The Effect of Population History on the Lengths of Ancestral Chromosome Segments

Nicola H. Chapman* and Elizabeth A. Thompson†,[1]

**Division of Medical Genetics, University of Washington, Seattle, Washington 98195 and*
†*Department of Statistics, University of Washington, Seattle, Washington 98195*

Manuscript received February 20, 2002
Accepted for publication June 10, 2002

ABSTRACT

An isolated population is a group of individuals who are descended from a founding population who lived some time ago. If the founding individuals are assumed to be noninbred and unrelated, a chromosome sampled from the population can be represented as a mosaic of segments of the original ancestral types. A population in which chromosomes are made up of a few long segments will exhibit linkage disequilibrium due to founder effect over longer distances than a population in which the chromosomes are made up of many short segments. We study the length of intact ancestral segments by obtaining the expected number of junctions (points where DNA of two distinct ancestral types meet) in a chromosome. Assuming random mating, we study analytically the effects of population age, growth patterns, and internal structure on the expected number of junctions in a chromosome. We demonstrate that the type of growth a population has experienced can influence the expected number of junctions, as can population subdivision. These effects are substantial only when population sizes are very small. We also develop an approximation to the variance of the number of junctions and show that the variance is large.

An isolated population is one that is descended from a small group of individuals (founders) and in which population growth is due almost exclusively to births within the population, rather than immigration from outside. Interest in the genetics of isolated populations has recently been revived among human geneticists, because of suggestions that such populations may be useful for disequilibrium-based mapping of susceptibility loci for complex disease. In particular, it is hoped that diseases for which there are several susceptibility loci in large outbred populations may be more homogeneous in small isolated populations. In addition, small recently founded populations may exhibit linkage disequilibrium over longer genetic distances than large outbred populations (Chapman and Wijsman 1998; Kruglyak 1999).

Isolated populations are fundamentally different from the large outbred populations that are usually assumed in the theoretical study of linkage disequilibrium and may differ from one another in several aspects of their history. Populations are founded at different times by founder groups of different sizes, experience different growth patterns, and may have varying levels of internal subdivision. Chapman and Thompson (2001) give a brief survey of the variety of histories and structures seen in human populations. It is important to understand the potential effects of these aspects of a population's history on disequilibrium, both to assess the utility of disequilibrium-based studies and to interpret the results of such studies.

Lonjou *et al.* (1999) presented observed disequilibria in two regions of the genome for a wide variety of human populations. In general, levels of disequilibrium in isolated populations were only slightly higher than in outbred populations. However, the pairs of loci they considered were very tightly linked (<0.2 cM apart) and therefore this result may simply reflect the large number of generations required to break down such associations. In this article, we address how the extent of linkage disequilibrium is affected by population history, rather than considering the magnitude of disequilibrium between two loci a particular distance apart.

We study the effects of population history on the number of junctions existing in a chromosome sampled from an isolated population. A junction is a point on the chromosome where DNA from two distinct ancestral chromosomes meet (Fisher 1949). Figure 1 shows examples of two chromosomes that might have been sampled from an isolated population. Different shadings represent different ancestral types. The top chromosome contains two junctions, and the chromosome is therefore made up of three segments. The bottom chromosome contains eight junctions and is made up of nine segments. A quantity of interest is the average length of contiguous ancestral segments remaining in the generation under study. If the chromosomes have broken into many short pieces relative to the founder population, disequilibrium due to founder effect will stretch over only short distances. Conversely, if the chromosomes are composed of a small number of large pieces, relative

[1]*Corresponding author:* Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195.
E-mail: thompson@stat.washington.edu

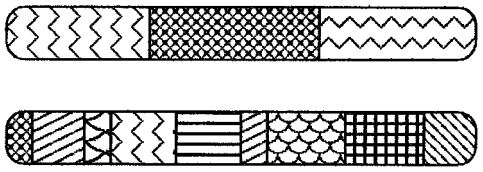


FIGURE 1.—Examples of chromosomes in isolated populations. Different shadings represent different ancestral types.

to the founder generation, disequilibrium will stretch over longer distances. If there are J junctions in a chromosome, there are $J + 1$ ancestral segments, and by Jensen's inequality,

$$E[\text{length of a segment}] = E\left[\frac{1}{J+1}\right] \geq \frac{1}{E[J]+1}. \quad (1)$$

Thus by obtaining the expected number of junctions in a length of chromosome, we obtain the expected number of contiguous segments and therefore a lower bound on their expected length.

A junction is formed when a crossover occurs between two chromosomes, at a point where they are not descendants of the same ancestral chromosome. That is, the chromosomes are not *identical by descent* (IBD) at that point. Once a junction is formed, it is transmitted as is any other gene (according to the laws of Mendelian inheritance). Since IBD is defined relative to some ancestral population, and since junctions require non-IBD to be formed, junctions are also defined relative to some ancestral population. In this article, junctions are defined relative to the founding generation; that is, this generation is assumed to consist of noninbred, unrelated individuals.

Some analogous questions regarding the lengths, number, and ancestral origins of chromosome segments have recently been considered by WIUF and HEIN (1997) and DERRIDA and JUNG-MULLER (1999). WIUF and HEIN (1997) consider, as do we, the moments of the number of ancestral chromosome segments, while DERRIDA and JUNG-MULLER (1999) focus on the number of distinct ancestors contributing to a current chromosome. The primary difference from this article is that these authors have considered the long-term equilibrium between the process of recombination and the IBD process modeled via the coalescent ancestry of chromosomes. Recombination increases the number of contributing ancestors, whereas coancestry decreases this number.

By contrast, in this article we consider IBD relative to a founder population at some defined time point in the past and the shorter-term effects of population structure. We study the formation and transmission of junctions in random-mating subdivisions of a monoecious population with discrete generations. We assume that during gamete formation, crossover events along the chromosome happen according to a Poisson process, which has rate one per morgan. This implies that the number of crossover events in a chromosome of length L has a Poisson distribution with mean L. The age of the population is assumed known, as is the size of the population at each generation. In subdivided populations, the generation of the split(s) and the sizes of the subpopulations are assumed known. We first present some theoretical results, including an expression for the expected number of junctions per morgan existing on a chromosome randomly sampled from a particular generation and two approximations to the variance of this quantity. We then apply these results to some example populations to illustrate the effects of population size, type of growth, and subdivision.

