Stepping stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum

Arkendra De* and Richard Durrett†

Departments of Statistics* and Mathematics†, Cornell U.

Ithaca NY 14853

Running Head: Consequences of Spatial Structure

Keywords: population subdivision, island model, stepping stone model, linkage disequilibrium, site frequency spectrum

Corresponding Author:

Rick Durrett

523 Malott Hall

Cornell University

Ithaca NY 14853

Phone (607) 255-8282

FAX (607) 255-7149

Email: rtd1@cornell.edu

ABSTRACT

The symmetric island model with *D* demes and equal migration rates is often chosen for the investigation of the consequences of population subdivision. Here we show that a stepping stone model has a more pronounced effect on the genealogy of a sample. For samples from a small geographical region commonly used in genetic studies of humans and Drosophila, there is a shift of the frequency spectrum that decreases the number of low frequency derived alleles, and skews the distribution of statistics of Tajima, Fu and Li, and Fay and Wu. Stepping stone spatial structure also changes the two locus sampling distribution, increases both linkage disequilibrium and the probability that two sites are perfectly correlated. This may cause a false prediction of cold spots of recombination and may confuse haplotype tests which compute probabilities based on a homogeneously mixing population.

INTRODUCTION

Homogeneously mixing populations of constant size are a convenient setting to develop the theory of population genetics. However, when one wants to understand patterns observed in data, one must consider the effects of population growth, bottlenecks, and population subdivision. When the consequences of population subdivision are investigated, symmetric island model with D demes and equal migration rates is the usual choice, and the case of two demes is especially popular. The island model is easy to analyze mathematically due to the fact that if two lineages are not in the same deme, then their relative location is not important. However, this has the unrealistic consequence that after one migration event, the lineage is distributed uniformly over the species range.

An alternative approach to modeling spatial structure which does not suffer from this defect is the stepping stone model. In this paper, we will investigate the consequences of modeling space as a two dimensional stepping stone model in which there is an L by L grid of colonies and migration only to neighboring colonies. The migration scheme is very simple, however, it results in what Wright called isolation by distance. In other words, it takes a number of migration events for the lineages to spread across the system. As we will see, this feature, which is certainly present in Drosophila and early human populations, causes a dramatic change in the coalescence structure of lineages.

The reason for this is intuitively clear. At small times the lineages have not had a chance to spread across the population, so the effective population size is reduced. The coalescence rate is increased reducing the number of low frequency derived alleles, skewing the site frequency spectrum, and increasing linkage disequilibrium. These effects occur in the island model as well, two lineages sampled from one deme have an increased coalescence rate until one of them migrates, at which point they behave like a random sample from the overall population. In contrast, as we will later explain, in the stepping stone model the effective population size increases roughly linearly in time.

The main point of this paper is to argue that spatial structure in the form of the stepping stone model has a different effect than the symmetric island model and can have a much greater impact on genealogies, so it should also be considered when one wants to asses the impact of spatial structure on estimation procedures or statistical tests. We begin by reviewing theoretical results of Cox and Durrett (2002), and Zähle, Cox, and Durrett (2005) for the coalescence time of a sample of size n, and contrast these results with the corresponding facts about the symmetric island model. The strange nonlinear time scaling needed to reduce genealogies in the stepping stone model to Kingman's coalescent, indicates that there is a strong effect on commonly used statistics, but the exact nature of the changes are difficult to analyze mathematically. Because of this, we turn to simulations to demonstrate the effect of stepping stone spatial structure on the decay of linkage disequilibrium, the site frequency spectrum, and the distribution of test statistics based on the site frequency spectrum.

THEORETICAL RESULTS

Wakeley, with various co-authors, has investigated the island model when the number of demes is large. Let N be the number of diploid individuals per colony. Wakeley (1998) has shown that for a scattered sample in which we get at most one sequence from each deme, in the limit as $D \to \infty$ the genealogy of a sample of size *n* is the same as that of a homogenously mixing population of size.

$$N_e = ND\left(1 + \frac{1}{M}\right) \quad \text{where} \quad M = 4Nm. \tag{1}$$

This formula for N_e is the same as the one of Nei and Takahata (1993).

The reason for the simplification for large *D* is easy to understand. At most times all lineages are in different demes, their actual locations are irrelevant, and the coalescence times will have the lack of memory property that characterizes the exponential. If we sample *n* chromosomes from one deme then there is an initial period called "scattering phase," which ends when all of the surviving lineages are in different demes. This initial phase is short compared to the coalescence time, so it is equivalent to a random reduction in the sample size. For more details, see page 1864 of Wakeley (1999).

Lessard and Wakeley (2004) have recently extended this analysis to the two-locus ancestral graph in a subdivided population. Wakeley and Lessard (2003) have applied these results to the study of LD in humans. They found that their model with a large number of demes fit the data for humans well (see Figure 2 on page 1049), in contrast to Reich *et al*. (2002) who did not get a good fit from the two island model.

The stepping stone model has been extensively studied since its introduction by Kimura in the 1950s. There have been many important contributions by Kimura and Weiss, Malécot, Maruyama, Nagylaki, Crow and Aoki, Slatkin, and others. Here we will focus on recent results of Cox and Durrett (2002), and Zähle, Cox, and Durrett (2005), referring the reader to the 2002 paper for more on the historical development.

Cox and Durrett (2002) investigated the Moran model in which each individual is replaced at rate 1, replacement comes from a different deme with the migration probability *m* (called ν by Cox and Durrett), and the probability a migrant to *x* comes from *y* is *q*(*y*-*x*) where *q* is a probability distribution with finite range that has the same symmetries as the two dimensional lattice. This symmetry assumption implies that the two coordinates are uncorrelated and have the same variance σ^2. The grid of colonies is an *L* by *L* square and the difference *y*-*x* is computed modulo *L*, i.e., we have periodic boundary conditions that identify opposite edges of the square. This assumption is a mathematical convenience that has been used in many previous studies, but is not necessary. The proofs in Cox and Durrett (2002), and Zähle, Cox, and Durrett (2005) extend easily to a flat universe with migration out of the system suppressed or reflected back in, and the qualitative behavior is the same.

Theorem 4 of Cox and Durrett (2002) shows that if we pick any two chromosomes and $Nm / \log L \to \infty$ as $L \to \infty$ then the coalescence time divided by NL^2 converges to an exponential distribution with mean 1. In words, if the per colony migration rate, *Nm*, is much larger than log *L* the system is essentially homogeneously mixing. This is a strong migration limit which corresponds to results of Nagylaki (1980) and Notohara (1993) for the island model. It should also be compared to the remark of Kimura and Maruyama (1971) that "marked local differentiation of gene frequencies can occur if $Nm < 1$" while "if $Nm \geq 4$ the whole population tends to behave as a panmictic unit." As we can see from the mathematical result, the cutoff between the two behaviors is not constant somewhere between 1 and 4, but depends on the number of demes and increases like log *L*.

To state results for the case in which the local population size N is not much larger than log L, we need the following rescaled migration rate

$$\alpha = \frac{2\pi}{\log L} \cdot Nm\sigma^2$$

This definition may look mysterious but as we will see α is the natural rescaled migration rate for the stepping stone model, which is the analogue of $M = 4Nm$ for the island model. Note that in the stepping stone model the variance σ^2 joins the product Nm to give the composite parameter that describes the strength of migration, but in contrast to the island model this quantity is divided by log L. The 2π which comes from the central limit theorem is included to make the next formula simple.

