DOI: 10.1534/genetics.106.063131

On the Genealogy of a Duplicated Microsatellite

Kangyu Zhang* and Noah A. Rosenberg†,1

**Program in Molecular and Computational Biology and Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113 and †Department of Human Genetics, Center for Computational Medicine and Biology, and the Life Sciences Institute, University of Michigan, Ann Arbor, Michigan 48109-2218*

Manuscript received August 18, 2006
Accepted for publication September 25, 2007

ABSTRACT

When a microsatellite locus is duplicated in a diploid organism, a single pair of PCR primers may amplify as many as four distinct alleles. To study the evolution of a duplicated microsatellite, we consider a coalescent model with symmetric stepwise mutation. Conditional on the time of duplication and a mutation rate, both in a model of completely unlinked loci and in a model of completely linked loci, we compute the probabilities for a sampled diploid individual to amplify one, two, three, or four distinct alleles with one pair of microsatellite PCR primers. These probabilities are then studied to examine the nature of their dependence on the duplication time and the mutation rate. The mutation rate is observed to have a stronger effect than the duplication time on the four probabilities, and the unlinked and linked cases are seen to behave similarly. Our results can be useful for helping to interpret genetic variation at microsatellite loci in species with a very recent history of gene and genome duplication.

GENE and genome duplications are important mechanisms for evolving genetic novelty (Ohno 1970; Lynch and Conery 2000; Zhang 2003). This fundamental role of duplication in the evolutionary process has led to the development of a variety of population-genetic models that utilize genetic polymorphisms in duplicated genes for understanding the evolutionary histories of duplicated gene families (Walsh 2003; Innan 2004). For example, these models have been used to study the process by which "concerted evolution" through gene conversion influences the similarities and differences among recently duplicated genes (Innan 2002, 2003; Teshima and Innan 2004), as well as the way in which "subfunctionalization" through specialization of duplicate copies of a gene after duplication can lead to preservation in a genome of two or more paralogs (Lynch and Force 2000; Ward and Durrett 2004).

Microsatellites, short and tandemly repetitive sequences that are widely dispersed in a variety of genomes, are among the most important genetic markers used by population and evolutionary biologists, due to their high variability in copy number (Ellegren 2004). Microsatellite variation has long been studied using stepwise mutation models (Goldstein *et al.* 1995; Slatkin 1995; Zhivotovsky and Feldman 1995), which specify the probabilities of various types of change in copy number for the basic repeated unit (Ohta and Kimura 1973; Calabrese and Sainudiin 2005). The simplest of these models assumes that the mutation rate is independent of copy number, that mutations alter copy number only by one unit, and that increases and decreases in copy number are equally likely. A variety of properties of microsatellite loci evolving according to this symmetric stepwise mutation model have been investigated, including the pairwise allele size difference between two randomly chosen alleles (Pritchard and Feldman 1996; Nielsen 1997; Blum *et al.* 2004), the within-population variance of the allele size distribution (Zhivotovsky and Feldman 1995; Blum *et al.* 2004), the expected homozygosity (Ohta and Kimura 1973; Kimmel and Chakraborty 1996; Pritchard and Feldman 1996; Nielsen 1997), and the probability that two alleles identical in size are identical by descent (Estoup *et al.* 2002).

Duplicated microsatellites have been known since shortly after the discovery of microsatellite loci, and they have been useful particularly in Y-chromosomal studies in humans (Mathias *et al.* 1994; Balaresque *et al.* 2007). However, despite the extensive history of studies both of gene duplication and of microsatellites as genetic markers, only recently have these two topics been combined in studies that use microsatellites as a tool for investigating genetic duplication (David *et al.* 2003; Antunes *et al.* 2006). Microsatellites, due to their large amount of variability, have the potential to be informative about duplication. As microsatellites are typically genotyped using PCR primers that amplify all segments of a genome that have the primers as flanking regions, PCR primers applied in a diploid individual to a microsatellite that lies in a duplicated autosomal region of the genome will amplify four genomic fragments. Because of the rapid changes that occur in microsatellite copy numbers

[1]*Corresponding author:* Department of Human Genetics, Center for Computational Medicine and Biology, and the Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109-2218.
E-mail: rnoah@umich.edu

over time, these four fragments may contain as many as four distinct copy number variants. Thus, the occurrence of three or four distinct alleles at a microsatellite locus, given a single set of PCR primers in a single diploid individual, can provide information about recent duplication of the region containing the microsatellite. This phenomenon suggests the potential of microsatellites as selectively neutral markers for studying gene and genome duplication.

To understand the effects that duplication may have on microsatellite variation, in this article, we develop a simple coalescent model that we use to explore the properties of duplicated microsatellites. Under the symmetric stepwise mutation model, we deduce the probability distribution of the number of distinct alleles that will be observed in one diploid individual at a duplicated microsatellite, conditional on a known genealogy, duplication time, and mutation rate, assuming that the duplication event predates the most recent common ancestor for the two allelic copies of each paralog. Randomness in the genealogy is then incorporated into the analysis by considering the distribution of genealogies both under a model in which the two paralogs are completely linked and using one in which they are completely unlinked. The effects of the model parameters—the duplication time and the mutation rate—are then evaluated using both numerical computation and a simulation-based approach.

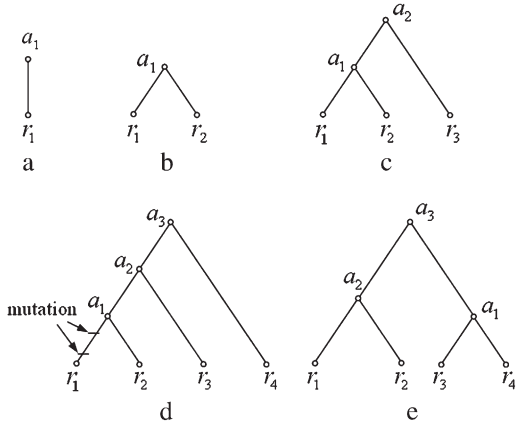

FIGURE 1.—Notation for trees. Trees with sample sizes ranging from 1 to 4 are shown. When the sample size is 1, 2, or 3, only one unlabeled, rooted, binary topology is possible, as shown in a–c. If the sample size is 4, two tree topologies are possible (d and e). Two downward mutations are indicated in d.

THEORY

Definitions: Several concepts used in our analysis are illustrated in Figure 1, and notation is described in Table 1. For a microsatellite locus and a node of a genealogical tree, we refer to the number of copies of the repeated unit at that node as the "allele state" of the node. Terminal nodes are labeled from left to right as r_i and internal nodes are labeled as a_i, ordered sequentially on the basis of their relative time back from the present. If $i < j$, then $a_i.t < a_j.t$; if $a_i.t = a_j.t$, then i and j are annotated according to their relative position from the left side of the diagram. As an example of the notation, consider Figure 1d, which has four terminal nodes r_1, r_2, r_3, r_4 and three internal nodes a_1, a_2, a_3. The allele state of terminal node r_1 is $r_1.s = 2$, the time backward from the present is $r_1.t = 0$, and the number of descendant nodes (not including the node itself) is $r_1.b = 0$. For the ancestral node a_1, suppose $a_1.s = 4$ and $a_1.t = 0.6$. From Figure 1d we can see that $a_1.b = 2$. Because the absolute difference in allele state between node a_1 and node r_1 is $|a_1.s - r_1.s| = 2$, assuming that mutations occur only through changes of $+1$ or -1 repeat unit, at least two mutation events must have happened along the branch connecting the two nodes.

Unlinked paralogs: Our aim in this section is to compute the probability distribution of the number of distinct alleles that will be observed in a single diploid individual for a duplicated microsatellite locus. The computation assumes that mutation follows the symmetric stepwise mutation model, that is, that microsatellite mutations occur by a single repeat unit, that mutations by $+1$ and by -1 unit are equally likely, and that the combined per-generation mutation rate for both types of mutations, μ, is independent of the current number of repeats. Figure 1a shows the simplest scenario, involving a single ancestral node and a single descendant node. Define the allele state difference between a_1 and r_1 as $d = |r_1.s - a_1.s|$ and the branch length as $t = a_1.t - r_1.t$. Following Equation 34 of WEHRHAHN (1975) and Equation 3 of WILSON and BALDING (1998), under the symmetric stepwise mutation model, the probability that the descendant node r_1 has allele state $r_1.s$ is

$$P(r_1.s \mid a_1.s, t, \theta) = e^{-t\theta/2} \sum_{k=0}^{\infty} \frac{(t\theta/4)^{2k+d}}{k!(k+d)!} = e^{-t\theta/2} I_d(t\theta/2). \quad (1)$$

This equation results from summing the Poisson probability of $2k + d$ mutations over all possible values of k, where k is the number of $+1$ mutations that are canceled by equally many -1 mutations. In (1), I_d denotes the dth-order modified Bessel function of the first kind (GRADSHTEYN and RYZHIK 1980). For simplicity, $P(r_1.s \mid a_1.s, t, \theta)$ is denoted in the following derivations as $V(r_1 - a_1, t)$, where r_1 and a_1 represent $r_1.s$ and $a_1.s$ (Table 1).

