The Annals of Statistics
1997, Vol. 25, No. 1, 138–168

MARKOV CHAINS FOR MONTE CARLO TESTS OF GENETIC EQUILIBRIUM IN MULTIDIMENSIONAL CONTINGENCY TABLES

BY LAURA C. LAZZERONI[1] AND KENNETH LANGE[2]

Stanford University and University of Michigan

Hardy–Weinberg equilibrium and linkage equilibrium are fundamental concepts in population genetics. In practice, testing linkage equilibrium in haplotype data is equivalent to testing independence in a large, sparse, multidimensional contingency table. Testing Hardy–Weinberg and linkage equilibrium simultaneously on multilocus genotype data introduces the additional complications of missing information and symmetry constraints on marginal probabilities. To avoid unreliable large-sample approximations for sparse contingency tables, one can use exact tests like Fisher's classical test that condition on observed marginal totals. Unfortunately, computing p-values for exact tests is often infeasible because of the large number of tables consistent with the marginal totals of an observed table. We develop here Markov chains for sampling from the appropriate conditional distributions for testing genetic equilibrium. These chains compare favorably with a parallel, independent-sampling method that we present. For n haplotype observations on J loci, the Markov chains converge to their stationary distributions in $[(J-1)n \ln n]/2 + O(n)$ steps and can be an efficient tool for estimating p-values. Our theoretical treatment of these results involves strong stationary stopping times, order statistics, large deviations and the embedding of Poisson processes. We include some general results on the application of strong stationary times to bounding the precision and bias of sample average estimators.

1. Introduction. The concepts of Hardy–Weinberg equilibrium and linkage equilibrium are central in population genetics theory [Crow (1988)]. These independence assumptions for genotype and haplotype frequencies simplify analyses such as gene mapping calculations, forensic applications of DNA fingerprinting and genetic counselling [Weir (1990)]. Such analyses routinely incorporate the product rules for genotype and haplotype frequencies implied by genetic equilibrium. Fortunately, the allele frequencies required for these calculations can be estimated more easily and precisely than either genotype or haplotype frequencies. Despite these simplifications, genetic disequilibrium also has its uses. For instance, it can point to the

Received September 1995; revised April 1996

[1] Research supported in part by USPHS Grants GM-08185 and GM-53275 and NSF Grant DMS-95-10516.

[2] Research supported in part by USPHS Grants GM-53275 and CA-16042.

AMS 1991 *subject classifications*. 62P10, 65C05.

Key words and phrases. Exact inference, strong stationary stopping time, large deviations, Hardy–Weinberg, linkage equilibrium, independent sampling, random transpositions.

association of particular alleles with an increased risk of disease, reveal population substructure or serve as a guide in positional cloning [Weir (1990)]. Thus, tests of genetic equilibrium are of fundamental importance in population genetics.

Methods for testing genetic equilibrium have the added bonus of extending to contingency tables in other application areas. When haplotype information is available, testing linkage equilibrium is equivalent to testing independence in a J-way contingency table. When haplotype information is missing, Hardy–Weinberg and linkage equilibrium can be tested simultaneously. The observed genotype structure then imposes special symmetry constraints on the marginal probabilities of the underlying contingency table. In either genetic setting, the contingency tables encountered in practice tend to be large and sparse. Because expected cell counts are low and many nuisance parameters must be estimated for such tables, p-values of tests based on traditional large-sample approximations can be unreliable [Elston and Forthofer (1977), Emigh (1980), Agresi (1992)].

In theory, exact tests can be constructed by conditioning inference upon marginal counts [Haldane (1954), Louis and Dempster (1987), Agresi (1992)]. Although this strategy eliminates nuisance parameters, it may fail because of the difficulty in evaluating conditional likelihoods determined by large numbers of cells subject to complicated marginal constraints. A useful alternative to deterministic computation is to use Monte Carlo methods to sample from the conditional distribution [Verbeek and Kroonenberg (1985)]. When the number of genes represented in the sample is large, independent Monte Carlo sampling for tests of genetic equilibrium is cumbersome [Guo and Thompson (1992)]. It may then be more efficient to use dependent samples generated by a Markov chain. Guo and Thompson (1992) suggest a Metropolis algorithm for generating samples suitable for testing Hardy–Weinberg equilibrium at a single locus. Diaconis and Sturmfels (1996) discuss a general method for constructing Markov chains on contingency tables. Finally, Kolassa and Tanner (1994) use the Gibbs sampler to sample from an approximation to the conditional distribution.

This paper develops Markov chains for estimating p-values of exact tests for Hardy–Weinberg and linkage equilibrium in multilocus data. Each Markov chain is designed so that its stationary distribution coincides with the null sampling distribution of the corresponding multidimensional contingency table conditioned on its margins. One advantage of our approach is that explicit bounds can be constructed for the variation distance of the chains from stationarity. Our analysis shows that the chains converge rapidly to their stationary distributions and require little work per step to execute. Rapid convergence is necessary for the chains to be competitive with a parallel, independent-sampling method that we also describe. The more quickly a Markov chain circulates through its state space, the lower the correlation of its sampled states and the smaller the number of required steps to achieve the same statistical precision attainable under independent sampling [Hastings (1970)].

Section 2 reviews some genetics terminology and introduces the concepts of Hardy–Weinberg and linkage equilibrium. Section 3 presents the general form of the data and the distributions to be used for conditional inference. Our description of the multilocus genotype distribution and its moments is new. In Section 4, we introduce two Markov chains for sampling from the conditional haplotype and genotype distributions and describe a related independent-sampling method. By embedding these Markov chains in a simpler uniform chain, in Section 5, we are able to deduce that the embedded chains have the correct stationary distributions for testing genetic equilibrium. In Section 6, we review how the tail probability of a strong stationary stopping time can be used to bound the variation distance of a Markov chain from stationarity. We present some new general results relating stationary times to the precision and bias of sample average estimators derived from running a chain.

In Section 7, we define for the uniform chain a strong stationary stopping time based on Matthews's (1988) strong uniform time for a sequence of random transpositions. In computing the expected value of the stopping time and bounding its tail probability, we encounter some interesting mathematics involving order statistics, embedding in Poisson processes and large-deviation estimates for sums of independent geometric random variables. Our analysis demonstrates that the haplotype chain reaches stationarity in $[(J-1)n \ln n]/2 + O(n)$ steps, where n is the sample size and J is the number of loci or factors. In contrast, we show that the genotype chain reaches stationarity in $Jn \ln n + O(n)$ steps. Section 8 discusses a practical application to real data. Our timed comparisons suggest that the Markov chain methods can be substantially more efficient than independent sampling.

2. Genetics background. Genes occur at sites, called loci, arranged sequentially along chromosomes. The variants of a gene at a locus are called alleles. It is helpful to think of loci as analogous to the factors of a contingency table and alleles at each locus as analogous to the levels of that factor. For $j = 1, \ldots, J$ and $k = 1, \ldots, K(j)$, we will write a_{jk} for allele k at locus j and p_{jk} for the relative frequency of allele a_{jk} in the population. (Note that subsequently we use "frequency" as shorthand for "relative frequency" or "proportion.") The autosomes, which include all chromosomes except the X and Y sex chromosomes, occur in homologous pairs. Thus, an individual's genotype at autosomal locus j consists of a pair of alleles a_{jk} and $a_{jk'}$ and is written as $a_{jk}/a_{jk'}$ with $k \le k'$. Alleles a_{jk} and $a_{jk'}$ are codominant if individuals with the three genotypes a_{jk}/a_{jk}, $a_{jk}/a_{jk'}$ and $a_{jk'}/a_{jk'}$ can be unambiguously distinguished. The first and last of these genotypes are homozygous, consisting of two copies of one allele; the middle genotype is heterozygous, consisting of one copy each of two alleles. We will confine our attention to autosomal loci with codominant alleles. When necessary, the terms "single-locus genotype" and "multilocus genotype" will be used to distinguish between genotypes at a single locus or at a set of loci, respectively. A haplotype consists of the alleles at a set of loci transmitted to a child

by one of his parents. Each individual naturally inherits one maternal and one paternal haplotype. It is convenient, for our purposes, to assume that the genes defining a haplotype reside at the ordered loci $1, \ldots, J$ on a single chromosome. This suggests the notation $a_{1k(1)}a_{2k(2)} \cdots a_{Jk(J)}$ for a haplotype with allele $a_{jk(j)}$ at each locus j.

A population is said to be in Hardy–Weinberg equilibrium with respect to locus j when the genotypes at that locus have population frequencies satisfying

$$\Pr(a_{jk}/a_{jk'}) = \begin{cases} 2p_{jk}\,p_{jk'}, & \text{for all } k < k', \\ (p_{jk})^2, & \text{for all } k = k'. \end{cases}$$

Under Hardy–Weinberg equilibrium, the two alleles transmitted to a random, noninbred person by his mother and father are independent and identically distributed [Cavalli-Sforza and Bodmer (1971)]. Inbreeding occurs when the parents are related. We further define Hardy–Weinberg equilibrium with respect to a set of loci to mean that the two haplotypes transmitted to a random, noninbred person by his mother and father are independent and identically distributed. Linkage equilibrium is in effect for a set of loci when haplotypes for those loci have population frequencies satisfying the independence rule

$$\Pr(a_{1k(1)}a_{2k(2)} \cdots a_{Jk(J)}) = \prod_{j=1}^{J} p_{jk(j)}.$$

The Hardy–Weinberg proportions often give a good description of observed genotype frequencies. In an infinitely large, randomly mating population, genotype frequencies reach Hardy–Weinberg equilibrium at an autosomal locus in a single generation. This mathematical result presupposes no selection, mutation or migration and identical initial allele frequencies in the two sexes [Cavalli-Sforza and Bodmer (1971)]. A population will also eventually reach linkage equilibrium under the same circumstances. However, this will take much longer than a single generation if the loci of the haplotype are closely spaced along a chromosome [Lange (1993)]. For this reason, violations of linkage equilibrium are more common than violations of Hardy–Weinberg equilibrium.