Theorem 5 of Cox and Durrett (2002) shows that two chromosomes sampled at random from the population have a coalescence time that is asymptotically, as $L \to \infty$, exponential with mean

$$(1+\alpha)\frac{L^2 \log L}{2\pi\sigma^2 m} = NL^2 \frac{1+\alpha}{\alpha}$$

This result is similar to one found by Barton *et al*. (2002) in a model with a grid of colonies with the local alleles frequencies modeled by diffusion processes. As stated in (9) of Charlesworth, Charlesworth, and Barton (2003) if the density of individuals ρ = 1 the mean coalescence time of two individuals is

$$\frac{L^2 \log(KL/\sigma)}{2\pi\sigma^2} + 2L^2$$

where K is a constant. The second term and the K/σ inside the logarithm are there to make the approximation more accurate for small L. They are not important as $L \to \infty$, so they should be dropped when comparing with our asymptotic result, but even with this there remains the difference of the factor of 1+α. It is difficult to say what causes this difference since the result cited by Charlesworth, Charlesworth, and Barton (2003) and attributed to Barton et al. (2002) does not appear in that paper.

Theorem 2 in Zähle, Cox, and Durrett (2005) extends the result for the coalescence time of two chromosomes by showing that a sample of n chromosomes chosen at random from the population has the same genealogy as a sample of size n from a homogeneously mixing population of size

$$N_e = NL^2\left(1+\frac{1}{\alpha}\right) \tag{2}$$

Note that this has the same form as N_e in the island model when the number of demes $D = L^2$ and the scaled migration rate $M = \alpha$. To illustrate the use of these formulas, consider our 10 x 10 grid with migration to nearest neighbors so $L = 10$ and $\sigma^2 = ½$. If the local population size is $N = 25$ and we choose $m = 0.1$ so that $4Nm = 1$ then $\alpha = \pi(0.25)/\log(10) = 0.341$ and $N_e = 2500(1.341)/0.341 = 9829$ versus the actual population size of 2500.

In many genetic studies, samples are not chosen at random from the population as a whole. For example, one of the samples in Sabeti *et al*. (2002) consists of 73 Beni individuals who are civil servants in Benin City, Nigeria. To capture this type of local sample in our framework, we assume that the n chromosomes are sampled at random from a L^β by L^β square of colonies. Theorem 3 in Zähle, Cox, and Durrett (2005) shows that we get the ordinary coalescent after a nonlinear time change in which times $L^{2\gamma}/4\pi m\sigma^2$ with $\beta \le \gamma \le 1$ correspond to time $\log((\gamma+\alpha)/(\beta+\alpha))$ in the coalescent, and then time proceeds at the usual linear rate for a population with the N_e given in (2). Here, we have changed Zähle, Cox, and Durrett's (2005) *2m* in the denominator to

$4\pi m\sigma^2$. This does not affect the limit theorem, but as we will see, it makes for a better approximation.

To help explain the time change, we note that the coalescence rate at time $s = L^{2\gamma} / 4\pi m\sigma^2$ (that is, $\gamma = \log(4\pi m\sigma^2 s) / 2\log L$) is

$$\frac{d}{ds}\log\left(\frac{\gamma+\alpha}{\beta+\alpha}\right) = \frac{1}{\gamma+\alpha}\cdot\frac{1}{2s\log L}$$

To make a connection with simulation results in Figure 4 of Wilkins (2004) we have graphed 1 over the coalescence rate versus time s in Figure 1. This shows that the coalescence occurs much more rapidly in the initial stages of the stepping stone model compared to a homogeneously mixing population. The reason for graphing 1 over the rate is that this quantity is almost linear in time. This can be seen intuitively by noting that at time s, the central limit theorem says that lineages will be spread over a region with radius of order $\sqrt{s}$ and hence area of order s, so the effective population size is of order s.

The remarks in the previous paragraph apply to times $s = L^{2\gamma} / 4\pi m\sigma^2$ with $\gamma \leq 1$. At times $\geq L^2/4\pi m\sigma^2$ the lineages have had time to spread across the entire space. At this point the coalescence time, which is of order $L^2 \log L$, is much larger than the time, of order $L^2/m\sigma^2$, needed for the a random walk on an L by L square that jumps at rate $2m$ and has variance σ^2 to equilibrate in the uniform distribution (Cox and Durrett 2002). Thus the relative positions of lineages are unimportant and the system behaves as if it were homogeneously mixing. Since we have changed Zähle, Cox, and Durrett's (2005) $2m$ in the denominator to $4\pi m\sigma^2$, 1 over the coalescence rate increases until it becomes constant in the second regime, and the transition is continuous.

Zähle, Cox, and Durrett (2005) were able to compute various quantities for samples of size 2 under the infinite sites model including the expected number of pairwise differences and the probability for no coalescence before recombination for two loci with a per generation recombination probability *r*. They showed that the latter quantity decayed more slowly in the stepping stone model compared to a homogenously mixing population, but it is hard to relate this to commonly used measures of linkage disequilibrium.

The limit theorem of Zähle, Cox, and Durrett (2005) is difficult to use for computations because the coalescence rate changes in time. For the ordinary coalescent one can easily compute the correlation between coalescence times at two loci for samples of size two by considering whether recombination or coalescence occurs first. One can find this result of Griffiths (1981) explained on pages 80-83 of Durrett (2002). However, in the situation of Zähle, Cox, and Durrett (2005) one must also consider the time at which the event occurs, and one can no longer obtain the answer by solving three equations in three unknowns. Thus, to investigate the effect of stepping stone population structure on samples of size $n > 2$, we turn to simulations.

METHODS

Model Simulation Parameters and Sampling Schemes: We simulated coalescent models with constant recombination and mutation rates across the locus in homogeneously mixing, island, and stepping stone models using Hudson's ms program (Hudson 2002). We fix the physical size of our locus to 100 kb and set both mutation and recombination rate to be 10^{-8} /nucleotide/generation. Our sample size is fixed for all models at n = 40 chromosomes. In the spatial simulations there are 100 demes and *m* is the probability that the new individual is a new migrant. In the stepping stone model, space is a 10 by 10 grid with periodic boundary conditions and with migration to the four nearest neighbor migration with equal probabilities.

In order to try to minimize the differences between the spatial structures, we use N_e formulas from Nei and Takahata (1993) and Zähle, Cox, Durrett (2005) given earlier as (1) and (2) to determine the number of diploids per colony, N so that the computed N_{e} for the population is roughly 10,000. Table 1 lists the scaled migration rate (*4Nm*), number of diploids per colony (N) , and computed effective population sizes (N_e). The definition of effective population size we are using here is ½ the average coalescence time of two lineages, so this makes the mean number of pairwise differences the same.

In order to understand the effect of sampling on the statistics, we employ two different sampling schemes for the island and stepping stone models: (i) Chromosomes are randomly selected from the population (random sampling). (ii) 40 chromosomes are sampled from one colony in the stepping stone model, or from one deme in the island model (local sampling). The second sampling strategy corresponds to sampling from one population.

Decay of Linkage Disequilibrium: Following Pritchard and Przeworski (2001), we compute the square of the correlation coefficient, which for two loci with two alleles A and a, and B and b is

$$r = \frac{p(AB) - p(A)p(B)}{\sqrt{p(A)p(a)p(B)p(b)}}$$

for all pairs of segregating sites for which the minor allele frequency is at least 0.2. In order to average results over the replications, we create bins of size 1000 (0.01 times the length of our region) based on interSNP distance, and average the r^2 observations in each bin. The number of simulations used to compute averages was 400,000. In addition to investigating the mean values of r^2 , we examine the distribution of r^2 values for distances in the bins [0.1,0.11], [0.3,0.31], and [0.5,0.51], where the distance is measure relative to the length of the locus. That we are examining r^2 for loci separated by roughly 10, 30, and 50Kb. The number of simulations used to determine the distribution of r^2 was 350,000.

Site Frequency Spectrum: Since the alignment and ancestral state are known, we can compute for each SNP the observed number of chromosomes i ($1 \leq i \leq 39$) that have the mutant nucleotide. This number is then divided by the total number of segregating sites from all 350,000 replications, to get the site frequency distribution.

SNP Density: We choose our physical sequence length to 10 kb. The number of segregating sites for 350,000 simulations was tabulated and normalized.