We now determine the probability distribution of the number of distinct alleles in an individual at a duplicated locus. Because both paralogs will be amplified by the same primers, this number can range from one to four, unlike in the case of a nonduplicated locus, for which it is either one or two. For now, we assume that the

TABLE 1

Notation

Symbol	Quantity represented
n	Number of alleles amplified (equal to twice the number of haploid genomes sampled, and equal to four throughout article).
N	Haploid effective population size [or scaling constant for the neutral coalescent model (Nordborg 2001)].
μ	Combined mutation rate for mutations of +1 and −1 repeat unit.
θ	Mutation rate parameter, $\theta = 2N\mu$.
q	A node of a tree, with three attributes: an allele state, a time (measured backward from the present), and a number of descendant terminal nodes extant in the present. These attributes are denoted $q.s$, $q.t$, and $q.b$, respectively.
$q.s$	Allele state attribute of a node q.
$q.t$	Time attribute of a node q, measured in units of N generations.
$q.b$	Number of descendant terminal nodes attribute of a node q.
a_k	Internal node (including the root): $k \in \{1, 2, \ldots, n-1\}$, $a_k.t > 0$, $a_k.b > 0$.
r_k	Terminal node: $k \in \{1, 2, \ldots, n\}$, $r_k.t = 0$, $r_k.b = 0$.
c_k	The event that k distinct alleles $k \in \{1, 2, \ldots, n\}$ are observed in a sample.
t_d	Time of the duplication event, measured in units of N generations.
$V(r_1 - a_1, t)$	The probability in time t of an allele state difference of $r_1.s - a_1.s$ between the allele state of node r_1 and that of its ancestor a_1. $V(k, t)$ refers to the probability $V(r_1 - a_1, t)$, when $k = r_1.s - a_1.s$.

two paralogs are completely unlinked, as might occur if a whole chromosome or genome was duplicated. We also assume that the duplication event happened longer ago than the time at which alleles at the two paralogous loci coalesce to their respective common ancestors and that the evolution of lineages at each paralog follows a coalescent model in a population of constant effective size N allelic copies. Mutation events on different branches of the genealogy of four alleles from the two paralogs (Figure 2) are assumed to occur independently.

As shown in Figure 2, define $t_d = a_3.t - r_1.t$, $t_{2a} = a_1.t - r_1.t$, and $t_{2b} = a_2.t - r_2.t$. We now must select a starting value for $a_3.s$. A sensible choice for mutation models with stationary distributions of allele size is to select $a_3.s$ according to the stationary distribution. The stepwise mutation model has no stationary distribution (Moran 1975), but because it also has no length dependence or allele size constraints, any choice for $a_3.s$ will produce the same probability distribution of the number of distinct alleles. We can therefore assume without loss of generality that at the time of duplication, $a_3.s = 0$. Then

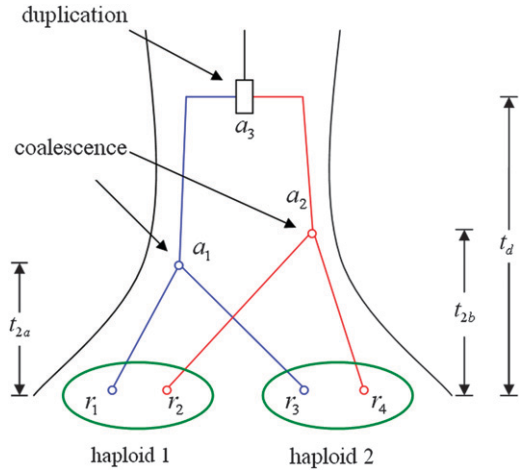

Figure 2.—Genealogy of two completely unlinked duplicated paralogous microsatellite loci, with sample size two haploid genomes. t_{2a} and t_{2b} have independent (truncated) exponential distributions. The genealogy of one paralog is in red, and the genealogy of the other paralog is in blue. The time t_d of the duplication event is assumed to be more ancient than the common ancestor of extant samples from a local population. Along the branches of this genealogy, looking backward in time, two coalescence events and one duplication event take place.

the probability of observing only one distinct allele in a sample—that is, the probability that all four sampled alleles have the same copy number—is

$$
\begin{aligned}
&P(c_1 \mid \theta, t_d, t_{2a}, t_{2b}) \\
&\quad = \sum_{k=-\infty}^{+\infty} P(r_1.s = r_2.s = r_3.s = r_4.s = k) \\
&\quad = \sum_{k=-\infty}^{+\infty} P(r_1.s = r_3.s = k) P(r_2.s = r_4.s = k) \\
&\quad = \sum_{k=-\infty}^{+\infty} \Bigg\{ \sum_{a_1.s=-\infty}^{+\infty} P(r_1.s = r_3.s = k \mid a_1.s) P(a_1.s) \\
&\qquad\qquad \times \sum_{a_2.s=-\infty}^{+\infty} P(r_2.s = r_4.s = k \mid a_2.s) P(a_2.s) \Bigg\}.
\end{aligned}
\tag{2}
$$

All probabilities are conditional on θ, t_d, t_{2a}, and t_{2b}, but for convenience, these quantities are not written in the following equations. Substituting (1) into (2) yields

$$
\begin{aligned}
&\sum_{a_1.s=-\infty}^{+\infty} P(r_1.s = r_3.s = k \mid a_1.s) P(a_1.s) \\
&\quad = \sum_{a_1.s=-\infty}^{+\infty} P(r_1.s = k \mid a_1.s) P(r_3.s = k \mid a_1.s) P(a_1.s \mid a_3.s) \\
&\quad = \sum_{a_1=-\infty}^{+\infty} V(a_1, t_d - t_{2a}) V(k - a_1, t_{2a})^2.
\end{aligned}
\tag{3}
$$

Similarly,

$$\sum_{a_2.s=-\infty}^{+\infty} P(r_2.s = r_4.s = k \mid a_2.s)P(a_2.s) = \sum_{a_2=-\infty}^{+\infty} V(a_2, t_d - t_{2b})V(k - a_2, t_{2b})^2. \quad (4)$$

Combining (2), (3), and (4) yields the probability of observing only one distinct allele,

$$P(c_1 \mid \theta, t_d, t_{2a}, t_{2b}) = \sum_{k=-\infty}^{+\infty} \left\{ \sum_{a_1=-\infty}^{+\infty} V(a_1, t_d - t_{2a})V(k - a_1, t_{2a})^2 \times \sum_{a_2=-\infty}^{+\infty} V(a_2, t_d - t_{2b})V(k - a_2, t_{2b})^2 \right\}. \quad (5)$$

This result can be viewed as a four-allele generalization of the classical computation for a nonduplicated locus of the probability under the stepwise model that two alleles have the same allele state (Ohta and Kimura 1973).

We now derive the probability of observing two distinct alleles. Using the tree topology (Figure 2) to identify the possible ways in which the genealogy could give rise to two distinct alleles,

$$\begin{aligned} P(c_2 \mid \theta, t_d, t_{2a}, t_{2b}) &= P(r_1.s = r_3.s, r_2.s = r_4.s, r_1.s \neq r_2.s) \\ &\quad + 2P(r_1.s = r_2.s = r_3.s, r_1.s \neq r_4.s) \\ &\quad + 2P(r_2.s = r_3.s = r_4.s, r_1.s \neq r_2.s) \\ &\quad + 2P(r_1.s = r_2.s, r_3.s = r_4.s, r_1.s \neq r_3.s). \end{aligned} \quad (6)$$

Applying the same conditional probability approach as was used in deriving the probability of observing only one distinct allele, we have

$$\begin{aligned} &P(r_1.s = r_3.s, r_2.s = r_4.s, r_1.s \neq r_2.s) \\ &= \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} P(r_1.s = r_3.s = k)P(r_2.s = r_4.s = l) \\ &= \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \left\{ \sum_{a_1.s=-\infty}^{+\infty} P(r_1.s = k \mid a_1.s)P(r_3.s = k \mid a_1.s)P(a_1.s) \right. \\ &\qquad \left. \times \sum_{a_2=-\infty}^{+\infty} P(r_2.s = l \mid a_2.s)P(r_4.s = l \mid a_2.s)P(a_2.s) \right\} \\ &= \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \sum_{a_1=-\infty}^{+\infty} \sum_{a_2=-\infty}^{+\infty} V(a_1, t_d - t_{2a})V(a_2, t_d - t_{2b}) \\ &\quad \times V(k - a_1, t_{2a})^2 V(l - a_2, t_{2b})^2. \end{aligned} \quad (7)$$