3. Distributions. For testing Hardy–Weinberg and linkage equilibrium, we will consider two types of data that can be obtained from a simple random sample of a population. The observations will consist either of haplotypes or of multilocus genotypes. We will use $\mathbf{i}$ to describe a particular arrangement of alleles comprising a distinct haplotype or multilocus genotype. Each arrangement $\mathbf{i}$ corresponds to a unique cell in a contingency table. If $n^{\mathbf{i}}$ denotes the number of observations of haplotype $\mathbf{i}$ for haplotype data or of multilocus genotype $\mathbf{i}$ for genotype data, then $n = \sum_{\mathbf{i}} n^{\mathbf{i}}$ is the total number of observations. Let $r_{jk}^{\mathbf{i}}$ be the number of copies of allele k at locus j appearing in type $\mathbf{i}$. For haplotype data, $r_{jk}^{\mathbf{i}} = 1$ or 0, depending on whether $\mathbf{i}$ possesses allele k

at locus j or not. For genotype data, $r_{jk}^{\mathbf{i}} = 2$ if $\mathbf{i}$ is homozygous for allele k at locus j; $r_{jk}^{\mathbf{i}} = 1$ if $\mathbf{i}$ is heterozygous with one copy of allele k at locus j; and $r_{jk}^{\mathbf{i}} = 0$ if $\mathbf{i}$ does not possess allele k at locus j. Clearly, $n_{jk} = \sum_{\mathbf{i}} r_{jk}^{\mathbf{i}} n^{\mathbf{i}}$ is the total number of copies of allele k at locus j in the sample. Finally, for genotype data, if we let $h^{\mathbf{i}}$ be the number of heterozygous loci in type $\mathbf{i}$, then $n_h = \sum_{\mathbf{i}} h^{\mathbf{i}} n^{\mathbf{i}}$ is the total number of heterozygous single-locus genotypes observed in the data.

Haplotype data used for testing linkage equilibrium of J loci form a standard J-way contingency table. When the conditions for genetic equilibrium noted above are met, the n haplotypes from a random sample of $n/2$ unrelated, noninbred individuals are independent. Given the genotype of an individual, his two haplotypes can sometimes be inferred from available family data [Goradia, Lange, Miller and Nadkarni (1992)]. For instance, if he has genotype a_{11}/a_{12} at locus 1 and his mother has genotype a_{11}/a_{11}, then he must have inherited allele a_{11} from his mother and allele a_{12} from his father. However, ambiguity can arise even when data is available on both parents. If the child and his parents all share the common genotype a_{11}/a_{12}, then additional information from other loci and other family members is needed to determine which allele came from which parent [Goradia, Lange, Miller and Nadkarni (1992)]. Sperm-typing can also be used to identify haplotypes in males [Lazzeroni, Arnheim, Schmitt and Lange (1994)].

Under linkage equilibrium, the probability of haplotype $\mathbf{i}$ is

$$p^{\mathbf{i}} = \prod_j \prod_k (p_{jk})^{r_{jk}^{\mathbf{i}}},$$

and the cell counts $\{n^{\mathbf{i}}\}$ follow a multinomial distribution with parameters $(n, \{p^{\mathbf{i}}\})$. The marginal allele totals $\{n_{jk}\}$ at any locus j likewise follow a multinomial distribution with parameters $(n, \{p_{jk}\})$. These marginal totals are independent from locus to locus. Conditional on the observed allele totals, the distribution of the cell counts is

$$\begin{aligned} \Pr\left(\{n^{\mathbf{i}}\} | \{n_{jk}\}\right) &= \frac{\binom{n}{\{n^{\mathbf{i}}\}} \prod_{\mathbf{i}} \left[\prod_j \prod_k (p_{jk})^{r_{jk}^{\mathbf{i}}}\right]^{n^{\mathbf{i}}}}{\prod_j \binom{n}{\{n_{jk}\}} \prod_k (p_{jk})^{n_{jk}}} \\ &= \frac{\binom{n}{\{n^{\mathbf{i}}\}}}{\prod_j \binom{n}{\{n_{jk}\}}} \end{aligned} \tag{1}$$

and does not depend on the unknown population allele frequencies. Lange (1993) demonstrates how to compute the moments of this generalization of the multivariate hypergeometric or Fisher–Yates distribution.

When the data consists of multilocus genotypes and information on surrounding family members is absent, the maternal or paternal origins of an

individual's alleles at heterozygous loci cannot be identified, and it is impossible to determine the underlying haplotypes. The resulting collapsing of genotype counts can best be illustrated in the single-locus setting. The complete data for a single locus could be arranged, if it were available, to form a square contingency table with columns and rows labeled, respectively, by the alleles transmitted by the mothers and the alleles transmitted by the fathers. If, as is ordinarily assumed, the two sexes have equal allele frequencies, the marginal probabilities of row j and column j would be the same. In contrast the observed genotype data are not ordered by parent and form an upper triangular table derived from the complete table by folding it along its main diagonal. Each off-diagonal count in this new table is the sum of two off-diagonal counts in the complete table.

Given Hardy–Weinberg and linkage equilibrium at a set of loci, the probability of multilocus genotype $\mathbf{i}$ is

$$p^{\mathbf{i}} = 2^{h^{\mathbf{i}}} \prod_j \prod_k (p_{jk})^{r_{jk}^{\mathbf{i}}}. \tag{2}$$

The multilocus genotype probability (2) incorporates a factor of 2 for each heterozygous single-locus genotype encountered. Consistent with independence, $p^{\mathbf{i}}$ reduces to the product of the probabilities of the constituent single-locus genotypes. As before, allele totals at the various loci follow independent multinomial distributions. Conditional upon these observed allele totals, the distribution for the multilocus genotypes of a random sample from a population in genetic equilibrium is

$$\begin{aligned} \Pr\bigl(\{n^{\mathbf{i}}\}|\{n_{jk}\}\bigr) &= \frac{\binom{n}{\{n^{\mathbf{i}}\}} \prod_{\mathbf{i}} \bigl[2^{h^{\mathbf{i}}} \prod_j \prod_k (p_{jk})^{r_{jk}^{\mathbf{i}}}\bigr]^{n^{\mathbf{i}}}}{\prod_j \binom{2n}{\{n_{jk}\}} \prod_k (p_{jk})^{n_{jk}}} \\ &= \frac{\binom{n}{\{n^{\mathbf{i}}\}} 2^{n_h}}{\prod_j \binom{2n}{\{n_{jk}\}}}. \end{aligned} \tag{3}$$

This generalizes the single-locus distribution first described by Levene (1949). It is noteworthy that allele frequencies again disappear in the conditional distribution.

To compute moments under the distribution (3), let

$$u^{\underline{m}} = u(u-1)\cdots(u-m+1)$$

be a falling factorial, and let $\{m^{\mathbf{i}}\}$ be a collection of nonnegative integers indexed by the genotypes $\mathbf{i}$. Setting $m = \sum_{\mathbf{i}} m^{\mathbf{i}}$, $m_{jk} = \sum_{\mathbf{i}} r_{jk}^{\mathbf{i}} m^{\mathbf{i}}$ and $m_h =$

$\sum_{\mathbf{i}} h^{\mathbf{i}} m^{\mathbf{i}}$, the falling factorial moment of the $\{n^{\mathbf{i}}\}$ corresponding to the $\{m^{\mathbf{i}}\}$ can be found by the computation

$$
\begin{aligned}
&E\left[\prod_{\mathbf{i}} (n^{\mathbf{i}})^{\underline{m^{\mathbf{i}}}}\right] \\
&\quad = \sum_{\Omega} \prod_{\mathbf{i}} (n^{\mathbf{i}})^{\underline{m^{\mathbf{i}}}} \frac{n!}{\prod_{\mathbf{i}} n^{\mathbf{i}}!} \frac{\prod_j \prod_k n_{jk}!}{[(2n)!]^J} 2^{n_h} \\
&\quad = \frac{n^{\underline{m}} \prod_j \prod_k (n_{jk})^{\underline{m_{jk}}} 2^{m_h}}{\left[(2n)^{\underline{2m}}\right]^J} \sum_{\Omega^*} \frac{(n-m)!}{\prod_{\mathbf{i}} (n^{\mathbf{i}} - m^{\mathbf{i}})!} \frac{\prod_j \prod_k (n_{jk} - m_{jk})!}{[(2n-2m)!]^J} 2^{n_h - m_h} \qquad (4) \\
&\quad = \frac{n^{\underline{m}} \prod_j \prod_k (n_{jk})^{\underline{m_{jk}}} 2^{m_h}}{\left[(2n)^{\underline{2m}}\right]^J},
\end{aligned}
$$

where Ω is the sample space and Ω^* is that subspace of Ω with $n^{\mathbf{i}} \geq m^{\mathbf{i}}$ for all $\mathbf{i}$. Equation (4) follows since $\prod_{\mathbf{i}} (n^{\mathbf{i}})^{\underline{m^{\mathbf{i}}}} \equiv 0$ on the complement of Ω^*, and the summation over Ω^* involves all probabilities of the multilocus genotype distribution with marginal allele counts $\{n_{jk} - m_{jk}\}$.

It is interesting that expected genotype counts under the conditional distribution differ from those estimated under the multinomial distribution using the standard maximum likelihood estimates $\{n_{jk}/2n\}$ of the allele frequencies. For a single locus, we can drop locus subscripts and compute

$$
E(n^{a_k/a_{k'}}) = \begin{cases} n \dfrac{n_k(n_k - 1)}{2n(2n-1)}, & k' = k, \\ 2n \dfrac{n_k n_{k'}}{2n(2n-1)}, & k' \neq k. \end{cases}
$$

Thus, fewer homozygotes and more heterozygotes are expected under the conditional distribution than under the estimated multinomial distribution. In contrast, the conditional expected haplotype counts agree with the estimates obtained from the multinomial distribution.

4. Markov chains for sampling. To estimate the p-value of an exact hypothesis test in this setting, one can run a Markov chain whose limiting distribution coincides with the conditional null distribution of the cell counts. The estimated p-value is simply the proportion of states encountered by the chain for which the value of the test statistic exceeds or matches its observed value. Following Besag and Clifford (1989), the same transition matrix can be used to construct an alternative Markov chain whose states under the null hypothesis are exchangeable with the observed table. This yields an exact p-value that is analogous to that proposed by Barnard (1963) for Monte Carlo tests based on an independent sample. In the sequel we adopt the first approach. However, the second implementation, while not equivalent to the first, is equally viable.

The states of our chains are tables of cell counts consistent with the observed marginal allele totals. For the haplotype chain, we consider the following transition mechanism. At each step, one of the loci (or factors) $j \in \{1, \ldots, J-1\}$ is selected with probability $1/(J-1)$. Two parents are then selected randomly with replacement from the n chromosomes in the current table. One child is created from the first parent's genes at all loci $j' \neq j$ and from the second parent's gene at locus j. The second child receives the remaining gene at each locus. The parents are then replaced by their children, and the cell counts corresponding to the parent and child haplotypes are updated accordingly. All other cell counts are unchanged. In effect, the alleles (or levels) assigned to the two parents have been exchanged at the selected locus. If at a given step the same parent is selected twice, the Markov chain remains in place.