Tajima's (1989) D statistic and **Fay and Wu's (2000) H** are calculated for each replication using Hudson's "sample_stats" program, which is included with the ms program. The median, and 2.5, and 97.5 percentiles are computed over 350,000 simulations.

RESULTS

To give a visualization of the impact of spatial structure on genetic data, Figure 2 gives sample outcomes for a homogeneously mixing population, an island model local sample with *4Nm* = 1, and a stepping stone local sample with *4Nm* = 10. Notice that there are more SNPs and many more haplotypes in the homogeneously mixing population compared to the two spatial samples.

Decay of Linkage Disequilibrium: As Figure 3 shows, when $4Nm$ = 1, there is a large difference in the rate of decay of r^2 between homogeneously mixing and the migration models. As expected, the stepping stone model has considerably more LD than the island model. When $4Nm$=1, the stepping stone local sample has $r^2 \approx 0.9$ at a distance of 100kb, which is considerably larger than values observed in the human genome, but of course our universe is only a 10 by 10 array of colonies. The random samples have a faster decay of r^2 than the local samples, but in contrast to the theoretical results quoted above, their behavior is not the same as that of homogeneously mixing population. One reason for this is that the sample size is n = 40, so n^2 = 1600 is much larger than the number of demes, 100, and the assumption of the limit theorem is not justified. A simple calculation for 40 lineages and 100 demes shows that the probability that all lineages will be in separate demes is 0.000122. Using a Poisson approximation with mean 0.4 for the number of lineages in a deme, we see that on the average 5.26 demes will have two lineages.

When $4Nm$=3, values of r^2 are somewhat reduced. As expected they are the smallest for the random samples, and largest for the local samples. When $4Nm$ = 10, the random samples are close to the homogeneously mixing case, but the curves for local samples are well above the homogeneously mixing decay curve.

Distribution of $\mathbf{r^2}$: Figure 4 compares the distribution of r^2 for a homogeneously mixing population and a stepping stone local sample with $4Nm$ = 10, for SNPs separated by 10-11, 30-31, or 50-51 Kb. Not only is the mean of r^2 larger in the stepping stone model, but there is a significant probability that r^2 = 1, even for SNPs separated by 50-51 Kb.

Site Frequency Spectrum: As Figure 5 shows, in the case of a random sample, the site frequency spectrum for island and stepping stone model with *4Nm* = 1 behave like a homogeneously mixing population. That is, as Fu (1995) has shown, the probability that *k* members of the sample have the mutant nucleotide is *c/k*, where *c* is a constant that makes the probabilities sum to 1. This is the behavior expected based on the theoretical results for random samples discussed earlier, which show that the genealogy of a random sample converges to that of Kingman's coalescent.

For the local samples, the site frequency spectra differ from the prediction for a homogeneously mixing population. We see from panels b-d in Figure 5 that both models have a significant reduction in the proportion of singletons, even when *4Nm*=10. This occurs because there is a greater initial coalescence rate due to the fact that the lineages are sampled from only one subpopulation. The reduction of singletons presents problems for the use of Fu and Li's (1993) *D* statistic, which looks for an excess in the frequency of such mutations compared to the neutral expectation.

For *4Nm* = 1 and *4Nm* = 3, both models exhibit an excess of intermediate and high frequency derived nucleotides compared to homogeneously mixing model. When *4Nm* = 10, the discrepancy at the high frequency end is almost gone but there are significant differences for rare alleles.

SNP Density: Figure 6 shows that for random samples the SNP density from both migration models match closely the one for a homogeneously mixing population, even when *4Nm* = 1. For local samples the SNP density is changed in the spatial models, see panels b-d, with the number of SNPs shifted toward smaller values, and the shift more pronounced for the stepping stone model than for the island model. Note that when

4Nm=1, more than 35% of segments of our 10Kb segments have zero SNPs in the stepping stone model, compared to 5% for the island model, while this almost never happens in a homogeneously mixing population. The skew in the distribution persists in local samples from both spatial models when *4Nm=3* and even *4Nm=10*.

Tajima's D is a statistic that is constructed by subtracting estimates of the scaled mutation rate based on the number of pairwise differences and the number of segregating sites. These components are related to the last two quantities we have investigated, so we should not expect much difference for a random sample, but a much larger one for a local sample. The results reported in Figure 7 follow this pattern. The medians for the random samples are close to the homogeneously mixing values. However, there are significant changes in the upper and lower cutoffs for the local samples and in most cases a dramatic shift toward positive values as shown by the changes in the median and the upper cutoff. This agrees with our previous observation that there is an excess of mutations at intermediate frequencies. In the stepping stone local samples, and to a lesser extent in the island local samples, this shift is accompanied by an increase in variability that causes the lower tail cutoffs to decrease. This is unfortunate for researchers looking for negative values of Tajima’s D as indications of positive selection. This use of Tajima’s *D* has also been shown to be misleading in bottlenecked populations (Thornton and Jensen, in press).

Fay and Wu's H is a statistic constructed by subtracting estimates of the scaled mutation rate based on the number of pairwise differences and another based on the homozygosity of derived variants. The difference is normalized so that the *H* statistic has variance 1. The homozygosity of derived variants is influenced most by variants at intermediate and high frequencies respectively. See pages 1406 and 1408 of Fay and Wu (2000). Since the most noticeable affect of spatial structure is to decrease the number of low frequency alleles, it is not surprising to see in Figure 8 that there are no systematic changes in the median of the *H* statistic, but the major effect is to increase the standard deviation and to expand the interval between the two cutoffs, which can result in spurious rejections of the neutral model.

DISCUSSION

The spatial distribution of individuals in a local sample in the island model or stepping stone model causes coalescence to occur more rapidly in the early stages of the genealogy of a local sample. This shifts the site frequency spectrum from rare alleles toward those of intermediate frequency, and we have shown through simulation that this alters the distribution of test statistics of Tajima (1989), Fu and Li (1993), and Fay and Wu (2000). Here we have contrasted a local sample from one population with a sample chosen randomly form the entire population, and shown that the local sample will have fewer alleles. In a different direction, Ptak and Przeworski (2002) have shown that taking small samples from a large number of geographic locations can increase the number of rare alleles.

Our simulation results have shown that stepping stone population structure produces a slow decay of linkage disequilibrium and dramatically increases the probability of perfect correlation, that is, $r^2 = 1$. This change in the two locus sampling distribution may cause trouble for likelihood methods, such as the ones McVean *et al.* (2004) have used to estimate recombination rates in humans. Low recombination rates could be assigned to intervals between markers with high correlation, when in reality this is due to spatial structure. In making the last comment, we are not suggesting that when spatial structure is taken into account, all cold spots will suddenly become warm. However, the qualitative picture of recombination rate variability may be significantly changed.

For a concrete example where spatial structure may have contributed to erroneously rejecting neutrality, we consider results of Hamblin and Aquadro (1996) for the glucose dehydrogenase gene based on a sample of 11 *Drosophila simulans* collected in 1984 in Raleigh, North Carolina. The only test that suggested a patterns of nonneutral evolution was Fu and Li's test. In that case observing one singleton out of 26 segregating sites had a probability of $p < 0.05$ on a two tailed test, but might not be significant if one took into account that local sampling could itself reduce the number of singleton mutations.

The effect of local population sampling in humans, particularly in European or North American populations may not be as dramatic, since as one goes back a few dozen generations the lineages disperse over a wide area. However, it could be more problematic when a population has occupied a small area for a long period of time. Sabeti *et al*. (2002) sample 73 Beni from Benin City, Nigeria and sequenced a region around the G6PD locus. To argue that selection had acted on this locus they examined the decay of the extended haplotype homozygosity (EHH), i.e., the probability that two randomly chosen chromosomes carrying the same core haplotype were identical by descent up to that point. To evaluate the likelihood of the observed data, they used Hudson's ms program to simulate homogeneously mixing populations of constant size, with exponential growth, a bottleneck, or a two island model. However, as we have shown, stepping stone spatial structure can dramatically reduce the decay of linkage disequilibrium.