THEORETICAL DEVELOPMENT

Mean number of junctions

Let J_t be the number of junctions present on a chromosome of length L, sampled at random from a population at generation t. Let $\boldsymbol{n} = \{n_0, n_1, \ldots n_t\}$, where n_j denotes the number of junctions formed in meioses from generation j. Finally, let $I_t(k, j) = 1$ if the kth junction formed in meioses from generation j is present on the chromosome selected at time t, and let $I_t(k, j) = 0$ otherwise. Then as a function of $\boldsymbol{n}$,

$$J_t = \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} I_t(k, j).$$

Taking the expectation conditional on $\boldsymbol{n}$,

$$E[J_t|\boldsymbol{n}] = \sum_{j=0}^{t-1} \sum_{k=1}^{n_j} E[I_t(k, j)].$$

Now $E[I_t(k, j)]$ is equal to the probability that junction k from generation j is present on the selected chromosome. Let l denote the locus where junction k formed, and consider the population at generation $j + 1$. One can think of locus l as having two alleles: One is junction k, and the other is *not* junction k. The frequency of k in generation $j + 1$ is exactly $1/(2N_{j+1})$, where N_{j+1} is the population size in generation $j + 1$, and is assumed known for all j. In a random-mating population, each of the $2N_{j+1}$ genes at locus l in generation $j + 1$ are equally likely to be the ancestor of locus l in the randomly selected chromosome. Therefore $E[I_t(k, j)]$, the probability that junction k from generation j is present on the selected chromosome, is equal to $1/(2N_{j+1})$, and thus

$$E[J_t|\boldsymbol{n}] = \sum_{j=0}^{t-1} \frac{n_j}{2N_{j+1}}.$$

Taking the expectation again,

$$E[J_t] = E[E[J_t|\boldsymbol{n}]] = \sum_{j=0}^{t-1} \frac{E[n_j]}{2N_{j+1}}, \quad (2)$$

and so we require $E[n_j]$.

Calculation of $E[n_j]$: Let $H_j(p)$ denote the proportion of the chromosome that is non-IBD in individual p of

generation j. Then

$$H_j(p) = \int_0^L \frac{I_j^p(x)}{L} dx,$$

where L denotes the length of the chromosome in morgans, and $I_j^p(x) = 1$ if the two haplotypes of individual p are non-IBD at point x on the chromosome, $I_j^p(x) = 0$ otherwise. Thus

$$\begin{aligned} E[H_j(p)] &= \int_0^L \frac{E[I_j^p(x)]}{L} dx \\ &= \int_0^L \frac{h_j}{L} dx \\ &= h_j, \end{aligned} \quad (3)$$

where h_j is the probability of non-IBD at a particular locus between the two haplotypes of an individual in generation j. Now $n_j = \sum_{m=1}^{2N_{j+1}} X_j(m)$, where $X_j(m)$ denotes the number of junctions formed in meiosis m from generation j. Since crossovers happen along the chromosome according to a Poisson process with rate one per morgan, conditional on $H_j(p_m)$, $X_j(m)$ has a Poisson distribution with mean $H_j(p_m)L$, where p_m denotes the parent of meiosis m, and L denotes the length of the chromosome in morgans. Therefore

$$\begin{aligned} E[X_j(m)] &= E[E[X_j(m)|H_j(p_m)]] \\ &= E[H_j(p_m)L] \\ &= h_j L, \end{aligned}$$

since the parent is simply a randomly chosen individual from generation j, and by Equation 3. Then

$$E[n_j] = E\left[\sum_{m=1}^{2N_{j+1}} X_j(m)\right] = 2N_{j+1}h_j L. \quad (4)$$

Substituting Equation 4 into Equation 2, we obtain

$$E[J_t] = \sum_{j=0}^{t-1} h_j \cdot L. \quad (5)$$

For the random-mating population considered here, $h_j = \prod_{i=0}^{j-1}(1 - (2N_i)^{-1})$ (Crow and Kimura 1970). This result allows calculation of the number of junctions expected in a chromosome, as a function of population sizes, thereby allowing the exploration of the effects of different patterns of population growth. In a subdivided population, the required population sizes are simply those within the subdivision of interest.

Equation 5 demonstrates that population history affects the expected number of junctions in a chromosome through the probability of non-IBD in each generation. This implies that in a large population where h_j remains close to one over many generations, the number of generations since the founding of the population is the most important factor in determining the expected number of junctions and therefore the lower bound on the expected length of intact ancestral segments. Growth patterns that result in small population sizes over long periods of time will result in the accumulation of IBD, and as a result fewer junctions will be expected in chromosomes from such populations. Similarly, chromosomes from populations in which there is extensive subdivision will be expected to carry fewer junctions and therefore have longer intact ancestral segments.

Variance of the number of junctions

Recall that n_i denotes the total number of junctions formed in all meioses from generation i.

Poisson approximation: We first consider a variance approximation on the basis of some simplifying assumptions. Specifically, suppose that

1. n_i has a Poisson distribution with mean $2N_{i+1}h_i L$.
2. n_i is independent of n_j for all $i \neq j$.
3. The presence of any one junction in the sampled chromosome from generation t is independent of the presence of any other junction in that chromosome. That is, Pr(*junction k formed in a meiosis from generation i exists in the chromosome sampled at generation t*|*junction l formed in a meiosis from generation j exists in the chromosome sampled at generation t*) = Pr(*junction k formed in a meiosis from generation i exists in the chromosome sampled at generation t), for any k, l, i, and j, where* $k \neq l$ *if* $i = j$.

Let $J_t(i)$ denote the number of junctions formed in generation i that exist in the randomly sampled chromosome from generation t. Then assumption 1, together with the fact that the probability that a junction formed in a meiosis from generation i exists in the chromosome sampled at generation t equals $1/(2N_{i+1})$, implies that $J_t(i)$ has a Poisson distribution with mean $h_i L$, for $0 \leq i \leq t-1$. Furthermore, assumptions 2 and 3 imply that $J_t(i)$ is independent of $J_t(j)$, for $i \neq j$. Therefore

$$J_t = \sum_{j=0}^{t-1} J_t(j) \sim \text{Poisson}\left(\sum_{j=0}^{t-1} h_j L\right).$$

For the Poisson distribution, the variance is equal to the mean and can therefore be calculated using Equation 5.