We can visualize the haplotype chain by arranging the n initial chromosomes in the data to form the rows of an $n \times J$ rectangular tableau, not to be confused with the table of cell counts of the chain. Thus, the allele at locus j of chromosome i appears in column j of row i of the tableau. The transition mechanism of the haplotype chain can be rephrased by selecting a column (locus) at random from the first $J-1$ columns of the tableau. The last column is left intact. After a column has been selected, two rows (chromosomes) are selected at random with replacement. Then the corresponding two alleles in the selected column are exchanged. If the same row is selected twice, this exchange leaves the tableau unchanged. To each rearrangement of the tableau, there corresponds a contingency table with the appropriate marginal counts. This table is created by counting for each of its cells the number of rows in the tableau with the appropriate haplotype.

In contrast, at each step of the genotype chain, one of the loci j in $\{1, \ldots, J\}$ is selected with probability $1/J$. Two parents are selected randomly with replacement from the n individuals in the current table. If the parents are distinct, then one gene is randomly extracted from the genotype at locus j of each parent. At each locus except j, the first child receives the genotype of the first parent, and the second child receives the genotype of the second parent. At locus j, the two selected genes are exchanged so that the first child receives the unselected gene of the first parent and the selected gene of the second parent. The second child receives the remaining two genes at locus j. The parents are replaced by their children and the corresponding cell counts updated. For example, at a single locus, two parents of genotype $a_{jk}/a_{jk'}$ can be removed and replaced by one child each of genotypes a_{jk}/a_{jk} and $a_{jk'}/a_{jk'}$. The count corresponding to the common parental genotype is decremented by 2 and both child genotypes counts are incremented by 1. If at a given step the same parent is selected twice, there is again no movement of the chain.

The genotype chain can be described in terms of a tableau with $2n$ rows and J columns created two adjacent rows at a time by laying down successively a pair of chromosomes for each of the n people of the genotype chain. The transition mechanism of the genotype chain is almost identical to the

transition mechanism of the haplotype chain. The only differences now are that the number of rows is doubled and that sampling from column J is permitted. The last column is included as a choice because rows are paired, and alleles originally paired at the last locus must be scrambled. When double sampling of a person occurs, the underlying exchange within the two rows produces no detectable change at the genotype level. For each arrangement of the tableau, the corresponding contingency table is constructed by taking each pair of rows, reading off a multilocus genotype and then counting the multilocus genotypes of the various kinds.

The haplotype and genotype chains can be contrasted with Markov chains proposed for single-locus genotype data by Guo and Thompson (1992) and for standard contingency tables by Diaconis and Sturmfels (1996). At any given step of these chains, cells are selected to participate in a transition with equal probability. As a consequence, these chains have a uniform limiting distribution on the set of tables consistent with the marginal totals. Other limiting distributions can be obtained by adding a Metropolis decision rule that accepts only some of the proposed transitions. In computer implementation of these chains, operations are defined in terms of the cells of the table, and memory requirements depend on the number of cells. In the haplotype and genotype chains used in this paper, cells are selected with probabilities proportional to the current cell counts. As we will show, the limiting distributions are given by the Fisher–Yates distributions (1) and (3), respectively. Operations are defined in terms of the observations, and memory requirements depend on the sample size. This is advantageous for sparse tables.

The tableau formulation of the haplotype and genotype chains can also be used in an obvious way to generate independent Monte Carlo samples from either distribution (1) or (3). For haplotypes, the n genes at each locus can be randomly permuted using a standard method [Nijenhuis and Wilf (1978)] carried out independently at each of the first $J - 1$ loci. This requires $(J - 1)(n - 1)$ random choices to generate an independent tableau. For genotypes, the $2n$ genes at each of the J loci must be randomly and independently permuted, requiring $J(2n - 1)$ random choices per independent tableau. Given the properties of the embedding map described in Section 5, these independently sampled tableaus generate independently sampled contingency tables with the correct distributions. This procedure extends the technique described by Boyett (1979) for two-way tables and by Guo and Thompson (1992) for single-locus genotypes.

5. Embedding of Markov chains. The limiting distributions of the haplotype and genotype chains are appropriate for permutation tests of genetic equilibrium. In this section, we show that these limiting distributions are the same as the null sampling distributions of the cell counts conditional on the marginal allele totals described in Section 3. Our arguments will make it clear that the independent Monte Carlo algorithms also correctly sample these null distributions.

To demonstrate that the haplotype and genotype chains have the limiting Fisher–Yates distributions (1) and (3), respectively, it suffices to show that each chain is irreducible, aperiodic and satisfies the detailed balance condition

$$\mu_u q_{uv} = \mu_v q_{vu}$$

for all states u and v [Kelly (1979)]. Here μ is the required stationary distribution and q_{uv} is the transition probability from u to v. We will address these issues indirectly by embedding each chain in a "uniform" chain having the uniform distribution as its unique stationary distribution. Both the haplotype and genotype chains will be embedded in such a way that their stationary distributions can be obtained by simple counting arguments.

One Markov chain can be embedded in another by constructing a map $f: C \to C'$ from the state space C of the original chain onto the state space C' of the embedded chain. This map partitions the states C into equivalence classes under the equivalence relation $x \sim y$ when $f(x) = f(y)$. If $R = (r_{uv})$ denotes the matrix of transition probabilities of the original chain, then it is natural to define the transition probabilities of the embedded chain by

$$q_{f(u)f(v)} = \sum_{w \sim v} r_{uw}.$$

For the embedding to be probabilistically consistent, it is necessary that

$$\sum_{w \sim v} r_{uw} = \sum_{w \sim v} r_{xw} \tag{5}$$

for all $x \sim u$. A distribution ν on the original chain induces a distribution μ on the embedded chain according to

$$\mu_{f(u)} = \sum_{w \sim u} \nu_w. \tag{6}$$

Mindful of these conventions, we have the following general results.

PROPOSITION 1. *The embedded Markov chain is irreducible if the original Markov chain is irreducible and aperiodic if the original chain is aperiodic. If the original chain is reversible with stationary distribution ν, then the embedded chain is reversible with induced stationary distribution μ given by* (6).

PROOF. To verify irreducibility, note that if an s-step transition probability satisfies $r_{uv}^{(s)} > 0$ in the original chain, then $q_{f(u)f(v)}^{(s)} > 0$ in the embedded chain as well. Aperiodicity follows by the same argument since if the greatest common divisor of the set $\{s: r_{uu}^{(s)} > 0\}$ is 1, then the greatest common divisor of the set $\{s: q_{f(u)f(u)}^{(s)} > 0\}$ is also 1. Finally, if ν satisfies detailed balance,

then the computation

$$\begin{aligned}\mu_{f(u)}q_{f(u)f(v)} &= \Big(\sum_{w\sim u}\nu_w\Big)\Big(\sum_{x\sim v}q_{ux}\Big)\\ &= \sum_{w\sim u}\nu_w\sum_{x\sim v}q_{wx}\\ &= \sum_{w\sim u}\sum_{x\sim v}\nu_x q_{xw}\\ &= \sum_{x\sim v}\nu_x\sum_{w\sim u}q_{xw}\\ &= \mu_{f(v)}q_{f(v)f(u)}\end{aligned}$$

shows that μ also satisfies detailed balance. □

Corresponding to the haplotype chain, we can define a uniform chain by pretending that a distinct label is attached to each of the Jn genes of the chain. This change makes all existing chromosomes unique and forces all cell counts of the haplotype table to equal either 0 or 1. Under the conditional haplotype distribution, it follows that each state $\{n^{\mathbf{i}}\}$ of the uniform chain has the same probability

$$\frac{\binom{n}{\{n^{\mathbf{i}}\}}}{\prod_j\binom{n}{\{n_{jk}\}}} = (n!)^{1-J}.$$

In the tableau formulation of the uniform chain, each transition involves a random transposition of two of the n distinct genes within one of the first $J-1$ columns. Verification of irreducibility, aperiodicity and reversibility is trivial for the uniform chain. Irreducibility follows because it is possible to achieve any combination of permutations of the genes within the first $J-1$ columns of the tableau by an appropriate sequence of exchanges. Aperiodicity follows because the chain remains in place whenever the same row is selected twice for an exchange. Finally, the uniform chain satisfies detailed balance for the uniform distribution because the transition probabilities $r_{uv} = r_{vu}$ are symmetric. This follows because whenever two parental chromosomes have been mated and replaced, then mating the resulting child chromosome at the same locus restores the chain to its previous state.

For the haplotype chain, the initial row position of a gene in the tableau constitutes the unique label that follows the gene as it is transferred from one row to another while remaining within the same column. The embedding map f aggregates and counts chromosomes sharing common haplotypes based on the allele types, not the unique labels. To check that the haplotype chain is consistently embedded in the uniform chain, suppose that $x \sim u$ in the

uniform chain. Because an outcome in the haplotype chain corresponding to a transition in the uniform chain is completely determined by the selected locus and the haplotypes of the selected pair of parental chromosomes, the number of such parental pairs in the uniform chain leading to a set of counts $\{n^{\mathbf{i}}\}$ is the same whether we start from state x or state u. Thus, the consistency condition (5) holds.