We are not the first to have suggested that spatial structure of the human population may have played a role in the spurious detection of positive selection. Mekel-Bobrov *et al*. (2005) and Evans *et al*. (2005) in studies of ASPM (abnormal spindle-like microcephaly associated) and microcephalin genes that the presence of haplotypes with unusually large frequencies were caused by positive selection. Currat *et al*. (2006) argued that human demographic models with structure followed by population growth could explain the observed haplotype frequency patterns. In reply, Mekel-Bobrov *et al*. (2006) argued that bottleneck in Currat *et al*.'s model was unrealistically long and narrow.

While one can debate the impact of spatial structure on genealogies that cannot be directly observed, there is clear evidence that some allele frequencies show pronounced spatial structure: for example, lactase persistence (Bersaglieri *et al*. 2004, Tishkoff *et al*. 2007), the CCR5-Δ32 deletion which leads to strong resistance against HIV-1 (Stephens *et al*. 1998), and the Duffy blood group (Hamblin, Thompson, and DiRienzo 2002). One final observation that argues for the importance of spatial structure in shaping patterns of variability is that of Rosenberg *et al*. (2002), who have shown that with information about a large number of microsatellite loci, one can classify most of the 1056 individuals in a sample from 52 populations into their correct geographical regions.

In the other direction, one might argue that patterns of nucleotide variability are shaped over longer time scales than microsatellites, so only recent mutations will show the effects of population structure. However, this paper has clearly shown that the "isolation by distance" in stepping stone model has a profound effect on patterns of variability, even when *Nm* is not much larger than 1, so it should also be considered when one wants to assess the impact of population subdivision on estimation procedures or statistical tests.

ACKNOWLEDGEMENTS

We would like to thank Jeff Jensen, Nadia Singh, and two anonymous referees for a large number of helpful suggestions. Both authors have been partially supported by a joint NIGMS/NSF grant (DMS 0201037).

REFERENCES

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., *et al*. (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J. Hum. Genet. 74, 1111-1120

Barton, N.H., DePaulis, F., and Etheridge, A.M. (2002) Neutral evolution in a spatially continuous population. Theor. Pop. Biol. 61, 31-48

Charlesworth, B., Charlesworth, D and Barton, N. H. (2003) The effects of genetic and geographic structure on neutral variation. Annu. Rev. Ecol. Evol. Syst. 34, 99-125

Cox, J.T. and Durrett, R. (2002) The stepping stone model: New formulas expose old myths. Ann. Appl. Prob. 12, 1348-1377

Currat, M., Excoffier, L., Maddison, W., Otto, S.P., Ray, N., *et al*. (2006) Comment on "Ongoing adaptive evolution of *ASPM,* a brain size determinant in *Homo sapiens*" and "*Microcephalin*, a gene regulating brain size, continues to evolve adaptively in humans" Science 313, Technical comment 172a

Durrett, R. (2002) *Probability Models for DNA Sequence Evolution.* Springer, New York

Evans, P.D., Gilbert, S.L., Mekel-Bobrov, N., Vallender, E.J., Anderson, J.R., *et al*. (2005) *Microcephalin*, a gene regulating brain size, continues to evolve adaptively in humans. Science. 309, 1717-1720

Fay, J.C., and Wu, C-I. (2000) Hitchhiking under positive Darwinian selection. Genetics 155 (2000), 1405-1413

Fu, Y.X. (1995) Statistical properties of segregating sites. Theor. Pop. Biol. 48, 172-197

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

Griffiths, R.C. (1981) Neutral two-locus multiple allele models with recombination. Theor. Pop. Biol. 19, 169-186

Hamblin, M.T., and Aquadro, C.F. (1996) High nucleotide sequence variation in a region of low recombination in Drosophila simulans is consistent with the background selection model. Mol. Biol. Evol. 13, 1133-1140

Hamblin, M.T., Thompson, E.E., and DiRienzo, A. (2002) Complex signatures of natural selection at the Duffy blood group locus. Am. J. Human Genetics. 70, 369-383

Hudson, R. (2001) Two-locus sampling distributions and their application. Genetics 159, 1805-1817

Hudson, R. (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18: 337-338

Kimura, M., and Maruyama, T. (1971) Patterns of neutral polymorphism in a geographically structured population. Genet. Res. 18, 125-131

Lessard, S., and Wakeley, J. (2004) The two-locus ancestral graph in a subdivided population: convergence as the number of demes grow in the island model. J. Math. Biol. 48, 275-292

McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. (2004) The fine scale structure of recombination rate variation in the human genome. Science. 304, 581-584

Mekel-Bobrov, N., Gilbert, S.L., Evans, P.D., Vallender, E.J., Anderson, J.R., *et al.* (2005) “Ongoing adaptive evolution of *ASPM,* a brain size determinant in *Homo sapies.*” Science. 309, 1720-1722

Mekel-Bobrov, N., Evans, P.D., Gilbert, S.L., Vallender, E.J., Hudson, R.R., *et al* (2006) Response to comment on "Ongoing adaptive evolution of *ASPM,* a brain size determinant in *Homo sapiens*" and "*Microcephalin*, a gene regulating brain size, continues to evolve adaptively in humans" Science 313, Technical comment 172b

Nagylaki, Nagylaki, T. (1980) The strong-migration limit in geographically structured populations. J. Math. Biology 9, 101-114

Nei, M., and Takahata, N. (1993) Effective population size, genetic diversity, and coalescence time in subdivided populations. J. Mol. Evol. 37, 240-244

Notohara, M. (1993) The strong-migration limit for the genealogical process in geographically structured populations. J. Math. Biol. 31, 115-122

Ptak, S.E. and Przeworski, M. (2002) Evidence for population growth in humans is confounded by fine-scale population structure. Trends in Genetics. 18, 559-563

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., *et al* (2002). Genetic structure of human populations. Science 298, 2381-2385

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832-837

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmuller, J., Varilly, P., *et al*. (2006) Positive natural selection in the human lineage. Science 312, 1614-1620

Stephens, L.C., Reich, D.E., Goldstein, D.B., Shin, H.D., Smith, M.W., *et al*. (1998) Dating the origin of the CCR5-Δ32 AIDS-resistance allele by the coalescence of haplotypes. Am. J. Human Genetics. 62, 1507-1515

Tajima, F. (1989) Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism Genetics 123, 585-595

Thornton, K.R., and Jensen, J.D. Controlling the false positive rate in multi-locus genome scans for selection. Genetics, in press

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbit, C.C., *et al*. Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics. 39, 31-40

Wakeley, J. (1998) Segregating sites in Wright's island model. Theor. Pop. Biol. 53, 166-174

Wakeley, J. (1999) Nonequilibrium migration inhuman history. Genetics 153, 1863-1871

Wakeley, J., and Lessard, S. (2003) Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. Genetics 164, 1043-1053

Wilkins, J. (2004) A separation-of-timescales approach to the coalescent in a continuous population. J. Math. Biol. 31, 115-122

Zähle,I., Cox, J.T., and Durrett, R. (2005) The stepping stone model, II. Genealogies and the infinite sites model. Ann. Appl. Probab. 15, 671-699

Model	**4Nm**	**N**	**Computed N_e**
Island	1	50	10,000
Island	3	75	10,000
Island	10	91	10,100
Stepping Stone	1	25	9,829
Stepping Stone	3	51	10,083
Stepping Stone	10	77	9,957

Table 1. Computed effective population size and scaled migration rates for island, and stepping stone models with our simulation parameters.