We also have the following three equations:

$$\begin{aligned} &P(r_1.s = r_2.s = r_3.s, r_1.s \neq r_4.s) \\ &= \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \sum_{a_1=-\infty}^{+\infty} \sum_{a_2=-\infty}^{+\infty} V(a_1, t_d - t_{2a})V(a_2, t_d - t_{2b}) \\ &\quad \times V(k - a_1, t_{2a})^2 V(k - a_2, t_{2b})V(l - a_2, t_{2b}), \end{aligned} \quad (8)$$

$$\begin{aligned} &P(r_2.s = r_3.s = r_4.s, r_1.s \neq r_2.s) \\ &= \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \sum_{a_1=-\infty}^{+\infty} \sum_{a_2=-\infty}^{+\infty} V(a_1, t_d - t_{2a})V(a_2, t_d - t_{2b}) \\ &\quad \times V(k - a_1, t_{2a})V(l - a_1, t_{2a})V(k - a_2, t_{2b})^2, \end{aligned} \quad (9)$$

$$\begin{aligned} &P(r_1.s = r_2.s, r_3.s = r_4.s, r_1.s \neq r_3.s) \\ &= \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \sum_{a_1=-\infty}^{+\infty} \sum_{a_2=-\infty}^{+\infty} V(a_1, t_d - t_{2a})V(a_2, t_d - t_{2b}) \\ &\quad \times V(k - a_1, t_{2a})V(l - a_1, t_{2a})V(k - a_2, t_{2b})V(l - a_2, t_{2b}). \end{aligned} \quad (10)$$

Combining Equations 7, 8, 9, and 10 we obtain

$$\begin{aligned} &P(c_2 \mid \theta, t_d, t_{2a}, t_{2b}) \\ &= \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \sum_{a_1=-\infty}^{+\infty} \sum_{a_2=-\infty}^{+\infty} V(a_1, t_d - t_{2a})V(a_2, t_d - t_{2b})[Q_2], \end{aligned} \quad (11)$$

where

$$\begin{aligned} Q_2 &= V(k - a_1, t_{2a})^2 V(l - a_2, t_{2b})^2 + 2V(k - a_1, t_{2a})^2 \\ &\quad \times V(k - a_2, t_{2b})V(l - a_2, t_{2b}) + 2V(k - a_1, t_{2a}) \\ &\quad \times V(l - a_1, t_{2a})V(k - a_2, t_{2b})^2 + 2V(k - a_1, t_{2a}) \\ &\quad \times V(l - a_1, t_{2a})V(k - a_2, t_{2b})V(l - a_2, t_{2b}). \end{aligned}$$

Applying the same approach, the probability of observing three distinct alleles is

$$\begin{aligned} &P(c_3 \mid \theta, t_d, t_{2a}, t_{2b}) \\ &= \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \sum_{\substack{m=-\infty \\ m\neq l,k}}^{+\infty} \sum_{a_1=-\infty}^{+\infty} \sum_{a_2=-\infty}^{+\infty} V(a_1, t_d - t_{2a}) \\ &\quad \times V(a_2, t_d - t_{2b})[Q_3], \end{aligned} \quad (12)$$

where

$$\begin{aligned} Q_3 &= V(k - a_1, t_{2a})^2 V(l - a_2, t_{2b})V(m - a_2, t_{2b}) \\ &\quad + V(k - a_1, t_{2a})V(l - a_2, t_{2b})^2 V(m - a_1, t_{2a}) \\ &\quad + 4V(k - a_1, t_{2a})V(k - a_2, t_{2b})V(l - a_1, t_{2a}) \\ &\quad \times V(m - a_2, t_{2b}), \end{aligned}$$

and the probability of observing four distinct alleles is

$$P(c_4 \mid \theta, t_d, t_{2a}, t_{2b}) = \sum_{k=-\infty}^{+\infty} \sum_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \sum_{\substack{m=-\infty \\ m\neq l,k}}^{+\infty} \sum_{\substack{n=-\infty \\ n\neq m,l,k}}^{+\infty} \sum_{a_1=-\infty}^{+\infty} \sum_{a_2=-\infty}^{+\infty} V(a_1, t_d - t_{2a}) \times V(a_2, t_d - t_{2b})[Q_4], \quad (13)$$

where $Q_4 = V(k - a_1, t_{2a})V(l - a_2, t_{2b})V(m - a_1, t_{2a})V(n - a_2, t_{2b})$.

Under the neutral coalescent model with N allelic copies in the population at a given locus, t_{2a} and t_{2b} follow independent exponential distributions with mean 1 time unit, where time is measured in units of N generations (Nordborg 2001). As a result, for unlinked paralogs—such as might be produced by genome duplication—their probability density functions are given by

$$f(t_{2a}) = e^{-t_{2a}} \quad \text{and} \quad f(t_{2b}) = e^{-t_{2b}}. \quad (14)$$

Using these two exponential distributions, we can integrate over values of t_{2a} and t_{2b} to derive the conditional probabilities of one, two, three, and four distinct alleles given only the mutation rate θ and the duplication time t_d. Because we assume that the most recent common ancestor for each paralog is more recent than the time of duplication, the calculation is conditional on the exponentially distributed coalescence times being smaller than t_d. Thus, for each i from 1 to 4 we have

$$P(c_i \mid \theta, t_d) = \frac{\int_0^{t_d}\int_0^{t_d} P(c_i \mid \theta, t_d, t_{2a}, t_{2b}) f(t_{2a}) f(t_{2b}) dt_{2a} dt_{2b}}{\int_0^{t_d}\int_0^{t_d} f(t_{2a}) f(t_{2b}) dt_{2a} dt_{2b}}. \quad (15)$$

Note that the denominator simplifies to $(1 - e^{-t_d})^2$.

Completely linked paralogs: For a model with completely linked loci, as might apply to a situation of tandem duplication, t_{2a} and t_{2b} are identical. Thus, the computations in this model can be viewed as a special case of those performed in the model with completely unlinked loci. Probabilities in this model are obtained by substituting t_{2a} and t_{2b} with a single variable, t_2, in Equations 2–13. Thus, for each i from 1 to 4 we have

$$P(c_i \mid \theta, t_d) = \frac{\int_0^{t_d} P(c_i \mid \theta, t_d, t_2) f(t_2) dt_2}{\int_0^{t_d} f(t_2) dt_2}. \quad (16)$$

The denominator simplifies to $1 - e^{-t_d}$.

METHODS OF COMPUTATION AND SIMULATION

We investigated the roles of t_d and θ in our duplicated microsatellite model (Figure 2) to understand the effects of the two parameters on the distribution of the number of distinct alleles in a single diploid individual. Using Mathematica (Wolfram Research, Champaign, IL), computations were performed with Equations 15 and 16. For computational efficiency, the sums indexed by a_1 and a_2 were replaced with a single sum indexed by $z = a_2 - a_1$, as described in the appendix. To make the computation feasible, infinite summations were truncated, and for all cases, each summation proceeded at least from −12 to +12. For large duplication times and mutation rates, the probability along a branch of a net allele size change >12 repeats in absolute value may have a nonnegligible probability. Denoting $P_y(c_2)$ as the probability $P(c_2)$ computed by truncating the sums from $-y$ to $+y$, the relative difference $|P_y(c_2) - P_{50}(c_2)|/P_{50}(c_2)$ tends to increase with increasing t_d and θ (results not shown). Thus, in the completely linked case, the truncation we used for the computation of $P(c_i)$ ($i = 1, 2, 3, 4$) at a given point (t_d, θ) employed at least as many terms as $\max(12, y^*)$, where y^* is a value that satisfies $|P_{y^*}(c_2) - P_{50}(c_2)|/P_{50}(c_2) < 0.01$ for some location (t_d^*, θ*) with $t_d^* \geq t_d$ and $\theta^* \geq \theta$. The same truncation as was used in the completely linked case was then used in the completely unlinked case.

In the completely unlinked case, the presence of a double integral (Equation 15) rather than a single integral makes the computational task more demanding than in the completely linked case. Joint analysis of t_d and θ also requires more computation than analysis of one parameter at fixed values of the other parameter. Thus, for all figures except Figures 3–5, which were obtained numerically, we used a simulation approach to study the joint effects of t_d and θ. In this approach, we first choose the coalescence times of the paralogs from an exponential distribution with mean 1, conditional on the exponential random variable being smaller than the duplication time. We then simulate mutations along the tree from the time of duplication forward, one branch at a time. A number of mutations is chosen for a given branch of length t, using a Poisson distribution with mean $\theta t/2$. After mutation events are placed, each mutation is specified as being a +1 or a −1 change in copy number, each with probability $\frac{1}{2}$. Allele states at the nodes corresponding to the paralog coalescences are recorded as the basis for simulation of mutations on the external branches. The simulated mutations on the external branches in turn lead to sizes for the four alleles, and each simulated tree is classified by the number of distinct alleles produced. The fractions of simulated trees with one, two, three, and four distinct alleles are then used as estimates of $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$. To ensure convergence in probability, 500,000 simulated trees were obtained for each combination of t_d and θ. The simulation results, which underlie Figures 6–12, were found to be consistent with exact computations at various sets of parameter values. For example, comparing the simulated values of $P(c_2)$ and the computations based on the exact formula in the linked case (Equation 16, with each term truncated from −40 to +40), using a fixed θ of 3.5 and varying t_d from 2 to 12 at intervals of 1 (with an extra point at