The above assumptions do not generally hold. Assumption 1 would hold if all of the individuals in generation i had the same proportion h_i of their genome non-IBD. In fact, this proportion varies across members of generation i and is equal to h_i only in expectation. This extra variability leads to extra-Poisson variation in the distribution of n_i. Assumption 2 does not hold, since, for example, knowing that n_i is very small relative to the number of meioses implies that the population is likely close to fixation, and therefore subsequent n_j $(j > i)$ must also be small. Junctions formed close to one another in the same meiosis are likely to be inherited together, and therefore assumption 3 is not generally true. The violation of these assumptions implies that the true variance is likely higher than that predicted by the Poisson

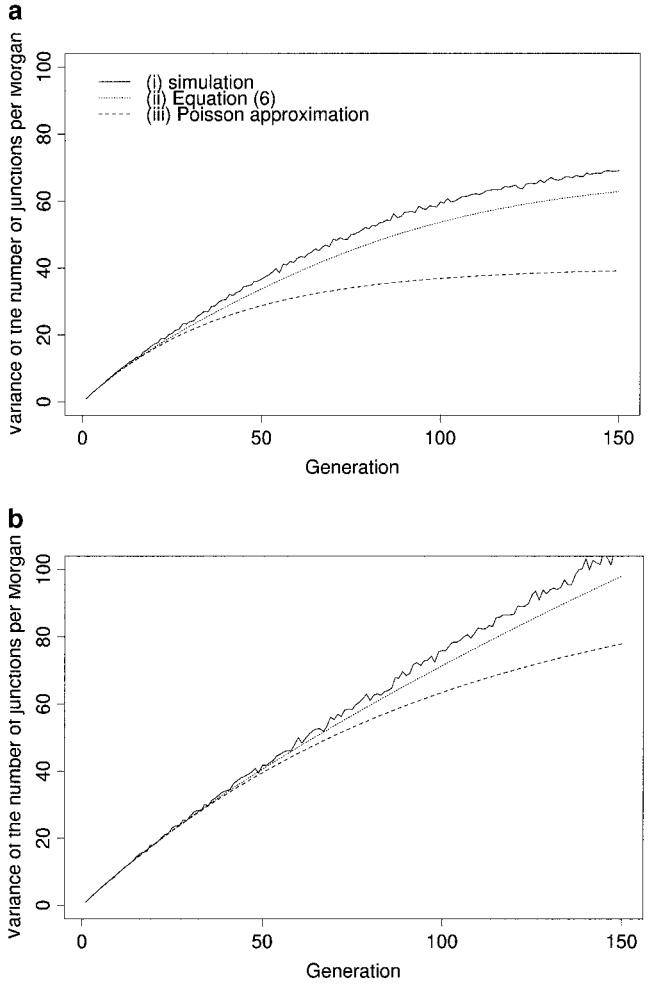


FIGURE 2.—Variance of the number of junctions in a randomly selected chromosome from a population of constant size $N = 20$ or $N = 50$, estimated by (i) simulation, (ii) Equation 6, and (iii) the Poisson approximation. (a) $N = 20$. (b) $N = 50$.

approximation. The assumptions are probably closer to the truth in larger populations.

Simulations: The performance of the Poisson approximation to the variance was investigated by simulation. Chromosome data were simulated for random-mating populations of constant size ($N = 20$ or $N = 50$) over 150 generations. Individual chromosomes were represented by a linked list of segments, where adjacent segments were of distinct ancestral types. Each individual in generation $i + 1$ was produced by randomly choosing (with replacement) two parents from generation i. A gamete from each of these parents was generated by simulating the locations of crossovers according to a Poisson process and constructing the gamete out of the appropriate segments of parental chromosomes. More details can be found in CHAPMAN (2001). In each simulation, a chromosome was randomly selected for the generation of interest, and the number of junctions existing in that chromosome was recorded. Variance estimates are based on 10,000 simulations.

TABLE 1

Ratio of the estimated variance (based on 10,000 simulations) to the theoretical mean as a function of population size (N) and generation (t)

$\hat{\sigma}^2/\mu$	$t = 20$	$t = 50$	$t = 125$	$t = 150$
$N = 20$	1.08	1.28	1.70	1.76
$N = 50$	0.99	1.06	1.29	1.34

Figure 2 shows the variance of the number of junctions per morgan in populations of constant size either $N = 20$ or $N = 50$, estimated by simulation and by the Poisson approximation. The Poisson variance is an underestimate of the true variance, especially for older generations, and the smaller population. Since for a true Poisson random variable, mean and variance are equal, Table 1 shows the ratio of the estimated variance ($\hat{\sigma}^2$, based on 10,000 simulations) to the theoretical mean (μ) as a function of N and t. Comparing the populations at times where $t/N = 1$ ($N = 20$, $t = 20$ and $N = 50$, $t = 50$) we see that the mean underestimates the variance by approximately the same amount: 6 or 8%. Similarly, comparing populations where $t/N = 2.5$ ($N = 20$, $t = 50$ and $N = 50$, $t = 125$) the mean underestimates the variance by 28 or 29%. This suggests that for a constant-sized population, t/N approximately determines the adequacy of the Poisson approximation to the variance. For values of $t/N > 1$, the Poisson approximation underestimates the true variance. The importance of the quantity t/N is not surprising, since for a population of constant size, $h_t = (1 - (2N)^{-1})^t \simeq \exp(-t/(2N))$. Larger values of t/N correspond to increasing amounts of IBD in the population, and in these situations, assumptions 1–3 may be further from the truth.