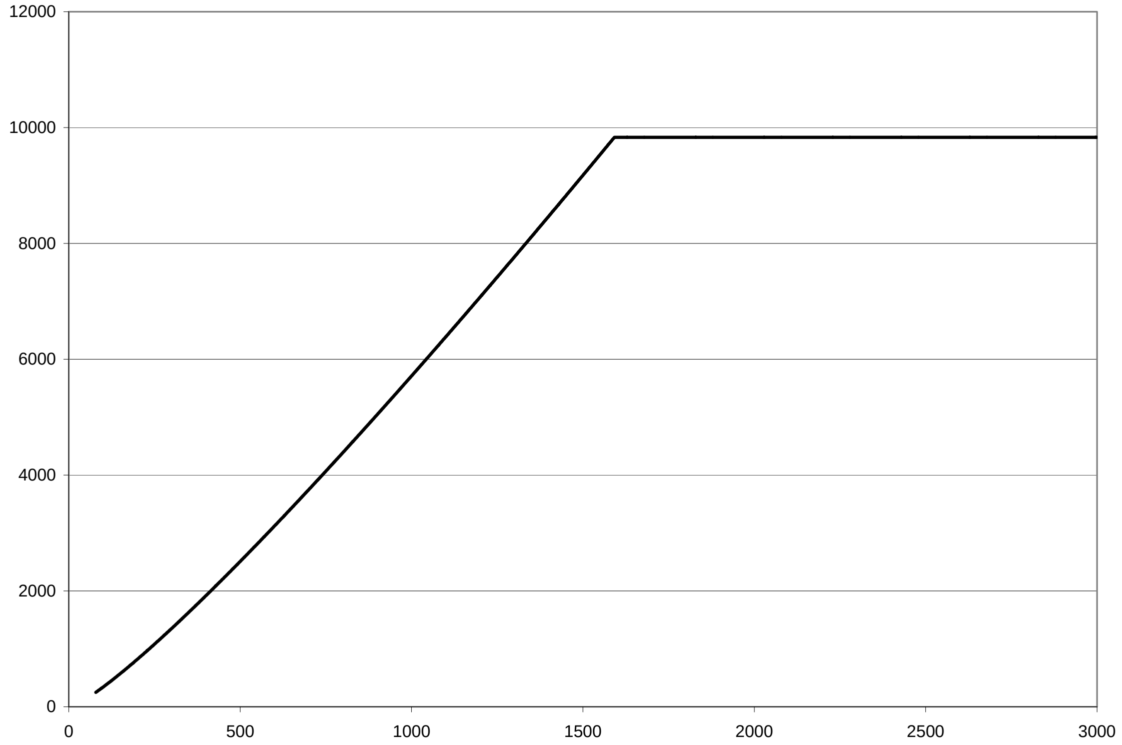


Figure 1. 1 over coalescence rate in the Zähle-Cox-Durrett result.

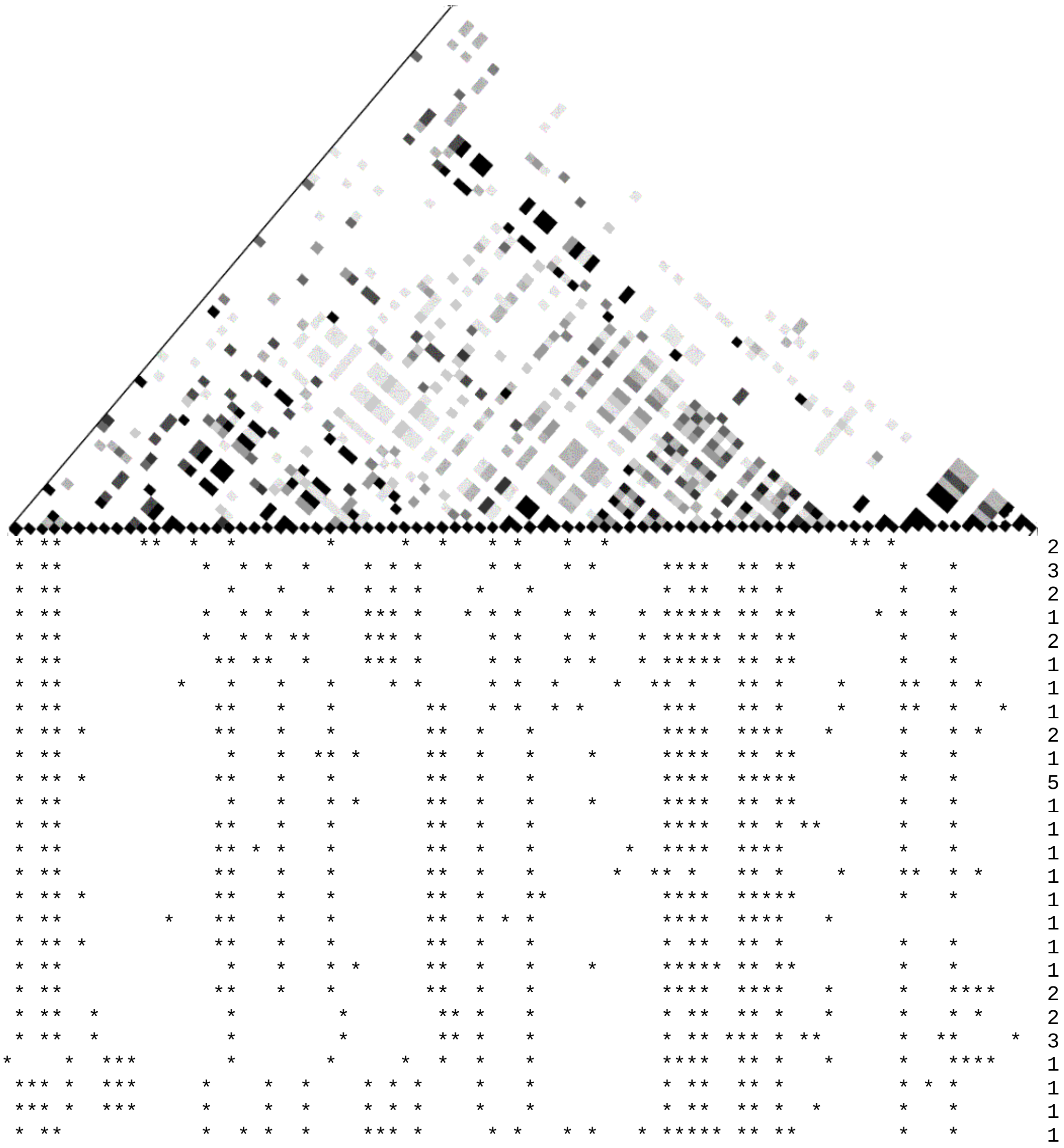
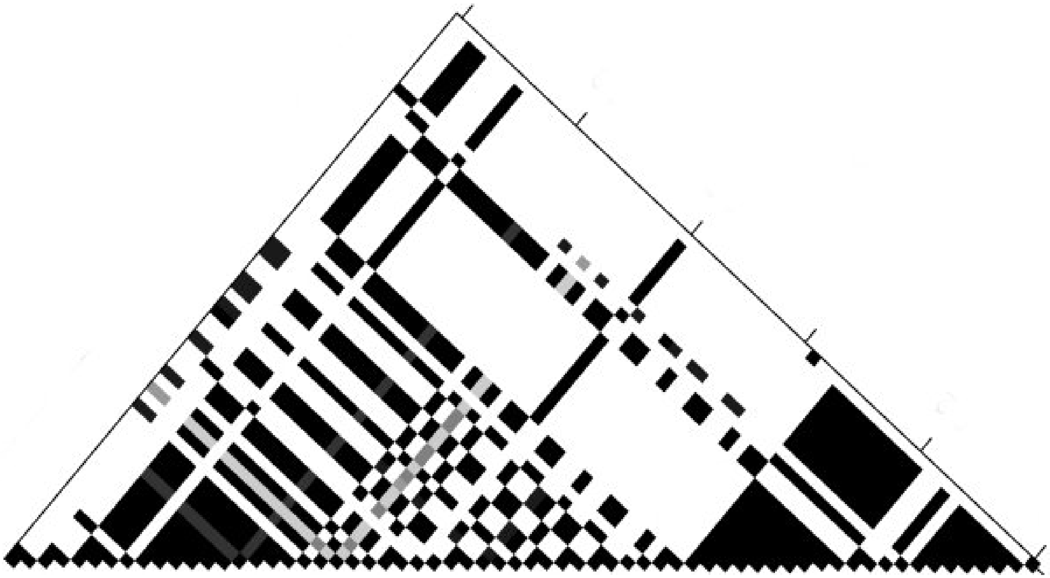
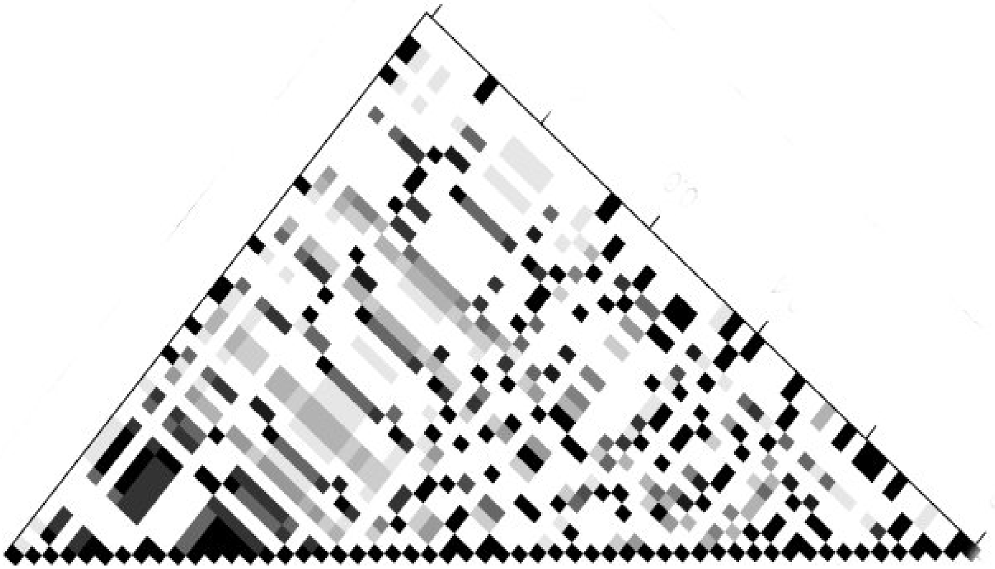