The counting argument for recovering the stationary distribution of the haplotype chain from the stationary distribution of the uniform chain is equally straightforward. Consider a set of counts $\{n^{\mathbf{i}}\}$ in the haplotype chain. These haplotypes can be assigned to rows in the uniform tableau in $\binom{n}{\{n^{\mathbf{i}}\}}$ ways. Within each such assignment, there are $\prod_{j=1}^{J}\prod_k n_{jk}!$ permutations of the genes of the various allele types among the available positions for each allele type. Because we consider only those arrangements satisfying the fixed arrangement of genes at locus J, there are a total of

$$\frac{\binom{n}{\{n^{\mathbf{i}}\}}\prod_{j=1}^{J}\prod_k n_{jk}!}{n!}$$

states of the uniform chain corresponding to the state $\{n^{\mathbf{i}}\}$ of the haplotype chain. Thus, we recover the conditional haplotype distribution

$$\mu_{\{n^{\mathbf{i}}\}} = \frac{\binom{n}{\{n^{\mathbf{i}}\}}\prod_{j=1}^{J}\prod_k n_{jk}!}{n!(n!)^{J-1}} = \frac{\binom{n}{\{n^{\mathbf{i}}\}}}{\prod_{j=1}^{J}\binom{n}{\{n_{jk}\}}}.$$

The genotype chain can be embedded similarly in a uniform chain in which the initial row number of a gene again acts as a label that follows the gene as it is transferred from row to row. Each transition consists of the random transposition of two of the $2n$ genes within one of the J columns. Now the embedding map f aggregates and counts individuals sharing common genotypes based on allele types rather than unique labels. A consistent embedding is achieved because two equivalent states in the uniform chain always involve the same number of potential selections leading to a given set of genotype counts. To recover the stationary distribution of the genotype chain, consider a typical set of genotype counts $\{n^{\mathbf{i}}\}$. These genotypes can be assigned to pairs of rows in the uniform tableau in $\binom{n}{\{n^{\mathbf{i}}\}}$ ways. Given the genotype assigned to a pair of rows, the two alleles at each heterozygous locus can be assigned to the two available rows in two ways. Homozygous loci entail no such choice. If there are n_h heterozygotes among the n original genotypes, then there are $\binom{n}{\{n^{\mathbf{i}}\}}2^{n_h}$ ways of assigning the $2n$ rows in the tableau. Finally, within each such assignment, there are $\prod_{j=1}^{J}\prod_k n_{jk}!$ permutations of the genes of the

various types among the available positions for each allele type. It follows that the stationary distribution is correctly given by

$$\mu_{\{n^{\mathbf{i}}\}} = \frac{\binom{n}{\{n^{\mathbf{i}}\}} 2^{n_h} \prod_{j=1}^{J} \prod_k n_{jk}!}{[(2n)!]^J}$$

$$= \frac{\binom{n}{\{n^{\mathbf{i}}\}} 2^{n_h}}{\prod_{j=1}^{J} \binom{2n}{\{n_{jk}\}}}.$$

6. Convergence rates and stationary times. We now estimate how fast the haplotype and genotype chains approach their stationary distributions from any arbitrary state using a strong stationary stopping time. Let μ^s be the distribution of either the embedded haplotype or genotype chain at step s, and let μ be its stationary distribution. Define ν^s and ν similarly for the corresponding uniform chain. One way of quantifying convergence of the embedded chain is to use the variation distance between μ^s and μ. This is defined by any of three equivalent expressions

$$\begin{aligned} \|\mu^s - \mu\| &= \tfrac{1}{2} \sum_w |\mu^s(w) - \mu(w)| \\ &= \tfrac{1}{2} \sup_{\|g\|=1} \left| \int g(w)\, d\mu^s(w) - \int g(w)\, d\mu(w) \right| \\ &= \sup_D |\mu^s(D) - \mu(D)|, \end{aligned} \tag{7}$$

where w is any state of the chain, $g(w)$ is any real-valued function satisfying the stated equality, $\|g\|$ is the supremum of $|g(w)|$ over all w and D is any subset of states [Aldous and Diaconis (1986)]. We define the variation distance $\|\nu^s - \nu\|$ of the uniform chain similarly.

Because the original Markov chain is embedded in the uniform chain, the third definition of variation distance in (7) implies

$$\|\mu^s - \mu\| \le \|\nu^s - \nu\|.$$

Furthermore, $\|\nu^s - \nu\|$ can be bounded by defining a strong stationary stopping time U for the uniform Markov chain. Let $\mathbf{W} = (w_1, w_2, \dots)$ be a sample path of the uniform chain. A stopping time U has the property that if $U(\mathbf{W}) = s$, then $U(\mathbf{W}^*) = s$ for all other sample paths $\mathbf{W}^* = (w_1^*, w_2^*, \dots)$ satisfying $w_1^* = w_1, \dots, w_s^* = w_s$. If in addition

$$\Pr(\mathbf{W}_s = w_s | U \le s) = \nu(w_s),$$

then U is said to be strongly stationary. It is straightforward to show [Diaconis (1988)] that

$$\|\nu^s - \nu\| \le \Pr(U > s). \tag{8}$$

Before constructing specific strong stationary times, we establish their value in bounding the precision and bias of sample mean estimators.

PROPOSITION 2. *Consider a stationary Markov chain* $\{\mathbf{W}_r\}$ *having stationary distribution* μ *and equipped with a strong stationary time* U *such that*

$$\Pr(\mathbf{W}_r = w_r | U \le r, \mathbf{W}_0 = w_0) = \mu(w_r) \tag{9}$$

and such that

$$\Pr(U = r | \mathbf{W}_0 = w_0) \tag{10}$$

does not depend on w_0. *If the random sequences* $X_r = g(\mathbf{W}_r)$ *and* $Y_r = h(\mathbf{W}_r)$ *have finite variances* σ_x^2 *and* σ_y^2, *respectively, then the correlation between* X_r *and* Y_{r+s} *satisfies*

$$\left|\operatorname{Corr}(X_r, Y_{r+s})\right| \le \Pr(U > s) \tag{11}$$

for all $s \ge 0$. *Hence, the variance of the sample mean* $S_m = (1/m)\sum_{r=0}^{m-1} X_r$ *satisfies*

$$\operatorname{Var}(S_m) \le \frac{\sigma_x^2}{m}\left[2E(U) - 1\right]. \tag{12}$$

PROOF. By the stationarity assumption, it suffices to take $r = 0$ in (11). If $Z = 1_{\{U > s\}}$, it follows from the identity

$$\begin{aligned}\mu(w_s) &= \Pr(\mathbf{W}_s = w_s | Z = 0)\Pr(Z = 0) + \Pr(\mathbf{W}_s = w_s | Z = 1)\Pr(Z = 1)\\ &= \mu(w_s) + \left[\Pr(\mathbf{W}_s = w_s | Z = 1) - \mu(w_s)\right]\Pr(Z = 1)\end{aligned}$$

that $\Pr(\mathbf{W}_s = w_s | Z = 1) = \mu(w_s)$ whenever $\Pr(Z = 1) > 0$. Conditioning on Z therefore yields

$$\begin{aligned}\operatorname{Cov}(X_0, Y_s) &= \operatorname{Cov}\left[E(X_0|Z), E(Y_s|Z)\right] + E\left[\operatorname{Cov}(X_0, Y_s|Z)\right]\\ &= \operatorname{Cov}(X_0, Y_s | Z = 1)\Pr(Z = 1)\end{aligned}$$

by virtue of property (9). The correlation inequality (11) then follows from

$$\begin{aligned}\left|\operatorname{Cov}(X_0, Y_s | Z = 1)\right| &\le \operatorname{Var}(X_0 | Z = 1)^{1/2}\operatorname{Var}(Y_s | Z = 1)^{1/2}\\ &= \sigma_x \sigma_y.\end{aligned}$$

Inequality (11) then implies that

$$\begin{aligned}\operatorname{Var}(S_m) &= \frac{1}{m^2}\left[\sum_{r=0}^{m-1}\operatorname{Var}(X_r) + 2\sum_{s=1}^{m-1}\sum_{r=0}^{m-1-s}\operatorname{Cov}(X_r, X_{r+s})\right]\\ &\le \frac{\sigma^2}{m}\left[1 + 2\sum_{s+1}^{m-1}\left(1 - \frac{s}{m}\right)\Pr(U > s)\right]\\ &= \frac{\sigma^2}{m}\left[2\sum_{s=0}^{m-1}\Pr(U > s) - 1 - \frac{2}{m}\sum_{s=1}^{m-1} s\Pr(U > s)\right].\end{aligned}$$

The variance bound (12) now follows from the well-known identity [Feller (1968) pages 265–266]

$$E(U) = \sum_{s=0}^{\infty} \Pr(U > s). \qquad \square$$

Because a strong stationary time is typically defined without reference to the initial state of the Markov chain, properties (9) and (10) are usually satisfied. When the chain is stationary as required in Proposition 2, S_m is an unbiased estimator of $\int g(w)\, d\mu(w)$. In the next proposition we drop the assumption of stationarity and investigate the bias of S_m.

PROPOSITION 3. *Let $\{\mathbf{W}_r\}$ be a Markov chain with strong stationary time U and stationary distribution μ. Then the absolute bias of the estimator $S_m = (1/m)\sum_{r=0}^{m-1} 1_D(\mathbf{W}_r)$ of $\mu(D)$ is bounded above by $E(U)/m$. If we replace $1_D(w)$ by an arbitrary function $g(w)$ with $\|g(w)\| = c$, then the absolute bias of S_m is bounded above by $2cE(U)/m$.*

PROOF. Let μ_r be the distribution of $\mathbf{W}_r$. In view of inequality (8) and the third definition of variation distance in (7), the bias of S_m satisfies

$$\begin{aligned} |\mathrm{Bias}(S_m)| &\le \frac{1}{m} \sum_{r=0}^{m-1} |\mu_r(D) - \mu(D)| \\ &\le \frac{1}{m} \sum_{r=0}^{m-1} \Pr(U > r) \\ &\le \frac{E(U)}{m}. \end{aligned}$$

The claimed bias inequality for an arbitrary bounded function $g(w)$ follows in similar manner from the second definition of variation distance in (7). $\square$

7. Stationary times for the Markov chains. For the haplotype chain, we will prove that the expected value of the stationary time U for the underlying uniform chain satisfies the explicit bound

$$\begin{aligned} E(U) \le (J-1)\Bigg[&\frac{n-1}{2}(\ln n + g) - \frac{1}{4} \\ &+ \frac{J-2}{\sqrt{2J-3}} \sqrt{\frac{\pi^2 n^2}{24} + \frac{n \ln n}{2} + hn} \Bigg] \\ &+ O\left(\frac{\ln n}{n}\right), \end{aligned} \tag{13}$$

where $\gamma \approx 0.577$ is Euler's constant, $g = \ln 2 + \gamma + 1 \approx 2.27$ and

$$h = \frac{\ln 2 + \gamma}{2} + \frac{11 - \pi^2}{12} \approx 0.73.$$

The right tail probability of U exhibits the sharp cutoff behavior

$$\begin{aligned} &\Pr\left[U \geq (J-1)n\left(c + \frac{\ln n + g}{2}\right)\right] \\ &\quad \leq 1.81(J-1)\exp\left(-\frac{6c}{5}\right)\left[1 + O\left(\frac{\ln n}{n}\right)\right] \end{aligned} \tag{14}$$

for every constant $c > 0$. For the genotype chain, we have the corresponding bounds

$$\begin{aligned} E(U) \leq J\Bigg[&\left(n - \frac{1}{2}\right)(\ln n + g') - \frac{1}{4} \\ &+ \frac{J-1}{\sqrt{2J-1}}\sqrt{\frac{\pi^2 n^2}{6} + n\ln n + h'n}\Bigg] \\ &+ O\left(\frac{\ln n}{n}\right) \end{aligned} \tag{15}$$

and

$$\begin{aligned} &\Pr\left[U \geq 2Jn\left(c + \frac{\ln n + g'}{2}\right)\right] \\ &\quad \leq 1.81J\exp\left(-\frac{6c}{5}\right)\left[1 + O\left(\frac{\ln n}{n}\right)\right]. \end{aligned} \tag{16}$$

Here, $g' = g + \ln 2 \approx 2.96$ and $h' = h + \ln 2 \approx 1.42$.