Relaxing assumptions 1 and 2: We now develop a second variance approximation, which does not require assumptions 1 and 2. Consider the calculation of $E[J_t^2]$. As a function of $\boldsymbol{n}$,

$$
\begin{aligned}
J_t^2 &= \left(\sum_{j=0}^{t-1}\sum_{k=1}^{n_j} I_t(k,j)\right)^2 \\
&= \sum_{j=0}^{t-1}\left\{\sum_{k=1}^{n_j} I_t(k,j)\right\}^2 + \sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\left\{\sum_{k=1}^{n_i} I_t(k,i)\right\}\left\{\sum_{l=1}^{n_j} I_t(l,j)\right\} \\
&= \sum_{j=0}^{t-1}\sum_{k=1}^{n_j} I_t(k,j) + \sum_{j=0}^{t-1}\sum_{k=1}^{n_j}\sum_{\substack{l=1,\\ l\neq k}}^{n_j} I_t(k,j)I_t(l,j) \\
&\quad + \sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} I_t(k,i)I_t(l,j).
\end{aligned}
$$

The first term in J_t^2 is a sum over all junctions. The second term is a sum over pairs of distinct junctions formed in the same generation, and the third term is

a sum over pairs of junctions formed in different generations. Applying conditional expectation,

$$E[J_t^2] = E[E[J_t^2|\boldsymbol{n}]]$$
$$= E\Bigg[\sum_{j=0}^{t-1}\sum_{k=1}^{n_j} E[I_t(k,j)] + \sum_{j=0}^{t-1}\sum_{k=1}^{n_j}\sum_{\substack{l=1,\\ l\neq k}}^{n_j} E[I_t(k,j)I_t(l,j)] + \sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} E[I_t(k,i)I_t(l,j)]\Bigg].$$

We argued previously that $E[I_t(k,j)] = 1/(2N_{j+1})$. By assumption 3,

$$E[I_t(k,j)I_t(l,j)] \simeq \frac{1}{2N_{j+1}} \cdot \frac{1}{2N_{j+1}},$$

and

$$E[I_t(k,i)I_t(l,j)] \simeq \frac{1}{2N_{i+1}} \cdot \frac{1}{2N_{j+1}}.$$

Then

$$E[J_t^2] = E[E[J_t^2|\boldsymbol{n}]]$$
$$\simeq E\Bigg[\sum_{j=0}^{t-1}\sum_{k=1}^{n_j}\frac{1}{2N_{j+1}}\Bigg] + E\Bigg[\sum_{j=0}^{t-1}\sum_{k=1}^{n_j}\sum_{\substack{l=1,\\ l\neq k}}^{n_j}\frac{1}{4N_{j+1}^2}\Bigg] + E\Bigg[\sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}\frac{1}{4N_{i+1}N_{j+1}}\Bigg]$$

and so

$$E[J_t^2] \simeq \sum_{j=0}^{t-1}\frac{E[n_j]}{2N_{j+1}} + \sum_{j=0}^{t-1}\frac{E[n_j(n_j-1)]}{4N_{j+1}^2} + \sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\frac{E[n_in_j]}{4N_{i+1}N_{j+1}}.$$

Therefore

$$\mathrm{Var}[J_t] = E[J_t^2] - E[J_t]^2$$
$$\simeq \sum_{j=0}^{t-1}\frac{E[n_j]}{2N_{j+1}} + \sum_{j=0}^{t-1}\frac{E[n_j(n_j-1)]}{4N_{j+1}^2} + \sum_{i=0}^{t-1}\sum_{\substack{j=0,\\ j\neq i}}^{t-1}\frac{E[n_in_j]}{4N_{i+1}N_{j+1}} - \Bigg(\sum_{j=0}^{t-1}h_jL\Bigg)^2. \qquad (6)$$

Expressions for $E[n_j^2]$ and $E[n_in_j]$ are developed in the APPENDIX (Equations A8 and A9), and $E[n_j]$ is given in Equation 4. The expectations in Equation 6 depend on the chromosome length and the population sizes over time, through the single-locus non-IBD probabilities (h_j, $j = 0 \ldots t-1$), and the two-locus non-IBD probabilities [$\Theta_j(\theta), \Gamma_j(\theta), \Delta_j(\theta)$ $j = 0, \ldots t-1$], which are described in the APPENDIX.

Figure 2 shows the variance of the number of junctions per morgan in populations of constant size either $N = 20$ or $N = 50$. The variance is estimated by simulation (10,000 iterations), Equation 6, and the Poisson approximation. For both populations, Equation 6 is much better than the Poisson-based variance approximation, particularly for later generations. It is interesting to note that in both examples, the Poisson approximation begins to fail at approximately the Nth generation, which is where the non-IBD proportion has been reduced to ~60%. For generations earlier than this, the two variance approximations are almost indistinguishable, and they are very close to the simulated variance (see Figure 3). This suggests that for young populations or older, larger populations, the Poisson variance approximation may be adequate. The Poisson approximation to the variance has an advantage over Equation 6, because it is so much easier to calculate.

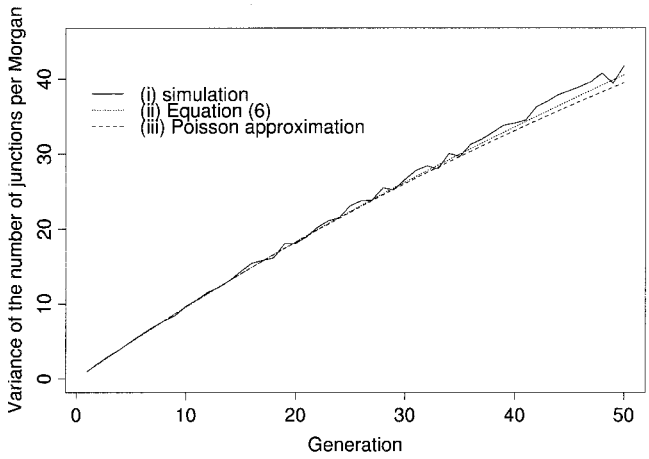


FIGURE 3.—Variance of the number of junctions in early generations from a population of constant size $N = 50$, estimated by (i) simulation, (ii) Equation 6, and (iii) the Poisson approximation.