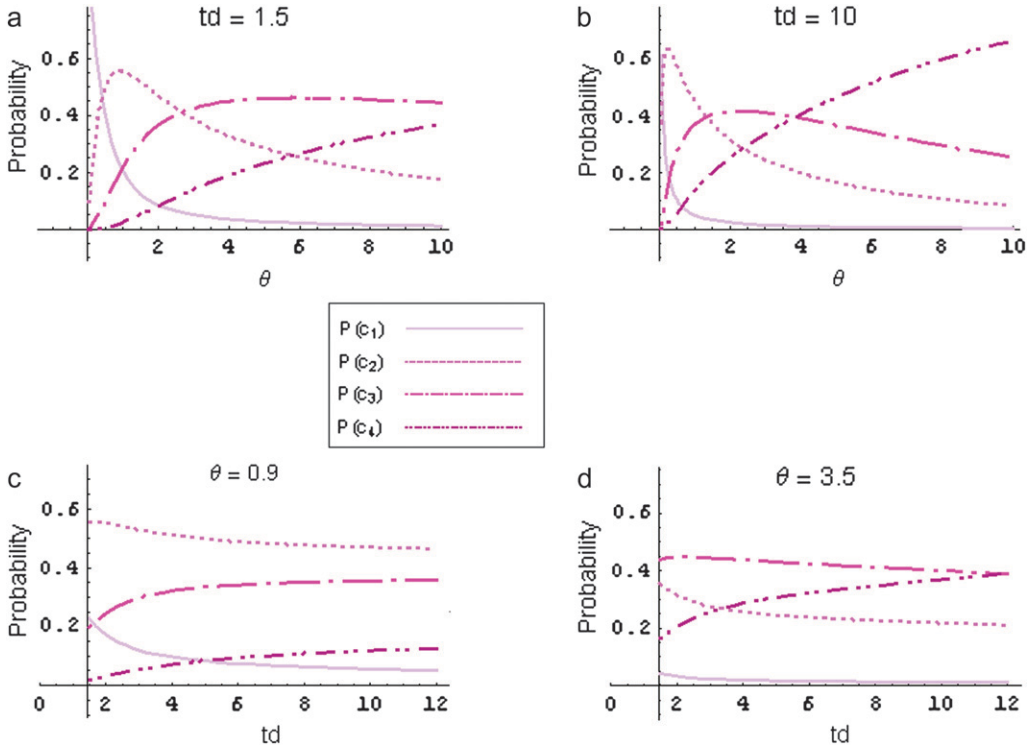

FIGURE 3.—Probability distribution of the number of distinct alleles for a pair of completely linked paralogous loci in a diploid individual. The probabilities of one, two, three, and four alleles are colored in four different shades. (a and b) The effect of varying θ from 0 to 10, with fixed small t_d (1.5) (a) and fixed large t_d (10) (b). (c and d) The effect of varying t_d from 1 to 12, with fixed small θ (0.9) (c) and fixed large θ (3.5) (d). In all graphs, $t_2 \sim \exp(1)$ (with the restriction that t_2 is strictly less than t_d).

$t_d = 1.5$), no significant difference is found ($P = 1.00$, two-sample Kolmogorov–Smirnov test). Similarly, no difference is found using a fixed t_d of 6, varying θ from 1 to 12 at intervals of 1 ($P = 1.00$, two-sample Kolmogorov–Smirnov test).

THE ROLE OF THE PARAMETERS

Figure 3 shows the role of the mutation rate θ and the duplication time t_d, in the case where the two paralogous loci are completely linked. We computed the distribution of the number of distinct alleles, varying θ for fixed values of t_d and varying t_d for fixed values of θ. Figure 3 shows the influence of θ for a "small" t_d of 1.5 units of coalescent time and for a "large" t_d of 10 units, with θ ranging from 0 to 10, and the influence of t_d for a small θ of 0.9 and a large θ of 3.5, with t_d ranging from 1.5 to 12.

The mutation parameter θ plays an important role in shaping the probability distribution of the number of distinct alleles, regardless of the value of t_d. When θ is near zero, mutations rarely occur and the current allele state is likely to be the same as the ancestral state. Thus, the scenario of a single distinct allele predominates. However, this pattern quickly disappears as θ increases. Eventually, for large values of θ, sufficiently many mutations occur that all four alleles are likely to be distinct.

When t_d is large (Figure 3b), as θ increases from 0 to 10, the probability that the number of distinct alleles is one decreases sharply from near 1 to near 0. The probability of two distinct alleles first increases to ~0.65, but then decreases slowly. The two-allele configuration is dominant for a short range, until $P(c_3)$ outpaces $P(c_2)$ near θ = 1.5. The probability of three distinct alleles increases quickly for small θ, and for θ from ~1.5 to ~4.0, the three-allele configuration is most probable, with a relatively stable probability over this range. The probability of four distinct alleles rises slowly but monotonically. Finally, at θ ≥ 4.0, the configuration with four distinct alleles becomes most probable.

When t_d is small (Figure 3a), compared to the case when t_d is large, the rate of decrease for one distinct allele and the rates of increase for two, three, and four distinct alleles are slower. At θ ~ 2.8, the configuration with three distinct alleles becomes most frequent and predominates as θ increases to 10. The difference in the speed at which the four possibilities change their order in the cases of large t_d and small t_d is primarily a result of long branches between the time of duplication and the times of coalescence of the individual paralogs in the former case and short branches separating these two events in the latter. When these branches are long, the increasing mutation rate will cause them to accumulate many mutations. These mutations will likely increase the number of distinct alleles to at least two, with nodes r_1 and r_3 having one allele state and nodes r_2 and r_4 having another. As θ increases further, mutations will occur on the shorter branches between the paralog coalescences and the present, further increasing the number of distinct alleles. Regardless of the duplication time, however, as the mutation rate becomes large for a fixed t_d, the probability of fewer than four distinct alleles becomes negligible.

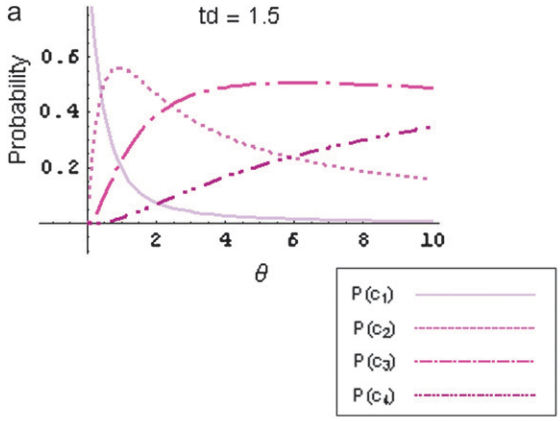

FIGURE 4.—Probability distribution of the number of distinct alleles for a pair of completely unlinked paralogous loci in a diploid individual. The probabilities of one, two, three, and four alleles are colored in four different shades. (a and b) The effect of varying θ from 0 to 10, with fixed small t_d (1.5) (a) and fixed large t_d (10) (b). (c and d) The effect of varying t_d from 1 to 12, with fixed small θ (0.9) (c) and fixed large θ (3.5) (d). In all graphs, $t_{2a} \sim \exp(1)$ and $t_{2b} \sim \exp(1)$ (with the restriction that t_{2a} and t_{2b} are strictly less than t_d).

The duplication time t_d plays a less important role than the mutation rate in shaping the probabilities of the four values for the number of distinct alleles. As t_d increases from 1.5 to 12 in Figure 3, c and d, the probabilities of one and two distinct alleles slowly decrease. When θ is small (Figure 3c), configurations with three and four distinct alleles slowly increase in probability, while the number of distinct alleles with the

