *hsp70* genes. J. Mol. Evol. **54:** 569–586.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Griffiths, R. C., and G. A. Watterson, 1990 The number of alleles in multigene families. Theor. Popul. Biol. **37:** 110–123.

Hey, J., 1991 A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. Theor. Popul. Biol. **39:** 30–48.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Innan, H., 2002 A method for estimating the mutation, gene conver-

sion and recombination parameters in small multigene families. Genetics **161:** 865–872.

Inomata, N., H. Shibata, E. Okuyama and T. Yamazaki, 1995 Evolutionary relationships and sequence variation of α-amylase variants encoded by duplicated genes in the *Amy* locus of *Drosophila melanogaster*. Genetics **141:** 237–244.

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

King, L. M., 1998 The role of gene conversion in determining sequence variation and divergence in the *Est-5* gene family in *Drosophila pseudoobscura*. Genetics **148:** 305–315.

Lazzaro, B. P., and A. G. Clark, 2001 Evidence for recent paralogous gene conversion and exceptional allelic divergence in the *Attacin* genes of *Drosophila melanogaster*. Genetics **159:** 659–671.

Lynch, M., and J. S. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science **290:** 1151–1155.

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

Nagylaki, T., 1984a Evolution of multigene families under interchromosomal gene conversion. Proc. Natl. Acad. Sci. USA **81:** 3796–3800.

Nagylaki, T., 1984b The evolution of multigene families under intrachromosomal gene conversion. Genetics **106:** 529–548.

Nordborg, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. J. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, New York.

Ohta, T., 1981 Genetic variation in small multigene families. Genet. Res. **37:** 133–149.

Ohta, T., 1982 Allelic and nonallelic homology of a supergene family. Proc. Natl. Acad. Sci. USA **79:** 3251–3254.

Ohta, T., 1983 On the evolution of multigene families. Theor. Popul. Biol. **23:** 216–240.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis. Genetics **123:** 585–595.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Communicating editor: J. Hey