To define a strong stationary stopping time for the uniform chain, we imagine placing a check mark on one of the genes during certain transitions. For the sake of concreteness, we consider the uniform chain corresponding to the haplotype chain. Obvious modifications of our arguments work for the genotype chain. At a transition involving locus j, a gene in column j is checked if certain conditions are met. Let U_j be the step when the last gene at locus j is checked. Then $U = \max_{j<J} U_j$ defines a stopping time. We borrow rules for checking genes from Matthews (1988), who defines a strong stationary time for a random permutation of n objects generated by a sequence of random transpositions. Which of Matthews's rules are in effect depends on the number of genes i previously checked at a locus. Let $\lceil x \rceil$ be the least integer greater than or equal to x. When $i < \lceil n/3 \rceil$ and both the first and the second gene currently selected at locus j are unchecked, then the first gene is checked. Once $\lceil n/3 \rceil$ genes have been checked, a second rule with two subrules prevails. First, if an unchecked gene is selected for exchange with itself or with a previously checked gene, then the unchecked

gene is checked. If checking does not take place by this mechanism, it can occur by an alternative one. Arrange all unchecked genes and all ordered pairs of checked genes in two lists. The first list has $n - i$ elements and the second list i^2 elements. If the kth pair of genes from the second list is selected, then check the kth gene from the first list. Note in this regard that $i^2 \geq n - i$ provided $n > 3$ and $i \geq \lceil n/3 \rceil$. With i genes currently checked at locus j, Matthews's rules imply that an additional gene at locus j will be checked with probability

$$p_i = \begin{cases} \dfrac{(n-i)^2}{(J-1)n^2}, & \text{for } i < \left\lceil \dfrac{n}{3} \right\rceil, \\ \dfrac{2(i+1)(n-i)}{(J-1)n^2}, & \text{for } i \geq \left\lceil \dfrac{n}{3} \right\rceil. \end{cases}$$

Matthews (1988) shows that when the last gene has been checked at locus j according to these rules, then the permutation of the labels in column j will be uniformly distributed with respect to the labels in the last column J. Subsequent random transpositions at locus j or at other loci do not alter this fact. Although the gene-checking times at the various loci are dependent, the resulting permutations at the loci are independent. At the last checking step U, all arrangements of the genes at loci $1, \ldots, J-1$ with respect to the genes at the last locus are equally likely, and U is a strong stationary time.

We now examine the stochastic behavior of the strong stationary stopping time for the uniform chain. Let U_{ji} be the number of steps after gene i at locus j is checked until gene $i+1$ at locus j is checked. Then $U_j = \sum_{i=0}^{n-1} U_{ji}$ defines the total waiting time until all genes are checked at locus j. The intralocus checking times U_{ji} are independent and geometrically distributed with success probability p_i. If $q_i = 1 - p_i$, then U_{ji} has expected value $1/p_i$ and variance q_i/p_i^2.

An upper bound to the expected value of $U = \max_{j<J} U_j$,

$$E(U) \leq E(U_1) + \sqrt{(J-2)\mathrm{Var}(U_1)}, \tag{17}$$

is available from the theory of order statistics for possibly dependent variables [David (1981)]. If the U_j were independent, continuous random variables, the superior bound

$$E(U) \leq E(U_1) + (J-2)\sqrt{\frac{\mathrm{Var}(U_1)}{2J-3}} \tag{18}$$

would apply [David (1981)]. One can construct an independent set of closely related waiting times by embedding the uniform chain in a Poisson process in such a way that the events of the Poisson process correspond to the steps of the chain [Blom and Holst (1991)]. The Poisson process can be constructed by first constructing $J-1$ independent Poisson processes having common intensity $1/(J-1)$. Each event in process j generates an exchange at locus j.

The superposition process formed by considering the events over all $J-1$ loci gives the requisite Poisson process with unit intensity. If Y_j is the waiting time in the superposition process until the completion of checking at locus j, then the Y_j are independent and identically distributed. Furthermore, $Y_j = \sum_{i=1}^{U_j} X_i$, where X_i is the exponential waiting time from event $i-1$ to event i of the superposition process. By construction $E(X_i) = \mathrm{Var}(X_i) = 1$. Exploiting the correspondence between the step $U = \max_{j<J} U_j$ and the time $Y = \max_{j<J} Y_j$, we have

$$E(Y) = E\left[E\left(\sum_{i=1}^{U} X_i \middle| U\right)\right] = E(U).$$

Since the Y_j are independent and identically distributed, this allows us to invoke inequality (18) with Y random variables replacing U random variables. However,

$$\begin{aligned} E(Y_j) &= E(U_j) \\ &= \sum_{i=0}^{n-1} \frac{1}{p_i} \\ &= (J-1)\left[\frac{n-1}{2}(\ln n + g) - \frac{1}{4}\right] + O\left(\frac{\ln n}{n}\right), \end{aligned}$$

using the asymptotic value (25) from the Appendix. Similarly,

$$\begin{aligned} \mathrm{Var}(Y_j) &= E\left[\mathrm{Var}(Y_j|U_j)\right] + \mathrm{Var}\left[E(Y_j|U_j)\right] \\ &= E(U_j) + \mathrm{Var}(U_j) \\ &= \sum_{i=0}^{n-1} \frac{1}{p_i^2} \\ &= (J-1)^2\left(\frac{\pi^2 n^2}{24} + \frac{n \ln n}{2} + hn\right) + O(\ln n), \end{aligned}$$

using the asymptotic value (26) from the Appendix. Combining these results with inequality (18) yields inequality (13), which to order $O(\ln n)$ can be rewritten as

$$E(U) \le (J-1)\left[\frac{n}{2}(\ln n + g) + \frac{J-2}{\sqrt{2J-3}} \frac{\pi n}{\sqrt{24}}\right] + O(\ln n).$$

In contrast, combining the same asymptotic values with inequality (17) leads to the inferior bound

$$E(U) \le (J-1)\left[\frac{n}{2}(\ln n + g) + \sqrt{J-2} \frac{\pi n}{\sqrt{24}}\right] + O(\ln n).$$

Thus, Poissonization of the U_j results in a second-order correction that can be substantial for moderately sized problems.

To find the tail probability of U we develop some large-deviation estimates for sums of geometrically distributed random variables. Our point of departure is the moment-generating function

$$\sum_{j=1}^{\infty} p_i q_i^{j-1} \exp(jt) = \frac{p_i \exp(t)}{1 - q_i \exp(t)}$$

of the geometric random variable U_{ji}. Applying Bernstein's inequality [Sen and Singer (1993)], we find the tail probability bound

$$\begin{aligned}
&\Pr\bigl[U_j \geq c(J-1)n + E(U_j)\bigr] \\
&\quad \leq \exp\bigl[-c(J-1)nt\bigr] \prod_{i=0}^{n-1} \exp\bigl[-tE(U_{ji})\bigr] \frac{p_i \exp(t)}{1 - q_i \exp(t)} \\
&\quad = \exp\bigl[-c(J-1)nt\bigr] \prod_{i=0}^{n-1} \frac{p_i \exp\bigl[t(1 - 1/p_i)\bigr]}{p_i - q_i\bigl[\exp(t) - 1\bigr]} \\
&\quad = \exp\bigl[-c(J-1)nt\bigr] \frac{\exp\bigl(-t \sum_{i=0}^{n-1}(q_i/p_i)\bigr)}{\prod_{i=0}^{n-1}\{1 - (q_i/p_i)\bigl[\exp(t) - 1\bigr]\}}
\end{aligned}$$

for any $t \in [0,1]$ and $c > 0$. From the inequality

$$\frac{1}{1-x} \leq \exp(x + x^2) \tag{19}$$

for $x \in [0, 3/5]$, we then conclude that

$$\begin{aligned}
&\Pr\bigl[U_j \geq c(J-1)n + E(U_j)\bigr] \\
&\quad \leq \exp\bigl[-c(J-1)nt\bigr] \\
&\qquad \times \exp\left\{-t \sum_{i=0}^{n-1} \frac{q_i}{p_i} + \sum_{i=0}^{n-1} \left[\frac{q_i}{p_i}(\exp(t) - 1) + \left(\frac{q_i}{p_i}\right)^2 (\exp(t) - 1)^2\right]\right\} \qquad (20) \\
&\quad = \exp\bigl[-c(J-1)nt\bigr] \\
&\qquad \times \exp\left[(\exp(t) - 1 - t) \sum_{i=0}^{n-1} \frac{q_i}{p_i} + (\exp(t) - 1)^2 \sum_{i=0}^{n-1} \left(\frac{q_i}{p_i}\right)^2\right].
\end{aligned}$$

Inequality (19) is proved in the Appendix along with the necessary bounds

$$\frac{q_i}{p_i}(e^t - 1) \leq \frac{3}{5} \tag{21}$$

for all i and the relevant t indicated below. In view of equations (25) and (26) in the Appendix, we can estimate the two sums appearing in inequality (20) by

$$\sum_{i=1}^{n-1} \frac{q_i}{p_i} = O(n \ln n) \tag{22}$$

and

$$\begin{aligned}\sum_{i=1}^{n-1}\left(\frac{q_i}{p_i}\right)^2 &= \sum_{i=1}^{n-1}\frac{1}{p_i^2} - 2\sum_{i=1}^{n-1}\frac{1}{p_i} + \sum_{i=1}^{n-1}1 \\ &= \frac{(J-1)^2\pi^2 n^2}{24} + O(n\ln n).\end{aligned} \tag{23}$$

We next make the crucial choice $t = 6/[5(J-1)n]$. For all nontrivial cases, $J \geq 2$ and $n \geq 2$, ensuring that $t \in [0,1]$. Since

$$\exp\left[\frac{6}{5(J-1)n}\right] = 1 + \frac{6}{5(J-1)n} + O\left(\frac{1}{n^2}\right),$$

it follows from equations (20), (22) and (23) that

$$\begin{aligned}\Pr\left[U_j \geq c(J-1)n + E(U_j)\right] &\leq \exp\left(-\frac{6c}{5}\right)\exp\left[\frac{3\pi^2}{50} + O\left(\frac{\ln n}{n}\right)\right] \\ &= \exp\left(\frac{3\pi^2}{50}\right)\exp\left(-\frac{6c}{5}\right)\left[1 + O\left(\frac{\ln n}{n}\right)\right] \\ &\approx 1.81\exp\left(-\frac{6c}{5}\right)\left[1 + O\left(\frac{\ln n}{n}\right)\right].\end{aligned}$$

Because $E(U_j) = (J-1)n(\ln n + g)/2 + O(\ln n)$,

$$\begin{aligned}&\Pr\left\{U \geq (J-1)\left[cn + \frac{n}{2}(\ln n + g)\right]\right\} \\ &\quad\leq \sum_{j=1}^{J-1}\Pr\left\{U_j \geq (J-1)n\left[c + O\left(\frac{\ln n}{n}\right)\right] + E(U_j)\right\} \\ &\quad\leq 1.81(J-1)\exp\left[-\frac{6c}{5} + O\left(\frac{\ln n}{n}\right)\right]\left[1 + O\left(\frac{\ln n}{n}\right)\right]\end{aligned}$$

for $c > 0$, which is equivalent to inequality (14).