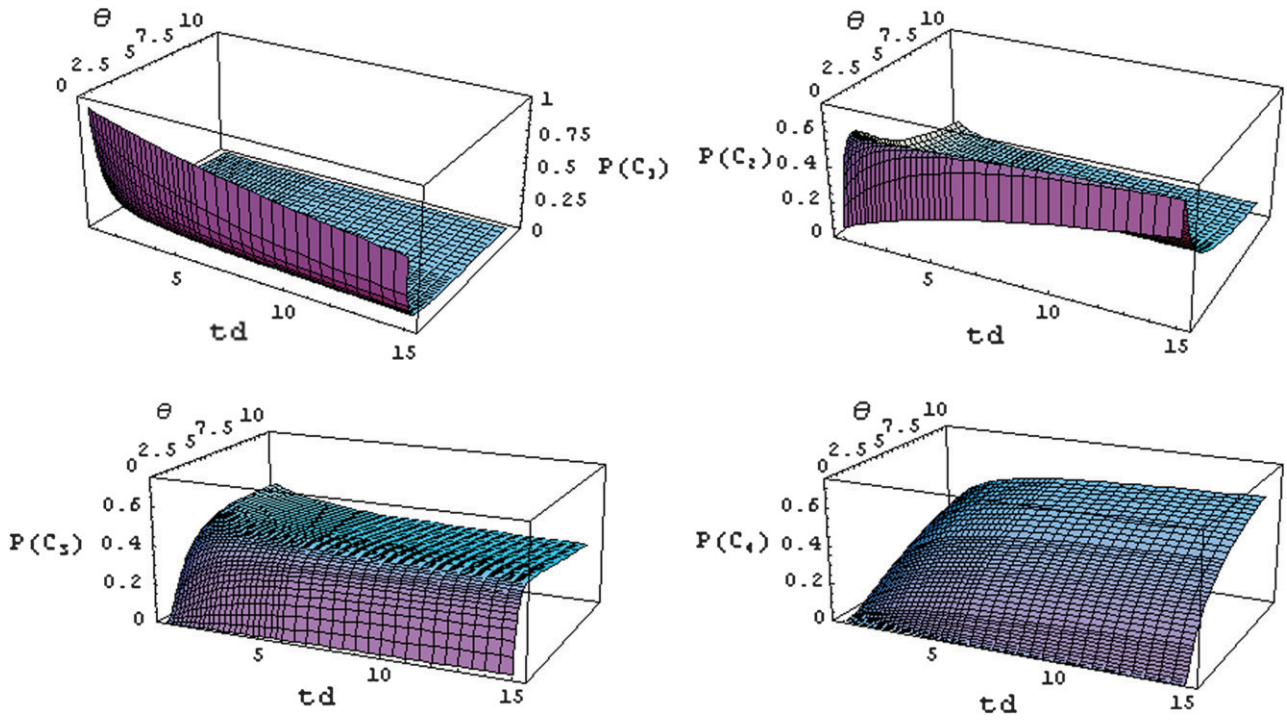

FIGURE 5.—Probability distribution of the number of distinct alleles at a pair of completely linked paralogous loci in a diploid individual. The time of duplication t_d ranges from 1.0 to 15.0, the mutation rate θ ranges from 0.05 to 11.0, and the time of the coalescence events, t_2, has exponential distribution with mean 1 (truncated to be strictly less than t_d); the four probabilities $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$ are displayed in the four different graphs. Each data point is computed numerically from a truncated summation as described in METHODS OF COMPUTATION AND SIMULATION.

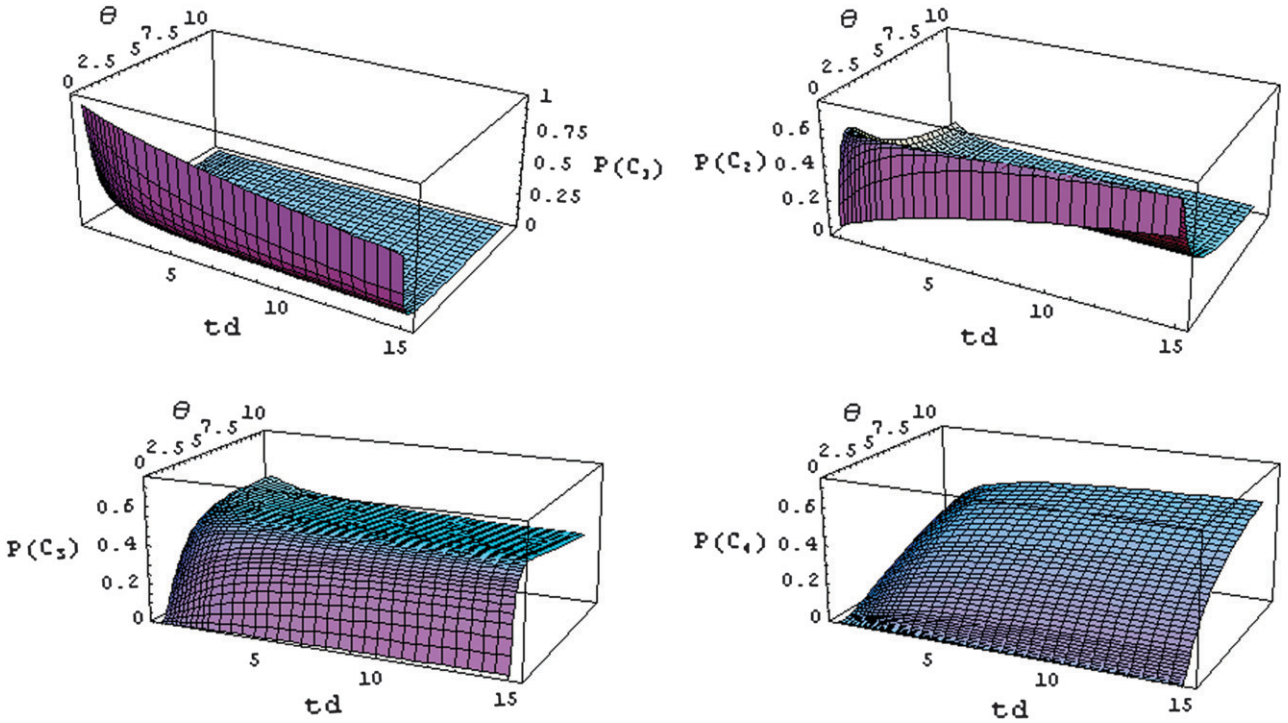

FIGURE 6.—Probability distribution of the number of distinct alleles at a pair of completely unlinked paralogous loci in a diploid individual. The time of duplication t_d ranges from 1.0 to 15.0, the mutation rate θ ranges from 0.05 to 11.0, and the times of two independent coalescence events, t_{2a} and t_{2b}, have exponential distributions with mean 1 (truncated to be strictly less than t_d); the four probabilities $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$ are displayed in the four different graphs. Each data point is based on 500,000 simulations.

highest probability is two. When θ is larger, however (Figure 3d), the probability of three alleles slowly decreases but remains highest until the right side of the graph, where $P(c_4)$ begins to dominate. The graphs illustrate that as t_d becomes large, for a fixed mutation rate, the probability of four alleles does not approach 1. Mutations will tend to occur on the long branches between duplication and paralog coalescence, so that the probability of having at least two distinct alleles increases. However, even for large t_d, coalescence times for the individual paralogs remain relatively small. Thus, for small mutation rates, mutations are unlikely to occur on the short branches between coalescence and the present; for large mutation rates, such mutations are more likely, although some probability exists that they do not occur. As t_d increases, the distribution of the number of alleles therefore approaches a distribution in which the probability is concentrated on two, three, and four distinct alleles and in which the relative probabilities of these configurations depend on the mutation rate. Similar behavior is observed in the case of completely unlinked loci (Figure 4).

Figures 5 and 6 show the probability surface for completely linked and unlinked paralogous loci with sample size two haploid genomes, covering the range of t_d from 1.0 to 15.0 and θ from 0.05 to 11.0. In both the linked and the unlinked cases, the probability $P(c_1)$ decreases quickly as θ increases, as the increasing number of mutations reduces the chance that alleles will be identical. $P(c_2)$ first increases as θ increases, reaching a local maximum. For t_d large relative to t_{2a} and t_{2b}, this maximum can be viewed as a consequence of the fact that there is some range of θ-values for which multiple mutations are likely to occur on the long internal branches, but not on the short external branches. In this range, $r_1.s$ and $r_3.s$ are likely to be equal due to the lack of external mutations—as are $r_2.s$ and $r_4.s$—but because many mutations will occur on the long internal branches, it is likely that $r_1.s$ and $r_3.s$ will differ from $r_2.s$ and $r_4.s$. As t_d increases, the range of θ-values that produces this phenomenon decreases in size, resulting in a narrower "ridge" on the surface for large t_d. $P(c_3)$ increases as θ increases, as mutations begin to occur on external branches. Finally, $P(c_4)$ is very small when θ is small, but it increases monotonically with increases in t_d and especially in θ, as it becomes increasingly likely that mutations will occur on all branches, internal and external. For θ sufficiently large, regardless of t_d, the configuration of four distinct alleles eventually predominates.