APPLICATION TO GROWING POPULATIONS WITH AND WITHOUT SUBDIVISION

To demonstrate the potential effects of different types of population growth on expected junction number and therefore intact segment length, we consider an example. Consider a population that has grown to 100 times its initial size, over a period of 100 generations. This example reflects the age of modern Finnish (NEVANLINNA 1972) and Japanese (BENEDICT 1989) populations. We consider initial population sizes (N_0) of 20, 100, and 500 individuals, and for each we consider five growth scenarios:

Linear growth: expansion by a constant number of individuals each generation.

Exponential growth: expansion by a constant percentage each generation. A 100-fold increase over 100 generations corresponds to a growth rate of 4.72% per generation.

Exponential growth with internal subdivision: population bifurcates whenever a population size of $2N_0$ is reached (first division at $t = 15$, subsequently every 15 generations).

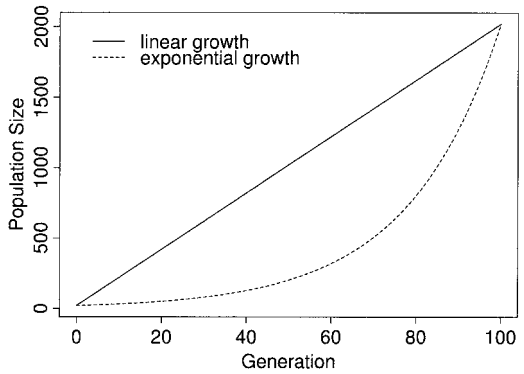


FIGURE 4.—Population sizes over time for linear and exponentially growing populations with $N_0 = 20$.

Exponential growth with internal subdivision: population bifurcates whenever a population size of $4N_0$ is reached (first division at $t = 30$, subsequently every 15 generations).

Exponential growth with internal subdivision: population bifurcates whenever a population size of $8N_0$ is reached (first division at $t = 45$, subsequently every 15 generations).

For a given value of N_0, all scenarios have the same total size at generation 100. All exponential growth scenarios have the same total number of individuals at all generations—the difference is in the extent of internal subdivision. Figure 4 shows the total population sizes over time for the population with $N_0 = 20$.

Table 2 shows the expected number of junctions in a chromosome selected from generation 100 (using Equation 5) and the corresponding lower bound on the expected length of intact ancestral segments (using Equation 1), for each of the five growth scenarios with $N_0 = 20$. In these populations, the type of growth has a pronounced effect on the expected number of junctions. Substantially more junctions are expected in the linearly growing population than in any of the exponentially growing populations. This is because the linearly growing population increases its size rapidly enough in the early generations that little IBD is accumulated. In contrast, all the exponentially growing populations remain small for a long period of time, during which IBD accumulates within the population. Thus fewer junctions are formed. For the same reason, increasing amounts of subdivision within the exponentially growing populations results in substantially fewer junctions being formed. Intact ancestral segments in the unsubdivided exponentially growing population are expected to be ~50% larger than in the linearly growing population. In the most subdivided exponential population, ancestral segments are expected to be twice as long as in the linearly growing population and almost 50% larger than those in the unsubdivided exponentially growing population. Thus different patterns of population growth can have a dramatic effect on expected number of junctions in a chromosome and therefore the length of ancestral segments.

TABLE 2

Expected number of junctions on a chromosome of length 1 M from generation 100 and the corresponding lower bound on expected segment length for each of the five growth scenarios with $N_0 = 20$

Type of growth	Subdivision	$E[J_{100}]$	Lower bound (cM) on E[segment length]
Linear	None	90.0	1.0
Exponential	None	64.9	1.5
	Split at $8N_0$	62.8	1.6
	Split at $4N_0$	57.3	1.7
	Split at $2N_0$	45.0	2.2

TABLE 3

Expected number of junctions on a chromosome of length 1 M from generation 100, for each of the five growth scenarios with $N_0 = 100$ and $N_0 = 500$

Type of growth	Subdivision	N_0 = 100	N_0 = 500
Linear	None	97.9	99.6
Exponential	None	91.7	98.3
	Split at $8N_0$	90.9	98.1
	Split at $4N_0$	88.9	97.7
	Split at $2N_0$	83.5	96.4

Table 3 shows the expected number of junctions on a chromosome of length 1 M from generation 100, for each of the five growth scenarios and the larger founding population sizes. For the larger populations ($N_0 = 100$ and $N_0 = 500$) the expected number of junctions in the linearly growing population is close to 100, which is what one would expect in an infinitely large population where IBD does not accumulate. This reflects the fact that little IBD accumulates in these populations because they start relatively large and grow quickly. The number of junctions expected in the exponentially growing populations is reduced relative to the linearly growing populations and further reduced in the subdivided populations. While these trends are the same as those observed in the smallest populations ($N_0 = 20$, see Table 2), the magnitude of the effects is much smaller. For example, when $N_0 = 500$, only 3% more junctions are expected in the linearly growing population than in the most subdivided exponentially growing population.

It is also important to consider the variability of the number of junctions in a chromosome. Table 4 shows the variance of the number of junctions in a chromo-

TABLE 4

Variance of the number of junctions in a chromosome randomly selected from generation 100, based on 10,000 simulations, Equation 6, or the Poisson approximation

		Exponential growth			
	Linear growth	No subdivision	$8N_0$	$4N_0$	$2N_0$
		$N_0 = 20$			
Simulation	92.04	86.94	81.85	77.78	68.86
Equation 6	90.68	80.50	76.95	70.93	60.53
Poisson	90.04	64.93	62.80	57.28	45.00
		$N_0 = 100$			
Simulation	—	—	94.07	91.67	86.24
Equation 6	97.93	92.06	91.24	89.34	84.60
Poisson	97.93	91.74	90.91	88.93	83.45
		$N_0 = 500$			
Simulation	—	—	—	—	—
Equation 6	99.58	98.29	98.11	97.66	96.38
Poisson	99.58	98.29	98.10	97.66	96.36

some of length 1 M from generation 100, estimated by simulation, Equation 6, and the Poisson approximation. Simulation-based estimates are available only for populations with $N_0 = 20$ and the subdivided populations with $N_0 = 100$, since simulation of the larger populations is too computationally demanding. For the populations with $N_0 = 20$, the Poisson approximation badly underestimates the variance. The approximation based on Equation 6 is much better, but still an underestimate. When $N_0 = 100$, both approximations are closer to the simulated values, and Equation 6 is still better. For the populations with $N_0 = 500$, the variance approximations are virtually identical, and we hypothesize that the variance is well estimated by either approximation for populations this large. The variance is always greater than or equal to the mean.