For the genotype chain, there are J columns, each of length $2n$, that must be permuted and we get similar results by substituting J and $2n$ for $J-1$ and n, respectively, in the proofs and in the results for the haplotype chain. Thus, for the genotype chain

$$\begin{aligned}E(Y_j) &= J\left[\left(n - \frac{1}{2}\right)(\ln n + \ln 2 + g) - \frac{1}{4}\right] + O\left(\frac{\ln n}{n}\right), \\ \operatorname{Var}(Y_j) &= J^2\left[\frac{\pi^2 n^2}{6} + n\ln n + (2h + \ln 2)n + O(\ln n)\right],\end{aligned}$$

and similar reasoning to the above yield inequalities (15) and (16).

8. A numerical example. We illustrate the use of the above Markov chains on chromosome 11 data collected on 24 Utah pedigrees. By considering children and grandchildren, Weir and Brooks (1986) reconstructed 8-locus haplotypes on 92 founders of these pedigrees. After deletion of two loci with a substantial number of missing observations and two individuals untyped at some remaining loci, the data consist of 180 haplotypes from 90 individuals. The six pertinent loci have 2, 2, 10, 5, 3 and 2 alleles, respectively.

Our tests are based on the χ^2 statistic

$$\begin{aligned}\chi^2 &= \sum_{\mathbf{i}} \frac{\left[n^{\mathbf{i}} - E(n^{\mathbf{i}})\right]^2}{E(n^{\mathbf{i}})} \\ &= \sum_{\mathbf{i}} \frac{(n^{\mathbf{i}})^2}{E(n^{\mathbf{i}})} - n\end{aligned} \tag{24}$$

with the proviso that expected counts $E(n^{\mathbf{i}})$ are computed under the conditional null distribution. Li (1955) first suggested using conditional expectations in the context of genotype data. Because closed-form expressions can be found for the expectation $E(\chi^2)$ (see the Appendix), the χ^2 statistic is especially valuable for evaluating the success of the various algorithms. Updating this statistic is simplified by the fact that if $m^{\mathbf{i}} = n^{\mathbf{i}} \pm a$, then the difference in the contribution of these two values of cell $\mathbf{i}$ to the χ^2 statistic is

$$\begin{aligned}\frac{(n^{\mathbf{i}})^2}{E(n^{\mathbf{i}})} - \frac{(m^{\mathbf{i}})^2}{E(n^{\mathbf{i}})} &= \frac{(n^{\mathbf{i}})^2}{E(n^{\mathbf{i}})} - \frac{(n^{\mathbf{i}} \pm a)^2}{E(n^{\mathbf{i}})} \\ &= \frac{\mp 2an^{\mathbf{i}} - a^2}{E(n^{\mathbf{i}})}.\end{aligned}$$

The p-value of an observed statistic χ^2_{obs} is by definition the probability $\Pr(\chi^2 \geq \chi^2_{\text{obs}})$ for an independent sample χ^2 from the null distribution. Fortunately p-values for the conditional null distributions (1) and (3) do not depend on unknown allele frequencies. Since a p-value is an expectation, the standard ergodic theorem for finite state Markov chains [Karlin and Taylor (1975)] justifies estimating $\Pr(\chi^2 \geq \chi^2_{\text{obs}})$ by taking a sample average of the indicator functions of the events $\chi^2 \geq \chi^2_{\text{obs}}$ over many steps of the appropriate chain.

To estimate p-values and the expectation $E(\chi^2)$, we chose the initial state of the Monte Carlo chains in two different ways. The observed Markov chain (OMC) starts from the observed table. Although this tactic yields biased estimates, bias diminishes as the total number of steps $m \to \infty$. The independent Markov chain (IMC) begins with an independent table drawn from the conditional null distribution. IMC estimates are unbiased because each table of the IMC chain is marginally distributed according to the stationary distribution. Independent Monte Carlo samples (IND) were also generated by the algorithm described in Section 4.

If Hardy–Weinberg equilibrium did not prevail for this set of loci, a test of linkage equilibrium could still be performed after randomly choosing one haplotype from each of the 90 individuals. Hardy–Weinberg equilibrium guarantees that the 180 haplotypes form an independent sample. Biologically, it is likely that most violations of Hardy–Weinberg for the set as a whole can be detected as violations of Hardy–Weinberg equilibrium for at least one locus within the set. In these data, none of the single-locus genotype χ^2 tests is significant. A test of linkage equilibrium was accordingly performed on the 180 six-locus haplotypes. The p-value of the observed $\chi^2_{\text{obs}} = 1517$ would be essentially 0 if the traditional large-sample approximation were appropriate for this sparse contingency table with $1200 = 2 \times 2 \times 10 \times 5 \times 3 \times 2$ cells. Each Monte Carlo method was independently implemented 100 times in Microsoft Fortran on a 486/66 DX2 personal computer. Each run produced one estimate $\hat{p}$ of the p-value and one estimate $\widehat{E(\chi^2)}$ of the expectation $E(\chi^2)$. To make the methods comparable in terms of computer time, the number of samples per run for each method was set so that each run would take about one minute of computing time, ignoring setup operations common to all of the methods.

Table 1 gives the average of the estimates $\hat{p}$ and $\widehat{E(\chi^2)}$ over all 100 runs, and the observed standard deviations of these estimates. The average $\hat{p}$ for all three methods is 0.13, sharply contradicting the large-sample result. The OMC method shows no obvious bias in estimating the theoretical expectation $E(\chi^2) = 1182$. The standard deviations of the estimates $\hat{p}$ and $\widehat{E(\chi^2)}$ suggest that both Markov chain methods are more efficient than independent sampling. Although the number of Markov chain iterations is much larger than the number of independent samples, this is compensated by the approximately 685 Markov chain iterations possible in the time it takes to generate a single independent sample.

If we ignore haplotype information, we can refit the same data using the genotype chain. The now much larger $3 \times 3 \times 54 \times 14 \times 6 \times 3$ contingency table has 90 individuals distributed over 122,472 cells. Both the observed statistic $\chi^2_{\text{obs}} = 134{,}823$ and its expectation $E(\chi^2) = 122{,}454$ are dramatically larger than for the haplotype table; note that we are using Li's (1955) convention for expected values in the definition of χ^2. Again the p-value of

TABLE 1
Monte Carlo results for chromosome 11 *haplotype data*

Method	Avg. $\hat{p}$	Avg. $\widehat{E(\chi^2)}$	Seconds per run	Iterations per run
OMC	0.1333 ± 0.0039	1179 ± 16	59.81	2.740×10^6
IMC	0.1337 ± 0.0035	1182 ± 14	59.87	2.740×10^6
IND	0.1332 ± 0.0057	1180 ± 22	59.98	3999

the observed statistic would be essentially 0 if the large-sample approximation were appropriate. However, inferring the presence of linkage disequilibrium, Hardy–Weinberg disequilibrium, or both for these data would be grossly misleading since all three Monte Carlo methods yield an average $\hat{p}$ of 0.14 as displayed in Table 2. Although the OMC method again shows no obvious bias in estimating $\widehat{E(\chi^2)}$, the observed standard deviations of $\hat{p}$ and $\widehat{E(\chi^2)}$ are proportionately much larger than for the haplotype chains. This is consistent with the smaller number of iterations possible per minute for the genotype chain. The genotype algorithms are slower because it is harder to recover genotype counts from the tableau than it is to recover haplotype counts. The disproportionate toll that this difficulty takes on independent sampling is clear; now 2159 Markov chain iterations are possible in the time required to generate a single independent sample. Possibly the same effect is showing up in the substantially smaller standard deviations under the Markov chain methods compared to independent sampling.

The empirical results for this example provide stronger support for the Markov chain methods than is available from the rough upper bounds provided by our theoretical analysis. For instance, the expected number of steps $E(U)$ until reaching stationarity is 4125 for the haplotype chain and 5071 for the genotype chain. Based on the numbers of iterations, given in Tables 1 and 2, Proposition 3 only ensures that the bias in the OMC-estimated p-values is no greater than 0.0015 and 0.0034, respectively. Similarly, Proposition 2 only guarantees that the ratio of the standard deviations of the IMC and IND estimators is no greater than 3.47 for equal computing times in the haplotype analysis. The empirical ratios of 0.62 for $\hat{p}$ and 0.64 for $\widehat{E(\chi^2)}$ reverse this unfavorable impression of the IMC method. For the genotype analysis, the theoretical bound on the ratio of standard deviations is 2.17, while the empirical ratios are 0.44 and 0.73, respectively.

Finally, as the current data show, some individuals will usually be untyped at some loci. We have simplified our presentation by discarding cases and loci with missing data. This is, of course, undesirable when there is substantial missing data. If we want to retain all of the data, then it is still possible to employ our sampling methods if the data are missing completely

TABLE 2
Monte Carlo results for chromosome 11 *genotype data*

Method	Avg. $\hat{p}$	Avg. $\widehat{E(\chi^2)}$	Seconds per run	Iterations per run
OMC	0.1428 ± 0.0044	$122{,}526 \pm 14{,}042$	59.75	1.490×10^6
IMC	0.1433 ± 0.0054	$122{,}299 \pm 20{,}674$	59.81	1.490×10^6
IND	0.1418 ± 0.0124	$116{,}405 \pm 28{,}464$	59.82	690

at random. The appropriate modification to the algorithms is simply to perform all transpositions or permutations only within those rows of the tableau for which the currently selected locus is observed. Of course, the definitions of the test statistics must be modified to accommodate missing data.