The general trend of increasing numbers of distinct alleles with increases in θ and t_d can be visualized in Figures 7a and 8a, which show the mean number of distinct alleles as functions of θ and t_d in the linked and unlinked cases. However, the variance of the number of distinct alleles, plotted in Figures 7b and 8b, is not

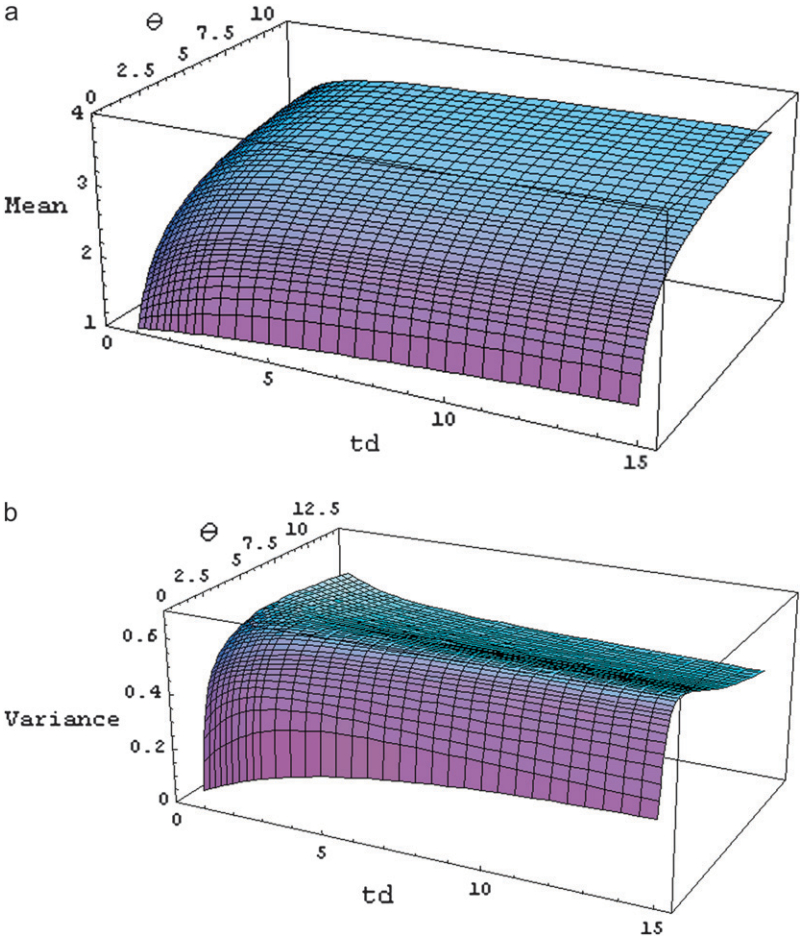

FIGURE 7.—Mean and variance of the number of distinct alleles at a pair of completely linked paralogous loci in a diploid individual. (a) Mean. (b) Variance.

monotonic in θ and t_d and instead has a ridge-like shape similar to that of $P(c_2)$ in Figures 5 and 6. This behavior reflects the fact that for a given value of t_d, at extreme values of θ where one configuration predominates, the number of distinct alleles has a low variance, whereas for intermediate values of θ where multiple configurations have nontrivial probabilities, the variance is relatively high.

Two additional ways of viewing the four probabilities—partitions of the parameter space by the relative order of the four probabilities and visualizations of the first-, second-, third-, and fourth-place probabilities—can help to identify features that are not easy to recognize in the graphs of the probability functions described above. For example, small differences between the unlinked case and the linked case, which are not conspicuous in the surface plots of Figures 5–8, can be viewed in the summaries of the parameter space plotted in Figures 9–12.

Figures 9 and 10 partition the parameter space by the relative order of the four probabilities $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$. Of the 24 possible orders in which these four probabilities could potentially occur, only 7 were observed at any point in the parameter space. The same sequence of transitions with increasing values of θ was always observed. Initially, $P(c_1) > P(c_2) > P(c_3) > P(c_4)$. To reach the eventual state in which $P(c_4) > P(c_3) > P(c_2) > P(c_1)$, the following transpositions then occur: $P(c_2)$ and $P(c_1)$, $P(c_1)$ and $P(c_3)$, $P(c_1)$ and $P(c_4)$, $P(c_2)$ and $P(c_3)$, $P(c_2)$ and $P(c_4)$, and $P(c_3)$ and $P(c_4)$. One difference between the linked and the unlinked cases is that for the unlinked case (Figure 10), the intermediate regions of the parameter space shown in blue have a greater area. In comparison with the linked case (Figure 9), which has a single coalescence time shared by both paralogs, in the unlinked case, there is a greater chance that one of the two coalescence times for the two paralogs will be extreme (either very large or very small). If one of these coalescence times is particularly large, θ must be smaller to have a chance of no mutations on the external branches between that coalescence and the present, so that situations with high values of $P(c_1)$ and $P(c_2)$ require smaller values of θ. Conversely, if one of the coalescence times is particularly small, θ must be larger to have a chance of some mutations on the external branches between that coalescence and the present. Thus, situations with a high value of $P(c_4)$ require larger values of θ. This same trend, in which the unlinked case has a greater proportion of the parameter space where the situation of three

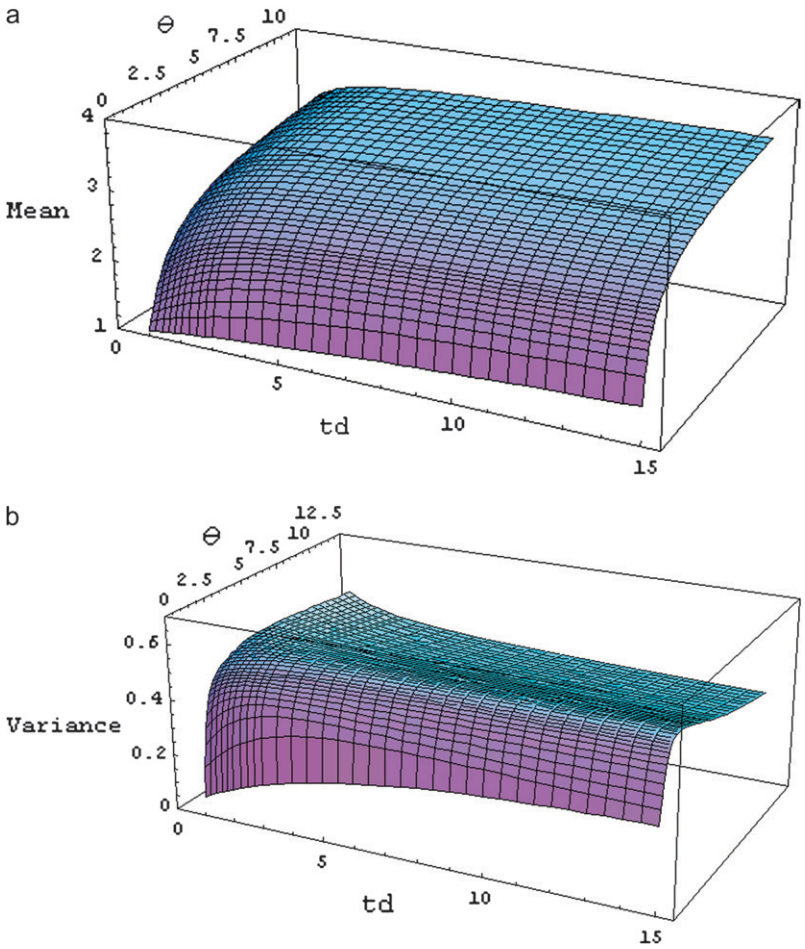

FIGURE 8.—Mean and variance of the number of distinct alleles at a pair of completely unlinked paralogous loci in a diploid individual. (a) Mean. (b) Variance.

distinct alleles predominates, is visible in Figures 11 and 12, which illustrate at each point in the parameter space the first, second, third, and fourth highest values among $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$. The part of the parameter space where $P(c_3)$ is in "first place" and the part where $P(c_4)$ is in "second place" are both larger in the unlinked case (Figure 12) than in the linked case (Figure 11). Note that in both the linked and the unlinked cases, as can be inferred from the list of the seven partitions of the parameter space in Figures 9 and 10, at every point in the parameter space, either $P(c_1)$ or $P(c_4)$ was always observed to be in fourth place.

DISCUSSION

In this article, we have introduced a model for studying evolution at duplicated microsatellite loci. The model incorporates a stepwise model for microsatellite mutation, together with a coalescent model for pairs of lineages of the same paralog, and was studied both in the case of unlinked paralogs (chromosome or genome duplication) and in the case of linked paralogs (tandem duplication). Using our model we have derived the distribution of the number of distinct alleles that will be amplified by the PCR primers of a duplicated microsatellite.

The two parameters of the model are the duplication time and the mutation rate. We found that the mutation rate θ has a strong influence on the probability distribution of the number of distinct alleles, in that the probabilities of one, two, three, and four alleles vary greatly with θ. As θ increases, the probability of four distinct alleles quickly becomes quite high. The duplication time has less of an influence: as the duplication time increases—although the probability of one distinct allele becomes negligible—the probabilities of two, three, and four alleles slowly approach values that depend on the mutation rate.