Figure 2, panel a. Homogeneously mixing data set.



```
  * **   ****  **   *  * * ***  *   *  *  * * **    * **         1
  * **   ****  **   *  * * ***  *   *  *  * * ** *  * **         5
  * **   *  *  **   * *    *       *    *  *    * ** *   ***     23
* * **   *  *  **   * * *  *       *    *  *    * ** *   ***     1
  * **  **  *  **   * *   *   ** *** *  **   *  * ** *   ***  ** 9
 * *  **     **  *** **   *   ** * * ** **   *  * ** *  ******   1
```

Figure 2, panel b. Island model data set.



```
    * *      * *  *                 * *      **         7
           *       *    *   ** *  * *                   24
           *       *    *   ** *  * *              *    1
**  *  * **** ** *   * *       *  ***          *        2
    * *      * *  *                 *  *****    *** *   1
        *  *  **    *   **     **    * ******   *** *   4
  ** *       * **     *   **  ** *      ****      * **  1
```

Figure 2, panel c. stepping stone model.

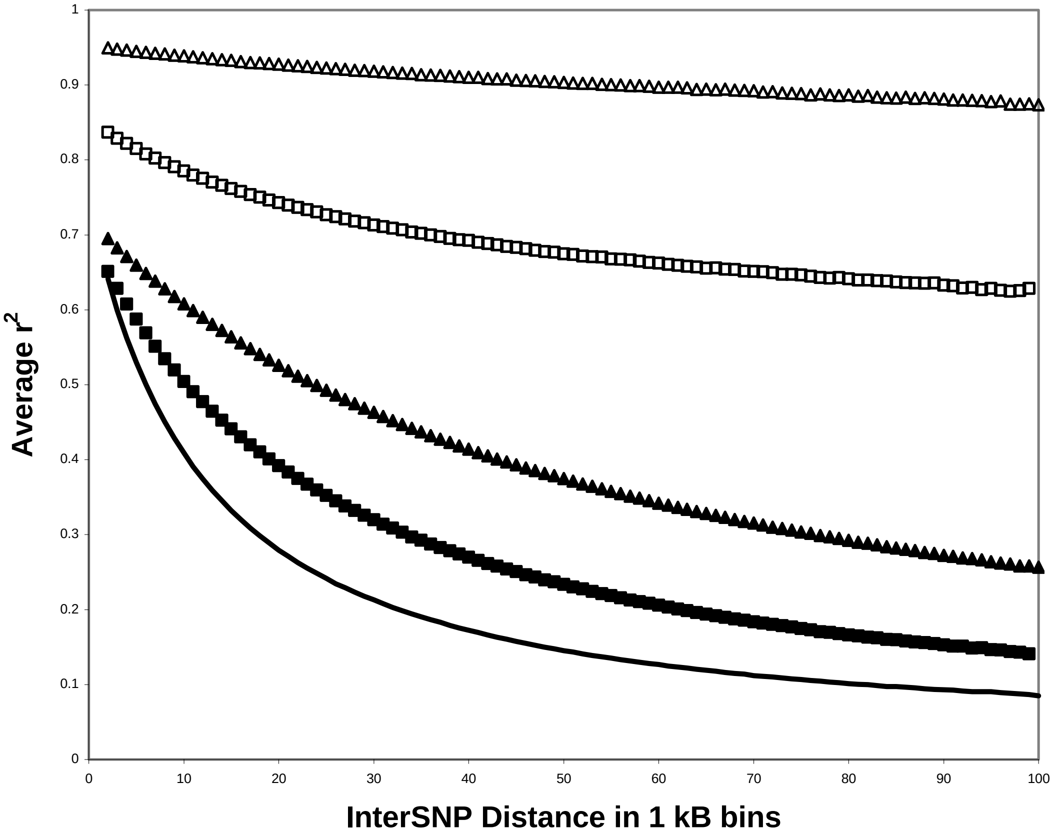


Figure 3, panel a, Decay of r^2 when *4Nm*=1

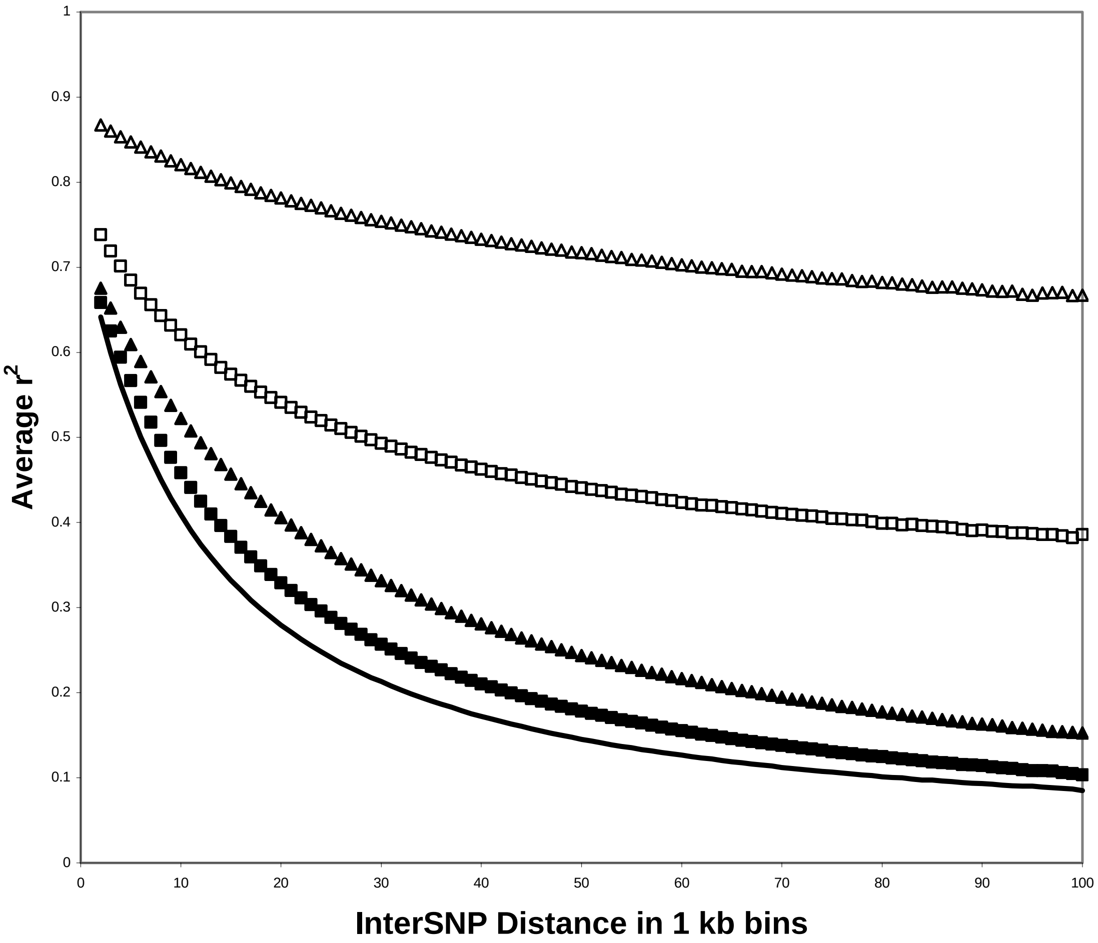


Figure 3, panel b, Decay of r^2 when *4Nm*=3.

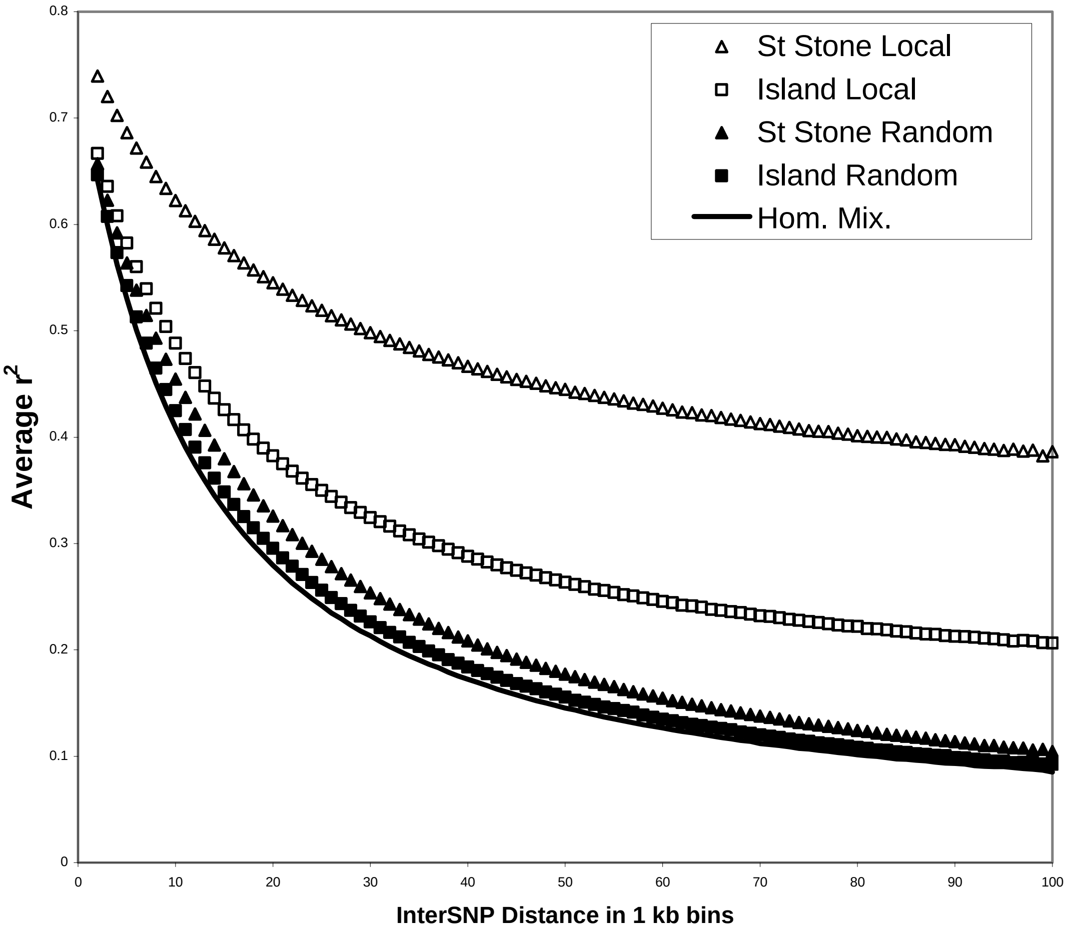


Figure 3, panel c, Decay of r^2 when *4Nm*=10.

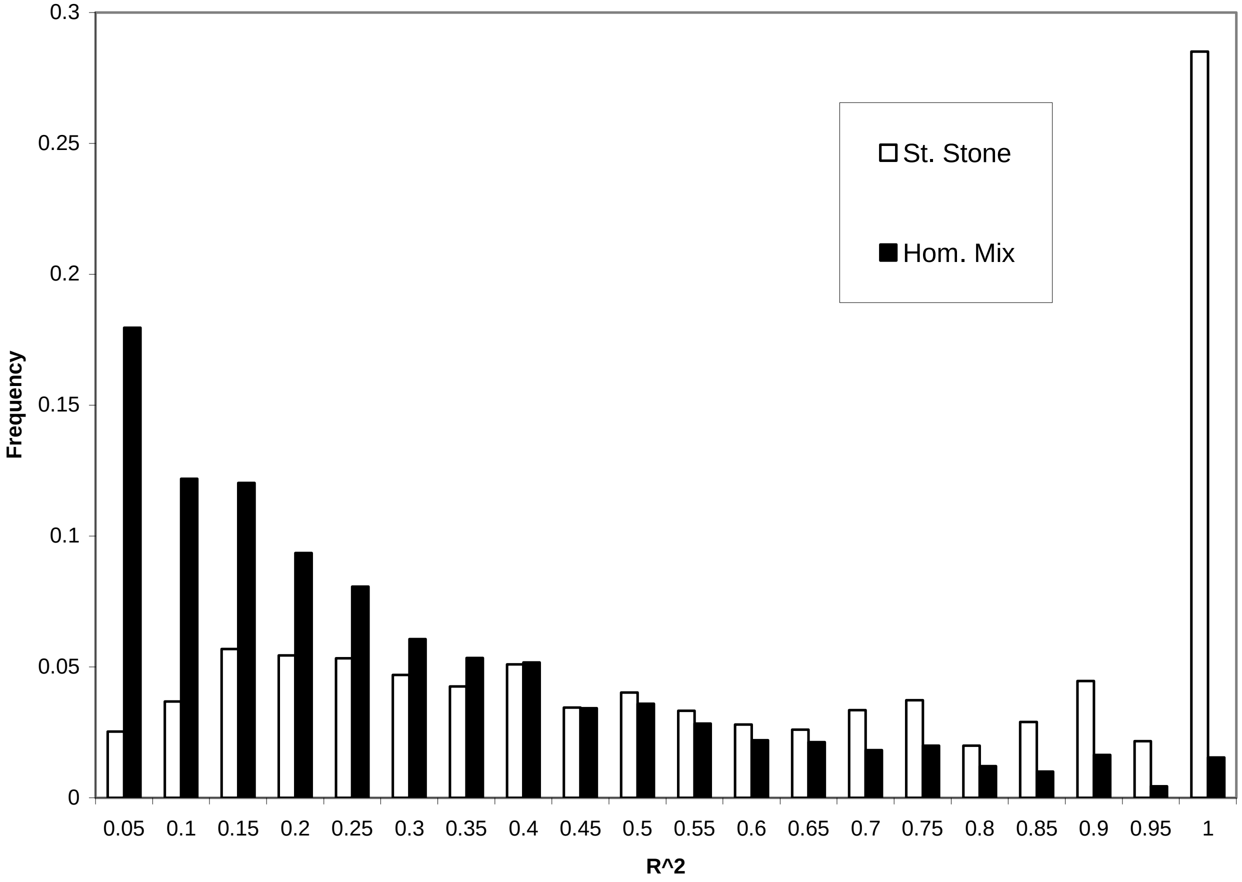


Figure 4, panel a, distance 10-11 Kb.