9. Discussion. Many recent advances in computational statistics rely on Markov chain methods such as the Gibbs sampler and the Metropolis algorithm to sample from complicated marginal distributions. Our example demonstrates that even when independent sampling methods are feasible, Markov chain algorithms can be more efficient. For this to occur, a Markov chain must either converge rapidly to its equilibrium or involve little work per step. The chains suggested here achieve an advantageous balance of these two criteria.

Besides serving as a fertile field of application, genetics has been for us a source of inspiration in designing chains useful for testing independence in multidimensional contingency tables. Genetics leads one naturally to think in terms of the arrangement of genes along a chromosome and the exchange of genes between chromosomes. This biological framework suggests the tableau incorporated in the uniform chain. Embedding the Markov chains in a uniform chain greatly simplifies our theoretical analysis. This simplification may exact a price since more complicated, but faster, stopping rules could possibly be designed that take into account the specific allele totals.

APPENDIX

A.1. Sums $\sum_{i=0}^{n-1}(1/p_i)$ and $\sum_{i=0}^{n-1}(1/p_i^2)$. In this section, we show that

$$\sum_{i=0}^{n-1} \frac{1}{p_i} = (J-1)\left[\frac{n-1}{2}(\ln n + g) - \frac{1}{4}\right] + O\left(\frac{\ln n}{n}\right) \tag{25}$$

and

$$\sum_{i=0}^{n-1} \frac{1}{p_i^2} = (J-1)^2\left(\frac{\pi^2 n^2}{24} + \frac{n \ln n}{2} + hn\right) + O(\ln n), \tag{26}$$

where $g = \ln 2 + \gamma + 1$ and $h = (\ln 2 + \gamma)/2 + (11 - \pi^2)/12$. The asymptotic expansions (25) and (26) are used in Section 7 to derive the expectation, variance and tail probability bound of the stopping time.

To evaluate the sum $\sum_{i=0}^{n-1} 1/p_i$, first choose an integer $0 \le b \le 2$ so that $(n+b)/3 = \lceil n/3 \rceil$. Then

$$\ln\left(\frac{n - b/2}{n+b}\right) = -\frac{3b}{2n} + O\left(\frac{1}{n^2}\right).$$

From the well-known asymptotic expansion of the harmonic series [Graham, Knuth and Patashnik (1989)]

$$\sum_{i=1}^{n} \frac{1}{i} = \ln n + \gamma + \frac{1}{2n} + O\left(\frac{1}{n^2}\right),$$

where $\gamma \approx 0.577$ is Euler's constant, we therefore deduce

$$\begin{aligned}
&\sum_{i=\lceil n/3 \rceil}^{n-1} \frac{1}{(i+1)(n-i)} \\
&\quad = \frac{1}{(n+1)} \sum_{i=(n+b)/3}^{n-1} \left(\frac{1}{i+1} + \frac{1}{n-i}\right) \\
&\quad = \frac{1}{(n+1)} \left(\sum_{i=1}^{n} \frac{1}{i} - \sum_{i=1}^{(n+b)/3} \frac{1}{i} + \sum_{i=1}^{(2n-b)/3} \frac{1}{i}\right) \\
&\quad = \frac{1}{n}(\ln n + d) - \frac{1}{n^2}\left(\ln n + d + \frac{1+6b}{4}\right) + O\left(\frac{\ln n}{n^3}\right),
\end{aligned} \tag{27}$$

where $d = \ln 2 + \gamma$.

Similarly, the Euler–Maclaurin expansion [Graham, Knuth and Patashnik (1989)]

$$\sum_{i=1}^{n} \frac{1}{i^2} = \frac{\pi^2}{6} - \frac{1}{n} + \frac{1}{2n^2} + O\left(\frac{1}{n^3}\right) \tag{28}$$

implies

$$\begin{aligned}
\sum_{i=1}^{(2n-b)/3} \frac{1}{i^2} &= \frac{\pi^2}{6} - \frac{3}{2n-b} + \frac{9}{2(2n-b)^2} + O\left(\frac{1}{n^3}\right) \\
&= \frac{\pi^2}{6} - \frac{3}{2n} + \frac{9-6b}{8n^2} + O\left(\frac{1}{n^3}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{i=0}^{\lceil n/3 \rceil - 1} \frac{1}{(n-i)^2} &= \sum_{i=1}^{n} \frac{1}{i^2} - \sum_{i=1}^{(2n-b)/3} \frac{1}{i^2} \\
&= \frac{1}{2n} + \frac{6b-5}{8n^2} + O\left(\frac{1}{n^3}\right).
\end{aligned} \tag{29}$$

It follows from expressions (27) and (29) together that

$$\begin{aligned}\sum_{i=0}^{n-1}\frac{1}{p_i} &= \sum_{i=0}^{\lceil n/3\rceil-1}\frac{(J-1)n^2}{(n-i)^2}+\sum_{i=\lceil n/3\rceil}^{n-1}\frac{(J-1)n^2}{2(i+1)(n-i)}\\ &= (J-1)\left[\frac{n}{2}(\ln n+d+1)-\frac{1}{2}\left(\ln n+d+\frac{3}{2}\right)\right]+O\left(\frac{\ln n}{n}\right)\\ &= (J-1)\left[\frac{n-1}{2}(\ln n+g)-\frac{1}{4}\right]+O\left(\frac{\ln n}{n}\right),\end{aligned}$$

where $g = d + 1 = \ln 2 + \gamma + 1$. This proves equality (25).

Next, we evaluate the sum $\sum_{i=0}^{n-1} 1/p_i^2$. Since expression (28) implies

$$\sum_{i=1}^{(n+b)/3}\frac{1}{i^2}=\frac{\pi^2}{6}-\frac{3}{n}+O\left(\frac{1}{n^2}\right),$$

it also follows that

$$\begin{aligned}&\sum_{i=\lceil n/3\rceil}^{n-1}\left[\frac{1}{(i+1)(n-i)}\right]^2\\ &\quad=\frac{1}{(n+1)^2}\sum_{i=(n+b)/3}^{n-1}\left(\frac{1}{i+1}+\frac{1}{n-i}\right)^2\\ &\quad=\frac{1}{(n+1)^2}\sum_{i=(n+b)/3}^{n-1}\left[\left(\frac{1}{i+1}\right)^2\right.\\ &\qquad\left.+2\left(\frac{1}{i+1}\right)\left(\frac{1}{n-i}\right)+\left(\frac{1}{n-i}\right)^2\right] \qquad (30)\\ &\quad=\frac{1}{(n+1)^2}\left[\sum_{i=1}^{n}\frac{1}{i^2}-\sum_{i=1}^{(n+b)/3}\frac{1}{i^2}+\sum_{i=1}^{(2n-b)/3}\frac{1}{i^2}\right.\\ &\qquad\left.+2\sum_{i=(n+b)/3}^{n-1}\left(\frac{1}{i+1}\right)\left(\frac{1}{n-i}\right)\right]\\ &\quad=\frac{\pi^2}{6n^2}+\frac{2}{n^3}\left(\ln n+d+\frac{1}{4}-\frac{\pi^2}{6}\right)+O\left(\frac{\ln n}{n^4}\right).\end{aligned}$$

The expansion

$$\sum_{i=1}^{n}\frac{1}{i^4}=\frac{1}{90}\pi^4-\frac{1}{3n^3}+O\left(\frac{1}{n^4}\right)$$

implies

$$\sum_{i=0}^{\lceil n/3\rceil-1}\left[\frac{1}{(n-i)^2}\right]^2 = \left(\sum_{i=1}^{n}\frac{1}{i^4} - \sum_{i=1}^{(2n-b)/3}\frac{1}{i^4}\right) = \frac{19}{24n^3} + O\left(\frac{1}{n^4}\right). \tag{31}$$

Finally, expressions (30) and (31) together yield

$$\begin{aligned}\sum_{i=0}^{n-1}\frac{1}{p_i^2} &= \sum_{i=0}^{\lceil n/3\rceil-1}\left[\frac{(J-1)n^2}{(n-i)^2}\right]^2 + \sum_{i=\lceil n/3\rceil}^{n-1}\left[\frac{(J-1)n^2}{2(i+1)(n-i)}\right]^2 \\ &= (J-1)^2\left[\frac{19n}{24} + \frac{\pi^2 n^2}{24} + \frac{n}{2}\left(\ln n + d + \frac{1}{4} - \frac{\pi^2}{6}\right)\right] + O(\ln n) \\ &= (J-1)^2\left(\frac{\pi^2 n^2}{24} + \frac{n\ln n}{2} + hn\right) + O(\ln n),\end{aligned}$$

where $h = d/2 + (11 - \pi^2)/12 = (\ln 2 + \gamma)/2 + (11 - \pi^2)/12$. This proves equality (26).

A.2. Inequalities (19) and (21). We next demonstrate the two inequalities (19) and (21) used to derive the tail probability bound.

First, to verify inequality (19),

$$\frac{1}{1-x} \le \exp(x + x^2),$$

for $x \in [0, 3/5]$, take logarithms. This produces the equivalent inequality

$$x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots \le x + x^2. \tag{32}$$

Inequality (32) holds because

$$\begin{aligned}\frac{x^3}{3}\left(1 + \frac{3}{4}x + \frac{3}{5}x^2 + \cdots\right) &\le \frac{x^3}{3}(1 + x + x^2 + \cdots) \\ &= \frac{x^3}{3}\,\frac{1}{1-x} \le \frac{x^2}{2}\end{aligned}$$

is valid on the interval $[0, 3/5]$.

Second, we show that if we let $t = 6/[5(J-1)n]$, then

$$[e^t - 1]\frac{q_i}{p_i} \le \frac{3}{5}$$

for all i and $n > 1$. For the uniform chain corresponding to the haplotype chain, the maximum of q_i/p_i occurs when p_i attains its minimum of $2/[(J -$

$1)n]$ at $i = n - 1$. Thus,

$$\max_i \frac{q_i}{p_i} = \frac{1 - 2/[(J-1)n]}{2/[(J-1)n]} \le \frac{(J-1)n - 2}{2}.$$

Because

$$e^t - 1 = \sum_{k=1}^{\infty} \frac{t^k}{k!} \le \frac{t}{1-t}$$

for all $0 \le t < 1$, at the value $t = 6/[5(J-1)n]$ we find that

$$\begin{aligned}[e^t - 1]\frac{q_i}{p_i} &\le \frac{6/[5(J-1)n]}{1 - 6/[5(J-1)n]}\left[\frac{(J-1)n-2}{2}\right] \\ &= \frac{6}{5(J-1)n - 6}\left[\frac{(J-1)n-2}{2}\right] \le \frac{3}{5}.\end{aligned}$$

Thus, inequality (21) is satisfied for all $n > 1$.