While our main goal has been to explore the properties of a relatively simple model, our results provide information about potential uses of duplicated microsatellites for inference. One application may be to infer θ and t_d for a partial genome duplication by genotyping a single individual at many duplicated microsatellites assumed to have the same mutation properties. In this strategy, the fractions of loci observed to have one, two, three, and four alleles are taken as estimates of $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$. However, the shapes of the likelihood surfaces in Figures 5 and 6 suggest that for many

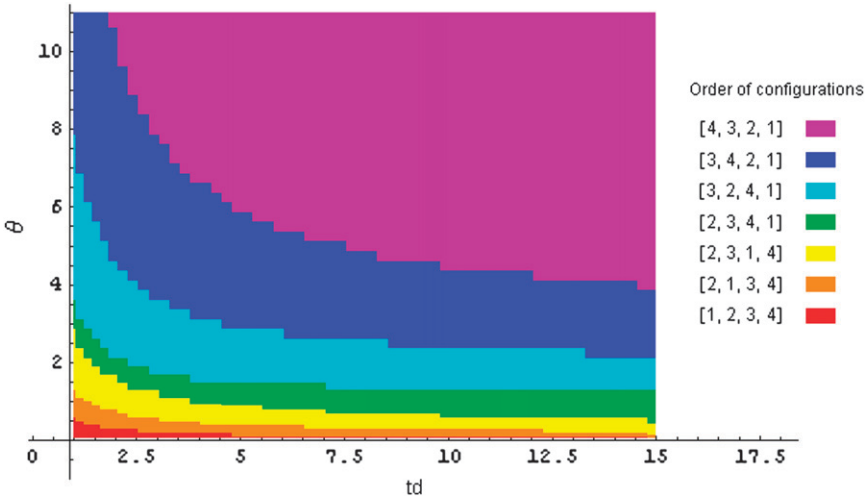

FIGURE 9.—Partition of the parameter space in the completely linked case, according to the order of the four probabilities $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$. For example, $[2, 1, 3, 4]$ refers to a situation where $P(c_2) > P(c_1) > P(c_3) > P(c_4)$.

combinations $[P(c_1), P(c_2), P(c_3)$, and $P(c_4)]$, the parameters will not both be identifiable, as points on a ridge with decreasing θ and increasing t_d may all produce similar probability distributions. If one of the two parameters is estimated by other means—such as by using nonduplicated microsatellites to estimate θ—there may be some possibility of estimating the other parameter. Another approach may be an expansion of the model to accommodate sample sizes larger than one diploid individual.

A second potential application concerns the linkage status of the two loci in a duplicate pair. Although some subtle differences are observable, the model illustrates that unlinked and linked paralogs produce similar distributions for the number of distinct alleles (Figures 5 and 6). This suggests that the frequency distribution across individuals of the number of distinct alleles—which can provide a valuable source of information for identifying that a locus is duplicated—is not very informative about the linkage relationship of the two paralogs.

In applications of the model, it will be important to determine how inferences depend on changes to the mutation scheme. We have focused on a simple mutation model due to its relative tractability, but observations of microsatellite loci suggest a variety of deviations from the model, including length-dependent mutation rates, multistep mutations, mutation-influencing interruptions, and different probabilities of upward and downward mutations (MATSUOKA *et al.* 2002; WHITTAKER *et al.* 2003; ELLEGREN 2004; SAINUDIIN *et al.* 2004; CALABRESE and SAINUDIIN 2005). Due to the particular choice of variable that we have studied—the number of distinct alleles in a single individual—some deviations from the stepwise mutation model may have only modest effects. For example, holding the overall mutation rate constant and assuming length-independent single-step mutation, asymmetry in the upward and downward mutation rates has no effect on the probability of identity for two alleles (KIMMEL and CHAKRABORTY 1996). Thus, it seems unlikely that such asymmetry would

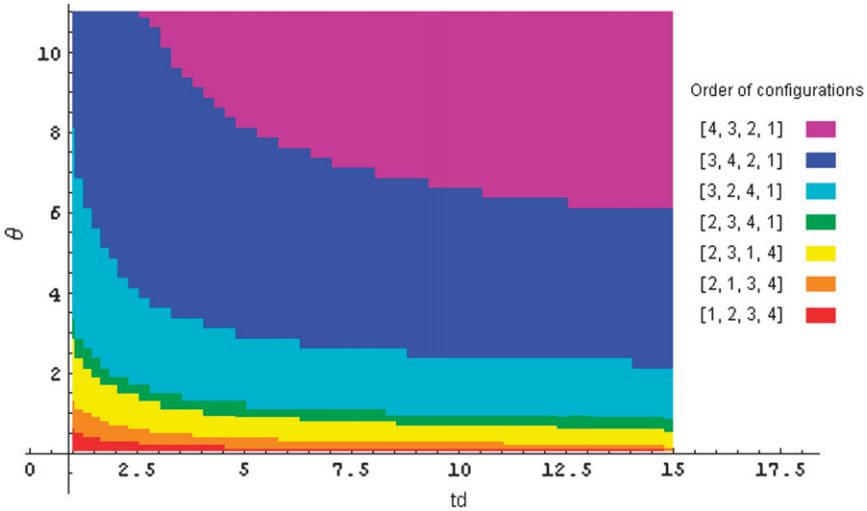

FIGURE 10.—Partition of the parameter space in the completely unlinked case, according to the order of the four probabilities $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$. For example, $[2, 1, 3, 4]$ refers to a situation where $P(c_2) > P(c_1) > P(c_3) > P(c_4)$.

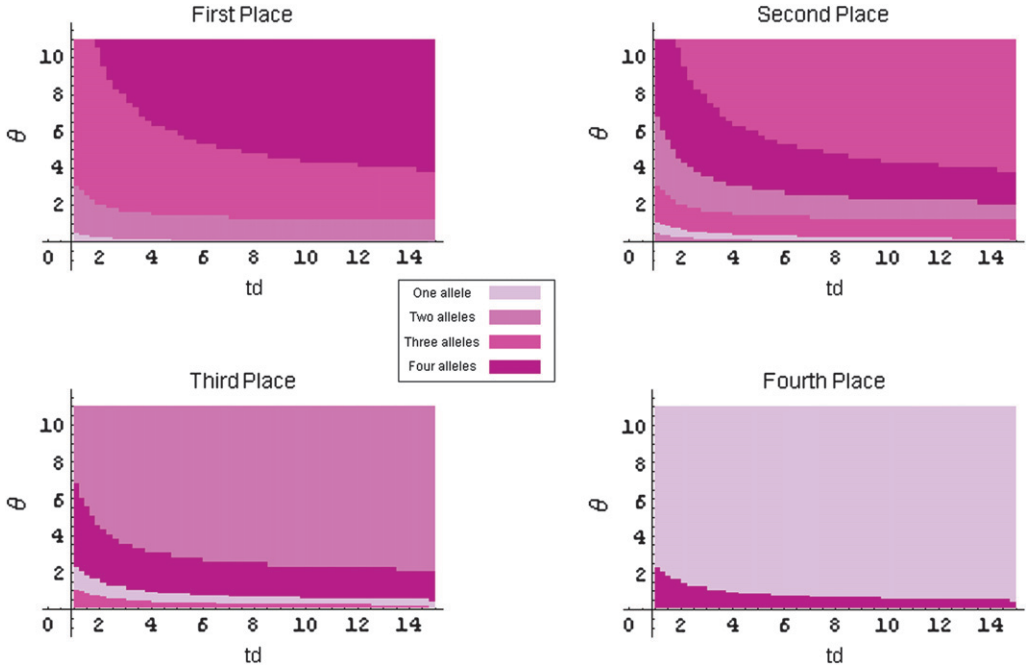

FIGURE 11.—First-, second-, third-, and fourth-place configurations for the linked case. For example, over a range of values of θ and t_d, the top left graph shows which configuration of alleles—one, two, three, or four distinct alleles—is most probable, that is, in "first place."

strongly influence the scenario of four alleles considered here.

It is noteworthy that for the model to be applicable, duplication must have occurred sufficiently recently that the microsatellite that experienced duplication is still polymorphic and actively mutating. Such scenarios may occur in various species of fish, in which large-scale duplications have occurred quite recently, and in which substantial numbers of polymorphic microsatellites are duplicated (DAVID *et al.* 2003, 2007; O'MALLEY *et al.* 2003). Between humans and chimpanzees, however, perhaps only ~6% of genes differ due to gene gains and losses (DEMUTH *et al.* 2006), and thus, microsatellites duplicated on the timescale for production of genetic variation within the human species are likely to be quite exceptional.