DISCUSSION

The theoretical development shows that the most important factor in determining the expected number of junctions in a chromosome, and therefore a lower bound for the average length of intact ancestral segments, is the time since founding of the population. In generation t of an infinitely large random-mating population, we expect t junctions per morgan in a chromosome. In finite populations, the expectation is $<t$, but the difference is substantial only if the historical population sizes have been small enough to result in the accumulation of IBD and therefore the production of fewer junctions. Similarly, different growth patterns and levels of subdivision affect the expected number of junctions in a substantial way only if population sizes are very small. Even when this is the case, the variance of the number of junctions in a chromosome is large, and so the existing number of junctions in a chromosome may differ substantially from that expected on the basis of known population history and structure.

These results allow us to predict that disequilibrium may persist over longer distances in smaller, more recently founded populations. Whether or not it does depends on the patterns of junction formation in many meioses, which we cannot observe. Studies of the extent of disequilibrium across the genome of an isolated population are therefore desirable. Only then can the utility of a large-scale disequilibrium mapping study be assessed.

We are grateful to a referee for drawing our attention to the related work of Wiuf and Hein (1997) and Derrida and Jung-Muller (1999). This work was supported in part by the Burroughs Wellcome Fund for the Program in Mathematical and Molecular Biology.

LITERATURE CITED

Benedict, R., 1989 *The Crysanthemum and the Sword.* Houghton Mifflin, Boston.

Chapman, N. H., 2001 Genome descent in isolated populations. Ph.D. Thesis, University of Washington, Seattle, WA.

Chapman, N. H., and E. A. Thompson, 2001 Linkage disequilibrium mapping: the role of population history, size and structure, pp. 413–437 in *Advances in Genetics*, Vol. 42. Academic Press, San Diego.

Chapman, N. H., and E. M. Wijsman, 1998 Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. Am. J. Hum. Genet. **63:** 1872–1885.

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

Derrida, B., and B. Jung-Muller, 1999 The genealogical tree of a chromosome. J. Stat. Phys. **94:** 277–298.

Fisher, R. A., 1949 *The Theory of Inbreeding.* Oliver and Boyd, Edinburgh.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

Lonjou, C., A. Collins and N. E. Morton, 1999 Allelic association between marker loci. Proc. Natl. Acad. Sci. USA **96:** 1621–1626.

Nevanlinna, H. R., 1972 The Finnish population structure—a genetic and genealogical study. Hereditas **71:** 195–236.

Weir, B. S., P. J. Avery and W. G. Hill, 1980 Effect of mating structure on variation in inbreeding. Theor. Popul. Biol. **18:** 396–429.

Wiuf, C., and J. Hein, 1997 On the number of ancestors to a DNA sequence. Genetics **147:** 1459–1468.

Communicating editor: M. Veuille

APPENDIX: CALCULATION OF SECOND-ORDER MOMENTS OF $H_i(p)$ AND n_i

To calculate the second-order moments of n_i, we require the second-order moments of $H_i(p)$, the proportion of the chromosome that is non-IBD in individual p of generation i.

Second-order moments of $H_i(p)$: To calculate second-order moments of $H_i(p)$, we consider some two-locus gene nonidentity measures described by Weir *et al.* (1980) and illustrated in Figure A1. Generally, we are interested in the probability that genes a and a' at locus x are non-IBD, and genes b and b' at locus y are also non-IBD. This probability is denoted Θ, Γ, or Δ ac-

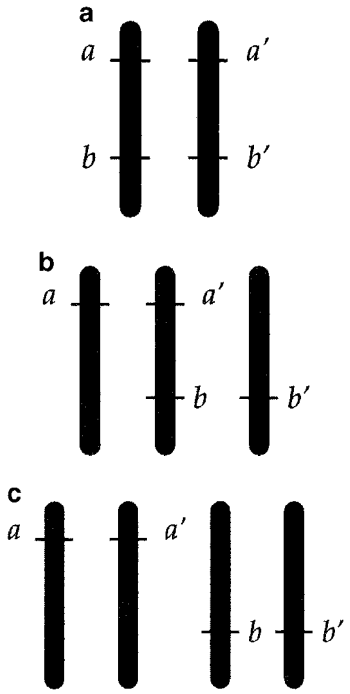


FIGURE A1.—Two-locus gene nonidentity measures. (a) Θ. (b) Γ. (c) Δ.

cording to the number of chromosomes in which the loci are being compared (see Figure A1). WEIR *et al.* (1980) consider the evolution of these probabilities over time for populations reproducing according to various schemes of random mating with discrete generations. Let $\nu_i = (\Theta_i, \Gamma_i, \Delta_i)^T$ denote the column vector of two-locus non-IBD probabilities at generation i. WEIR *et al.* (1980) show that $\nu_{i+1} = \Omega \cdot \nu_i$, where Ω is a transition matrix that depends on the recombination fraction θ between the loci and the size (N_i) of the population at generation i. Therefore ν_i depends on the population sizes up to and including generation $i - 1$ and the recombination fraction θ. We denote the probabilities of interest by $\Theta_i(\theta)$, $\Gamma_i(\theta)$, and $\Delta_i(\theta)$.

Calculation of $E[H_i(p)^2]$: Consider $E[H_i(p)^2]$, the expected value of the square of the non-IBD proportion in an individual in generation i.

$$\begin{aligned} E[H_i(p)^2] &= E\left[\int_0^L \frac{I_i^p(x)}{L}dx \cdot \int_0^L \frac{I_i^p(y)}{L}dy\right] \\ &= \frac{1}{L^2}\int_0^L\int_0^L E[I_i^p(x) \cdot I_i^p(y)]\,dxdy \\ &= \frac{1}{L^2}\int_0^L\int_0^L \Theta_i(\theta_{|x-y|})\,dxdy \\ &= \frac{2}{L^2}\int_0^L (L - s)\Theta_i(\theta_s)\,ds \equiv \overline{\Theta}_i. \end{aligned} \quad \text{(A1)}$$

In this equation, θ_s denotes the recombination fraction between two loci a distance s morgans apart, and line 4 is obtained from line 3 by a change of variables $s = |x - y|$ and integration. $\overline{\Theta}_i$ is too complicated to evaluate exactly. WEIR *et al.* (1980) discuss its estimation by numerical integration.