A.3. Expectation of the χ^2 statistic. Finally, let us compute the expected value of the χ^2 statistic (24) under distributions (1) and (3). Ignoring those cells $\mathbf{i}$ with expected count $E(n^{\mathbf{i}}) = 0$, we have

$$\begin{aligned}&E\left\{\sum_{\mathbf{i}} \frac{[n^{\mathbf{i}} - E(n^{\mathbf{i}})]^2}{E(n^{\mathbf{i}})}\right\} \\ &\quad = \sum_{\mathbf{i}}\left\{\frac{E\left[(n^{\mathbf{i}})^{\underline{2}}\right]}{E(n^{\mathbf{i}})} + \frac{E(n^{\mathbf{i}})}{E(n^{\mathbf{i}})} - \frac{2E(n^{\mathbf{i}})^2}{E(n^{\mathbf{i}})} + \frac{E(n^{\mathbf{i}})^2}{E(n^{\mathbf{i}})}\right\} \\ &\quad = \sum_{\mathbf{i}} \frac{E\left[(n^{\mathbf{i}})^{\underline{2}}\right]}{E(n^{\mathbf{i}})} + C - n,\end{aligned} \tag{33}$$

where C is the total number of cells. In the haplotype case, we eliminate those alleles with 0 representatives and compute

$$C = \prod_{j=1}^{J} K(j).$$

Using the expectations given by Proposition 4 of Lange (1993),

$$\begin{aligned}\sum_{\mathbf{i}} \frac{E\left[(n^{\mathbf{i}})^{\underline{2}}\right]}{E(n^{\mathbf{i}})} &= \sum_{k(1)=1}^{K(1)} \cdots \sum_{k(j)=1}^{K(J)} \frac{\left(\prod_j [n_{jk(j)}]^{\underline{2}}\right)/\left(n^{\underline{2}}\right)^{J-1}}{\left(\prod_j n_{jk(j)}\right)/n^{J-1}} \\ &= \frac{1}{(n-1)^{J-1}} \prod_{j=1}^{J} \sum_{k(j)=1}^{K(j)} [n_{jk(j)} - 1] \\ &= \frac{1}{(n-1)^{J-1}} \prod_{j=1}^{J} [n - K(j)].\end{aligned}$$

The genotype case is more delicate. If we again eliminate those alleles absent in the sample and let $s(j)$ be the number of alleles at locus j represented in the sample by a single copy, then it is possible to form $K(j) - s(j)$ different homozygous genotypes and $\binom{K(j)}{2}$ different heterozygous genotypes at locus j. This yields $K(j)[K(j) + 1]/2 - s(j)$ possible genotypes at locus j and

$$C = \prod_{j=1}^{J} \left\{ \frac{K(j)[K(j) + 1]}{2} - s(j) \right\}$$

possible multilocus genotypes in all.

To compute the requisite expectations, consider a particular multilocus genotype $\mathbf{i}$ having genotype $a_{jk(j)}/a_{jk(j')}$ at locus j. According to formula (4),

$$E(n^{\mathbf{i}}) = \frac{n}{\left[(2n)^{\underline{2}}\right]^J} \prod_{j=1}^{J} \begin{cases} [n_{jk(j)}]^{\underline{2}}, & \text{if } k'(j) = k(j), \\ 2n_{jk(j)}n_{jk'(j)}, & \text{if } k'(j) \neq k(j), \end{cases}$$

$$E\left[(n^{\mathbf{i}})^{\underline{2}}\right] = \frac{n^{\underline{2}}}{\left[(2n)^{\underline{4}}\right]^J} \prod_{j=1}^{J} \begin{cases} [n_{jk(j)}]^{\underline{4}}, & \text{if } k'(j) = k(j), \\ 4[n_{jk(j)}]^{\underline{2}}[n_{jk'(j)}]^{\underline{2}}, & \text{if } k'(j) \neq k(j). \end{cases}$$

Hence,

$$\begin{aligned} \sum_i \frac{E\left[(n^{\mathbf{i}})^{\underline{2}}\right]}{E(n^{\mathbf{i}})} &= \sum_{k(1) \le k(1)'} \cdots \sum_{k(J) \le k(J)'} \frac{E\left[(n^{\mathbf{i}})^{\underline{2}}\right]}{E(n^{\mathbf{i}})} \\ &= \frac{(n-1)}{\left[(2n-2)^{\underline{2}}\right]^J} \\ &\quad \times \prod_{j=1}^{J} \sum_{k(j) \le k'(j)} \begin{cases} \left[n_{jk(j)} - 2\right]^{\underline{2}}, & \text{if } k(j) = k'(j) \\ 2\left[n_{jk(j)} - 1\right]\left[n_{jk'(j)} - 1\right], & \text{if } k'(j) \neq k(j) \end{cases} \\ &= \frac{(n-1)}{\left[(2n-2)^{\underline{2}}\right]^J} \\ &\quad \times \prod_{j=1}^{J} \left\{ \sum_{k(j)=1}^{K(j)} \sum_{k'(j)=1}^{K(j)} \left[n_{jk(j)} - 1\right]\left[n_{jk'(j)} - 1\right] \right. \\ &\qquad \left. - \sum_{k(j)=1}^{K(j)} \left[3n_{jk(j)} - 5\right] \right\} \end{aligned}$$

$$= \frac{(n-1)}{\left[(2n-2)^2\right]^J} \prod_{j=1}^{J} \left\{ \left[\sum_{k(j)=1}^{K(j)} \left(n_{jk(j)} - 1\right) \right]^2 - 6n + 5K(j) \right\}$$

$$= \frac{(n-1)}{\left[(2n-2)^2\right]^J} \prod_{j=1}^{J} \left\{ \left[2n - K(j)\right]^2 - 6n + 5K(j) \right\}.$$

In particular, for a single locus with K alleles of which s appear exactly once, the above expected value (33) reduces to $(n-1)K(K-1)/(2n-3) - s$.

REFERENCES

AGRESTI, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.* **7** 131–177.

ALDOUS, D. and DIACONIS, P. (1986). Shuffling cards and stopping times. *Amer. Math. Monthly* **93** 333–348.

BARNARD, G. (1963). Discussion of "The spectral analysis of point processes" by M. S. Bartlett. *J. Roy. Statist. Soc. Ser. B* **25** 294.

BESAG, J. and CLIFFORD, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76** 633–642.

BLOM, G. and HOLST, L. (1991). Embedding procedures for discrete problems in probability. *Math. Sci.* **16** 27–40.

BOYETT, J. M. (1979). Random $R \times C$ tables with given row and columns totals. *J. Roy. Statist. Soc. Ser. C* **28** 329–332.

CAVALLI-SFORZA, L. L. and BODMER, W. F. (1971). *The Genetics of Human Populations*. Freeman, San Francisco.

CROW, J. E. (1988). Eighty years ago: the beginnings of population genetics. *Genetics* **119** 473–476.

DAVID, H. A. (1981). *Order Statistics*. Wiley, New York.

DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. IMS, Hayward, CA.

DIACONIS, P. and STURMFELS, B. (1996). Algebraic algorithms for sampling from conditional distributions. Unpublished manuscript.

ELSTON, R. and FORTHOFER, R. (1977). Testing for Hardy–Weinberg equilibrium in small samples. *Biometrics* **33** 536–542.

EMIGH, T. (1980). A comparison of tests for Hardy–Weinberg equilibrium. *Biometrics* **36** 627–642.

FELLER, W. (1968). *An Introduction to Probability and Its Applications* **1**, 3rd ed. Wiley, New York.

GORADIA, T. M., LANGE, K., MILLER, P. L. and NADKARNI, P. M. (1992). Fast computation of genetic likelihoods on human pedigree data. *Human Heredity* **42** 42–62.

GRAHAM, R. L., KNUTH, D. E. and PATASHNIK, O. (1989). *Concrete Mathematics*. Addison-Wesley, Reading, MA.

GUO, S. and THOMPSON, E. (1992). Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics* **48** 361–372.

HALDANE, J. (1954). An exact test for randomness of mating. *Journal of Genetics* **52** 631–635.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

KARLIN, S. and TAYLOR, H. M. (1975). *A First Course in Stochastic Processes*. Academic Press, New York.

KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, New York.

KOLASSA, J. E. and TANNER, M. A. (1994). Approximate conditional inference in exponential families via the Gibbs sampler. *J. Amer. Statist. Assoc.* **89** 697–702.

LANGE, K. (1993). A stochastic model for genetic linkage equilibrium. *Theoret. Population Biol.* **44** 129–148.

LAZZERONI, L. C., ARNHEM, N., SCHMITT, K. and LANGE, K. (1994). Multipoint mapping calculations for sperm-typing data. *American Journal of Human Genetics* **55** 431–436.

LEVENE, H. (1949). On a matching problem arising in genetics. *Ann. Math. Statist.* **20** 91–94.

LI, C. C. (1955). *Population Genetics*. Univ. Chicago Press.

LOUIS, E. and DEMPSTER, E. (1987). An exact test for Hardy–Weinberg and multiple alleles. *Biometrics* **43** 805–811.

MATTHEWS, P. (1988). A strong uniform time for random transpositions. *J. Theoret. Probab.* **1** 411–423.

NIJENHUIS, A. and WILF, H. S. (1978). *Combinatorial Algorithms*. Academic Press, New York.

SEN, P. K. and SINGER, J. M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall, New York.

VERBEEK, A. and KROONENBERG, P. M. (1985). A survey of algorithms for exact distributions of test statistics in $r \times c$ contingency tables with fixed margins. *Comput. Statist. Data Anal.* **3** 159–285.

WEIR, B. S. (1990). *Genetic Data Analysis*. Sinauer Associates, Sunderland.

WEIR, B. S. and BROOKS, L. D. (1986). Disequilibrium on human chromosome 11p. *Genetic Epidemiology* (*Supplement*) **1** 177–183.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: laura@playfair.stanford.edu

DEPARTMENT OF BIOSTATISTICS
AND DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
E-MAIL: klange@sph.umich.edu