Although our model has been limited to a single individual, a single duplication event, and relatively simple assumptions about mutation, it provides a beginning for understanding the factors that affect patterns of variation at a duplicated microsatellite. With the increasing frequency of observations of duplicated microsatellites

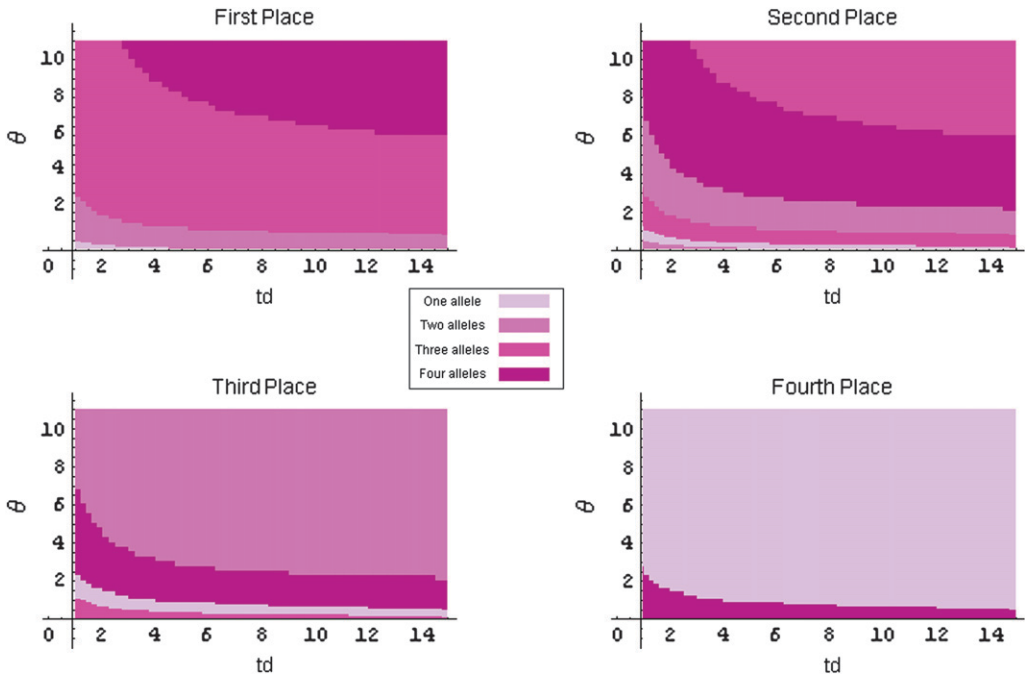

FIGURE 12.—First-, second-, third-, and fourth-place configurations for the unlinked case. For example, over a range of values of θ and t_d, the top left graph shows which configuration of alleles—one, two, three, or four distinct alleles—is most probable, that is, in "first place."

in organisms where recent gene and genome duplications have played an important role in evolution (David *et al.* 2003, 2007; O'Malley *et al.* 2003), further development of coalescent-based models of duplicated microsatellites may lead to new tools for inference about the evolutionary process in these organisms.

We thank D. Balding, L. David, M. Jakobsson, B. Padhukasahasram, V. Plagnol, S. Tavaré, J. Wall, and K. Zhao for suggestions and M. Uyenoyama and an anonymous reviewer for careful comments on an earlier version of this manuscript. K.Z. is partially supported by National Institutes of Health (NIH) grant R01GM074163 to X. J. Zhou and by the Burroughs Wellcome Fund Mathematics and Molecular Biology trainee program. This work was supported in part by National Science Foundation grant DEB-0716904, by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences, and by an Alfred P. Sloan Research Fellowship.

LITERATURE CITED

Antunes, A., K. Gharbi, P. Alexandrino and R. Guyomard, 2006 Characterization of *transferrin*-linked microsatellites in brown trout (*Salmo trutta*) and Atlantic salmon (*Salmo salar*). Mol. Ecol. Notes **6:** 547–549.

Balaresque, P., A. Sibert, E. Heyer and B. Crouau-Roy, 2007 Unbiased interpretation of haplotypes at duplicated microsatellites. Ann. Hum. Genet. **71:** 209–219.

Blum, M. G. B., C. Damerval, S. Manel and O. Francois, 2004 Brownian models and coalescent structures. Theor. Popul. Biol. **65:** 249–261.

Calabrese, P., and R. Sainudiin, 2005 Models of microsatellite evolution, pp. 289–305 in *Statistical Methods in Molecular Evolution*, edited by R. Nielsen. Springer, New York.

David, L., S. Blum, M. W. Feldman, U. Lavi and J. Hillel, 2003 Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. Mol. Biol. Evol. **20:** 1425–1434.

David, L., N. A. Rosenberg, U. Lavi, M. W. Feldman and J. Hillel, 2007 Genetic diversity and population structure inferred from the partially duplicated genome of domesticated carp, *Cyprinus carpio* L. Genet. Sel. Evol. **39:** 319–340.

Demuth, J. P., T. De Bie, J. E. Stajich, N. Cristianini and M. W. Hahn, 2006 The evolution of mammalian gene families. PLoS ONE **1:** e85.

Ellegren, H., 2004 Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. **5:** 435–445.

Estoup, A., P. Jarne and J.-M. Cornuet, 2002 Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol. Ecol. **11:** 1591–1604.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman, 1995 An evaluation of genetic distances for use with microsatellite loci. Genetics **139:** 463–471.

Gradshteyn, I. S., and I. M. Ryzhik, 1980 *Table of Integrals, Series, and Products*. Academic Press, London.

Innan, H., 2002 A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. Genetics **161:** 865–872.

Innan, H., 2003 The coalescent and infinite-site model of a small multigene family. Genetics **163:** 803–810.

Innan, H., 2004 Theories for analyzing polymorphism data in duplicated genes. Genes Genet. Syst. **79:** 65–75.

Kimmel, M., and R. Chakraborty, 1996 Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor. Popul. Biol. **50:** 345–367.

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

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459–473.

Mathias, N., M. Bayes and C. Tyler-Smith, 1994 Highly informative compound haplotypes for the human Y chromosome. Hum. Mol. Genet. **3:** 115–123.

Matsuoka, Y., S. E. Mitchell, S. Kresovich, M. Goodman and J. Doebley, 2002 Microsatellites in *Zea*—variability, patterns of mutations, and use for evolutionary studies. Theor. Appl. Genet. **104:** 436–450.

Moran, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

Nielsen, R., 1997 A likelihood approach to populations samples of microsatellite alleles. Genetics **146:** 711–716.

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

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

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

O'Malley, K. G., T. Sakamoto, R. G. Danzmann and M. M. Ferguson, 2003 Quantitative trait loci for spawning date and body weight in rainbow trout: testing for conserved effects across ancestrally duplicated chromosomes. J. Hered. **94:** 273–284.

Pritchard, J. K., and M. W. Feldman, 1996 Statistics for microsatellite variation based on coalescence. Theor. Popul. Biol. **50:** 325–344.

Sainudiin, R., R. T. Durrett, C. F. Aquadro and R. Nielsen, 2004 Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics **168:** 383–395.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.

Teshima, K. M., and H. Innan, 2004 The effect of gene conversion on the divergence between duplicated genes. Genetics **166:** 1553–1560.

Walsh, B., 2003 Population-genetic models of the fates of duplicate genes. Genetica **118:** 279–294.

Ward, R., and R. Durrett, 2004 Subfunctionalization: How often does it occur? How long does it take? Theor. Popul. Biol. **66:** 93–100.

Wehrhahn, C. F., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.

Whittaker, J. C., R. M. Harbord, N. Boxall, I. Mackay, G. Dawson *et al.*, 2003 Likelihood-based estimation of microsatellite mutation rates. Genetics **164:** 781–787.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Zhang, J., 2003 Evolution by gene duplication: an update. Trends Ecol. Evol. **18:** 292–298.

Zhivotovsky, L. A., and M. W. Feldman, 1995 Microsatellite variability and genetic distances. Proc. Natl. Acad. Sci. USA **92:** 11549–11552.

Communicating editor: M. K. Uyenoyama

APPENDIX

The probability distribution of the number of distinct alleles in the scheme of Figure 2 is equivalent to the corresponding distribution in a scheme in which the branch connecting a_3 and a_1 is contracted to a single point, the branch connecting a_3 and a_2 is extended to have length $2t_d - t_{2a} - t_{2b}$, and no changes are made to the branches connecting internal nodes a_1 and a_2 to descendant nodes r_1, r_2, r_3, and r_4.

The probabilities $P(c_1)$, $P(c_2)$, $P(c_3)$, and $P(c_4)$ in Equations 5 and 11–13 each include terms of the form

$$\sum_{a_1=-\infty}^{+\infty} \sum_{a_2=-\infty}^{+\infty} V(a_1, t_d - t_{2a}) V(a_2, t_d - t_{2b}). \qquad \text{(A1)}$$

By making the substitution $z = a_2 - a_1$ and interchanging the order of summation, the sum can be transformed to

$$\sum_{z=-\infty}^{+\infty} \sum_{a_1=-\infty}^{+\infty} V(a_1, t_d - t_{2a}) V(a_1 + z, t_d - t_{2b}). \quad \text{(A2)}$$

Because of symmetry in the mutation process, $V(a_1, t_d - t_{2a}) = V(-a_1, t_d - t_{2a})$. Thus, the inner sum can be seen to equal $V(z, 2t_d - t_{2a} - t_{2b})$, as it can be obtained by considering a single branch of length $2t_d - t_{2a} - t_{2b}$ traveling up the genealogy from node a_1 to node a_3 and then down to node a_2, rather than by considering separate branches from a_3 to a_1 and to a_2. Therefore, for computational efficiency, we can replace double sums of the form of Equation A1 in Equations 5, 11, 12, and 13 with the single sum

$$\sum_{z=-\infty}^{+\infty} V(z, 2t_d - t_{2a} - t_{2b}). \quad \text{(A3)}$$

A similar idea was used by Pritchard and Feldman (1996) for other calculations under symmetric stepwise mutation.