Calculation of $E[H_i(p) \cdot H_i(p')]$: Consider the product of the non-IBD proportions of two distinct individuals p and p' in the ith generation.

$$\begin{aligned} E[H_i(p) \cdot H_i(p')] &= E\left[\int_0^L \frac{I_i^p(x)}{L}dx \cdot \int_0^L \frac{I_i^{p'}(y)}{L}dy\right] \\ &= \frac{1}{L^2}\int_0^L\int_0^L E[I_i^p(x) \cdot I_i^{p'}(y)]\,dxdy \\ &= \frac{1}{L^2}\int_0^L\int_0^L \Delta_i(\theta_{|x-y|})\,dxdy \\ &= \frac{2}{L^2}\int_0^L (L - s)\Delta_i(\theta_s)\,ds \equiv \overline{\Delta}_i. \end{aligned} \quad \text{(A2)}$$

Calculation of $E[H_i(p) \cdot H_j(p')]$: Finally, we examine the product of the non-IBD proportions of two individuals: p from the ith generation, and p' from the jth generation. We assume that $i < j$. Then

$$\begin{aligned} E[H_i(p) \cdot H_j(p')] &= E\left[\int_0^L \frac{I_i^p(x)}{L}dx \cdot \int_0^L \frac{I_j^{p'}(y)}{L}dy\right] \\ &= \frac{1}{L^2}\int_0^L\int_0^L E[I_i^p(x) \cdot I_j^{p'}(y)]\,dxdy \\ &= \frac{1}{L^2}\int_0^L\int_0^L \Pr(a_i \neq a_i';\ b_j \neq b_j')\,dxdy, \end{aligned} \quad \text{(A3)}$$

where a_i and a_i' denote the genes at locus x in person p of generation i, b_j and b_j' denote the genes at locus y in person p' of generation j, and $\neq$ indicates non-IBD. To have $b_j \neq b_j'$, b_j and b_j' must be descended from different individuals in generation $j - 1$. This implies that

$$\Pr(a_i \neq a_i';\ b_j \neq b_j') = \left(1 - \frac{1}{2N_{j-1}}\right) \cdot \Pr(a_i \neq a_i';\ b_{j-1} \neq b_{j-1}'), \quad \text{(A4)}$$

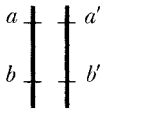
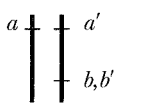
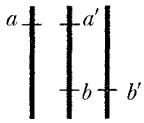
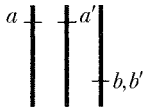
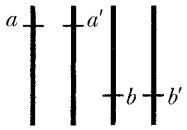
where b_{j-1} and b_{j-1}' denote the ancestors at generation $j - 1$ of b_j and b_j', respectively. Applying (A4) iteratively, we obtain

$$\Pr(a_i \neq a_i';\ b_j \neq b_j') = \prod_{k=1}^{j-i-1}\left(1 - \frac{1}{2N_{i+k}}\right) \cdot \Pr(a_i \neq a_i';\ b_{i+1} \neq b_{i+1}'), \quad \text{(A5)}$$

where b and b' denote genes at locus y on distinct chromosomes in generation $i + 1$. The probability on the right-hand side of Equation A5 depends on the relationship between the chromosomes carrying a_i, a_i', and the ancestors b_i and b_i' of b_{i+1} and b_{i+1}'. Table A1 shows the possible configurations of b_i and b_i', the probability of each configuration, calculated using the random-

TABLE A1

Possible configurations of a, a', b, and b'

Configuration	Probability	$\Pr(a \neq a';\ b \neq b')$
	$2 \cdot \frac{1}{2N_4} \cdot \frac{1}{2N_i}$	$\Theta_i(\theta)$
	$2 \cdot \frac{1}{2N_4} \cdot \frac{1}{2N_i}$	0
	$2 \cdot \frac{1}{2N_i} \cdot \frac{2N_i - 2}{2N_i} \cdot 2$	$\Gamma_i(\theta)$
	$\frac{2N_i - 2}{2N_i} \cdot \frac{1}{2N_i}$	0
	$\frac{2N_i - 2}{2N_i} \cdot \frac{2N_i - 3}{2N_i}$	$\Delta_i(\theta)$

mating model, and the desired probability $\Pr(a_i \neq a_i';\ b_{i+1} \neq b_{i+1}')$ conditional on that configuration.

The probability required in Equation A3 is then obtained by summing over the possible configurations and substituting that quantity into Equation A5. Therefore

$$\begin{aligned}\Pr(a_i \neq a_i';\ b_j \neq b_j') = &\prod_{k=1}^{j-i-1}\left(1 - \frac{1}{2N_{i+k}}\right)\\ &\cdot\left[\frac{1}{2N_i^2}\Theta_i(\theta_{|x-y|}) + \frac{2(N_i-1)}{N_i^2}\Gamma_i(\theta_{|x-y|}) + \frac{(N_i-1)(2N_i-3)}{2N_i^2}\Delta_i(\theta_{|x-y|})\right].\end{aligned} \tag{A6}$$

Substituting Equation A6 into Equation A3, we find

$$\begin{aligned}E[H_i(p)\cdot H_j(p')] &= \frac{1}{L^2}\int_0^L\int_0^L \prod_{k=1}^{j-i-1}\left(1 - \frac{1}{2N_{i+k}}\right)\\ &\quad\cdot\left[\frac{1}{2N_i^2}\Theta_i(\theta_{|x-y|}) + \frac{2(N_i-1)}{N_i^2}\Gamma_i(\theta_{|x-y|}) + \frac{(N_i-1)(2N_i-3)}{2N_i^2}\Delta_i(\theta_{|x-y|})\right]dxdy\\ &= \prod_{k=1}^{j-i-1}\left(1 - \frac{1}{2N_{i+k}}\right)\\ &\quad\cdot\left[\frac{1}{2N_i^2}\frac{2}{L^2}\int_0^L (L-s)\Theta_i(\theta_s)\,ds + \frac{2(N_i-1)}{N_i^2}\frac{2}{L^2}\int_0^L (L-s)\Gamma_i(\theta_s)\,ds + \frac{(N_i-1)(2N_i-3)}{2N_i^2}\frac{2}{L^2}\int_0^L (L-s)\Delta_i(\theta_s)\,ds\right]\\ &= \prod_{k=1}^{j-i-1}\left(1 - \frac{1}{2N_{i+k}}\right)\\ &\quad\cdot\left[\frac{1}{2N_i^2}\overline{\Theta}_i + \frac{2(N_i-1)}{N_i^2}\overline{\Gamma}_i + \frac{(N_i-1)(2N_i-3)}{2N_i^2}\overline{\Delta}_i\right],\end{aligned} \tag{A7}$$

where

$$\overline{\Gamma}_i \equiv \frac{2}{L^2}\int_0^L (L-s)\Gamma_i(\theta_s)\,ds.$$

Second-order moments of n_i: To calculate $E[n_i^2]$ and $E[n_i n_j]$, we use the formula $n_i = \sum_{m=1}^{2N_{i+1}} X_i(m)$, where $X_i(m)$ denotes the number of junctions formed in meiosis m from generation i. Since crossovers happen along the chromosome according to a Poisson process with rate one per morgan, given $H_i(p_m)$, $X_i(m)$ has a Poisson distribution with mean $H_i(p_m)L$, where p_m denotes the parent of meiosis m, and L denotes the length of the chromosome in morgans. Therefore

$$\begin{aligned}E[X_i(m)^2] &= E[E[X_i(m)^2|H_i(p_m)]]\\ &= E[H_i(p_m)L + (H_i(p_m)L)^2]\\ &= E[H_i(p_m)]L + E[H_i(p_m)^2]L^2\\ &= h_iL + \overline{\Theta}_iL^2,\end{aligned}$$

since the parent is simply a randomly chosen individual from generation i.

To calculate $E[n_i^2]$, we also consider $E[X_i(m)X_i(m')]$, the expected value of the product of the numbers of junctions formed in two different meioses from the same generation:

$$\begin{aligned}E[X_i(m)X_i(m')] &= E[E[X_i(m)X_i(m')|H_i(p_m)H_i(p_{m'})]]\\ &= E[H_i(p_m)H_i(p_{m'})L^2]\end{aligned}$$

since conditional on the proportion non-IBD in each of the parents, the numbers of junctions formed in each meiosis are independent. With probability $1/N_i$, both meioses are from the same parent. Otherwise they are from distinct individuals in the ith generation. Therefore

$$\begin{aligned}E[X_i(m)X_i(m')] &= E[H_i(p_m)H_i(p_{m'})L^2]\\ &= \frac{1}{N_i}E[H_i(p)^2L^2] + \frac{N_i-1}{N_i}E[H_i(p)H_i(p')L^2]\end{aligned}$$

$$= \frac{1}{N_i}\overline{\Theta}_i L^2 + \frac{N_i - 1}{N_i}\overline{\Delta}_i L^2$$

by Equations A1 and A2. Then

$$\begin{aligned} E[n_i^2] &= E\left[\left(\sum_{m=1}^{2N_{i+1}} X_i(m)\right)^2\right] \\ &= E\left[\sum_{m=1}^{2N_{i+1}} X_i(m)^2\right] + E\left[\sum_{m=1}^{2N_{i+1}} \sum_{\substack{m'=1,\\ m'\neq m}}^{2N_{i+1}} X_i(m) X_i(m')\right] \\ &= 2N_{i+1}(h_i L + \overline{\Theta}_i L^2) \\ &\quad + 2N_{i+1} \cdot (2N_{i+1} - 1)\left(\frac{1}{N_i}\overline{\Theta}_i + \frac{N_i - 1}{N_i}\overline{\Delta}_i\right)L^2. \end{aligned} \tag{A8}$$

Calculation of $E[n_i n_j]$ requires that we first obtain $E[X_i(m)X_j(m')]$, the expected value of the product of the numbers of junctions formed in two meioses occurring in two different generations. Now,

$$\begin{aligned} E[X_i(m)X_j(m')] &= E[E[X_i(m)X_j(m')\,|H_i(p_m)H_j(p_{m'})]] \\ &= E[H_i(p_m)H_j(p_{m'})L^2] \\ &= \prod_{k=1}^{j-i-1}\left(1 - \frac{1}{2N_{i+k}}\right) \\ &\quad \times \left[\frac{1}{2N_i^2}\overline{\Theta}_i + \frac{2(N_i - 1)}{N_i^2}\overline{\Gamma}_i \right. \\ &\quad \left. + \frac{(N_i - 1)(2N_i - 3)}{2N_i^2}\overline{\Delta}_i\right]L^2 \end{aligned}$$

for $i < j$, by Equation A7. Then

$$\begin{aligned} E[n_i n_j] &= E\left[\sum_{m=1}^{2N_{i+1}} X_i(m) \cdot \sum_{m'=1}^{2N_{j+1}} X_j(m')\right] \\ &= E\left[\sum_{m=1}^{2N_{i+1}} \sum_{m'=1}^{2N_{j+1}} X_i(m)X_j(m')\right] \\ &= \sum_{m=1}^{2N_{i+1}} \sum_{m'=1}^{2N_{j+1}} E[X_i(m)X_j(m')] \\ &= 4N_{i+1}N_{j+1}\prod_{k=1}^{j-i-1}\left(1 - \frac{1}{2N_{i+k}}\right)\left[\frac{1}{2N_i^2}\overline{\Theta}_i + \frac{2(N_i - 1)}{N_i^2}\overline{\Gamma}_i \right. \\ &\quad \left. + \frac{(N_i - 1)(2N_i - 3)}{2N_i^2}\overline{\Delta}_i\right]L^2. \end{aligned} \tag{A9}$$