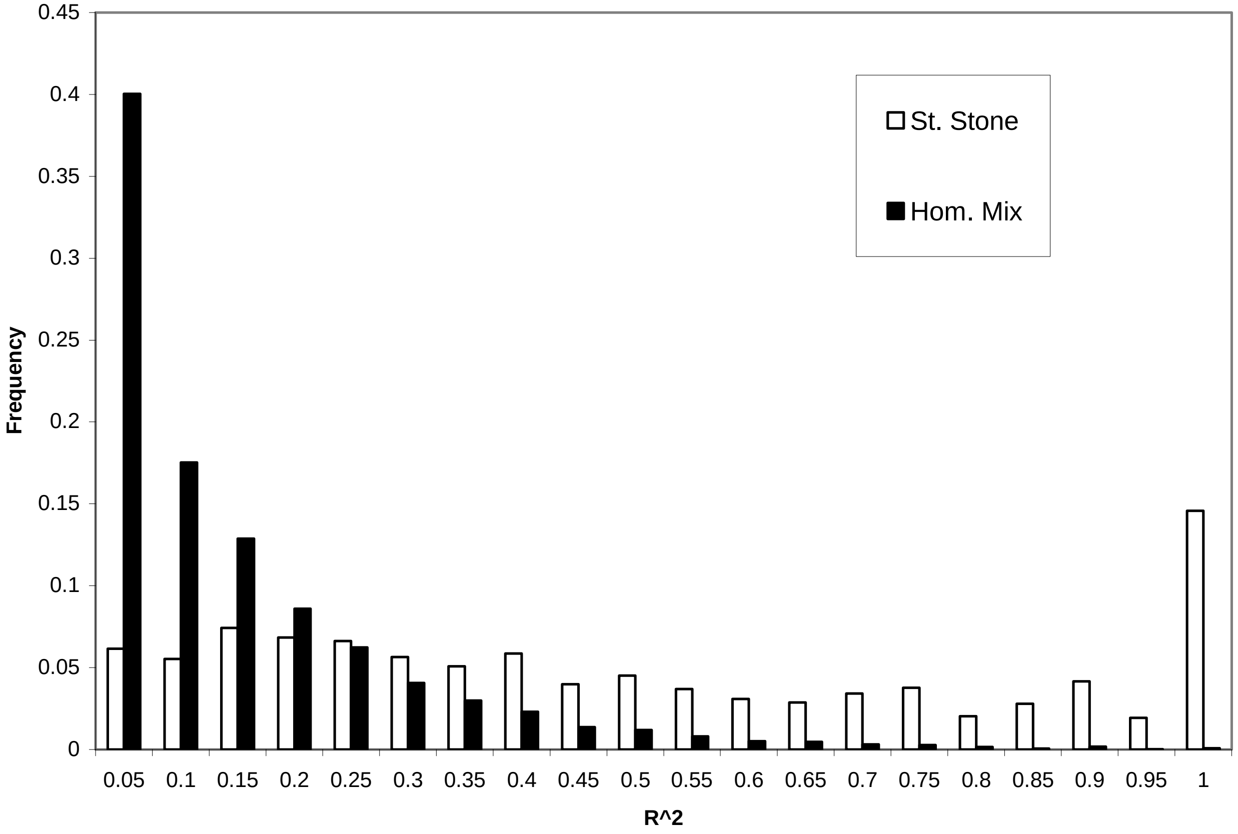


Figure 4, panel b, distance 30-31 Kb.

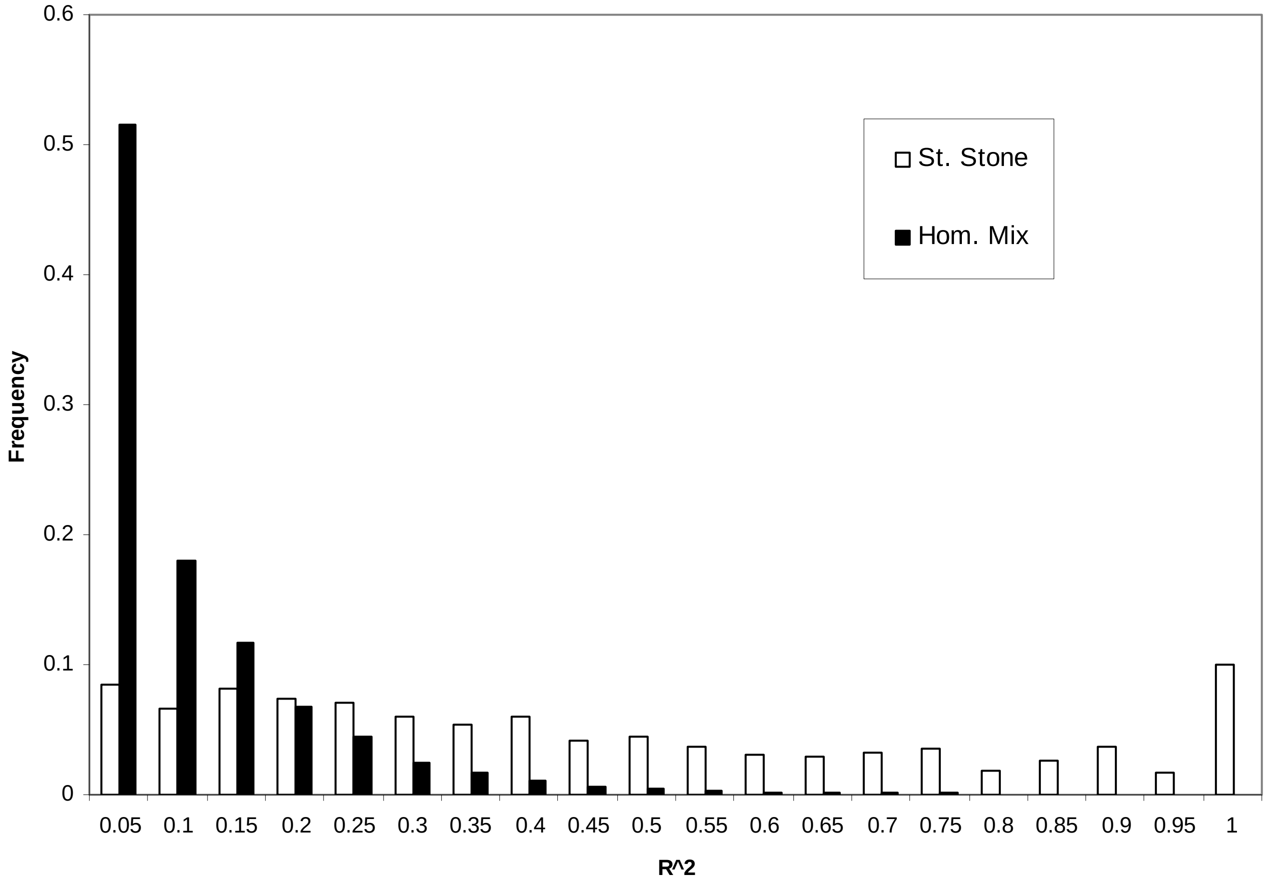


Figure 4, panel c, distance 50-51 Kb.

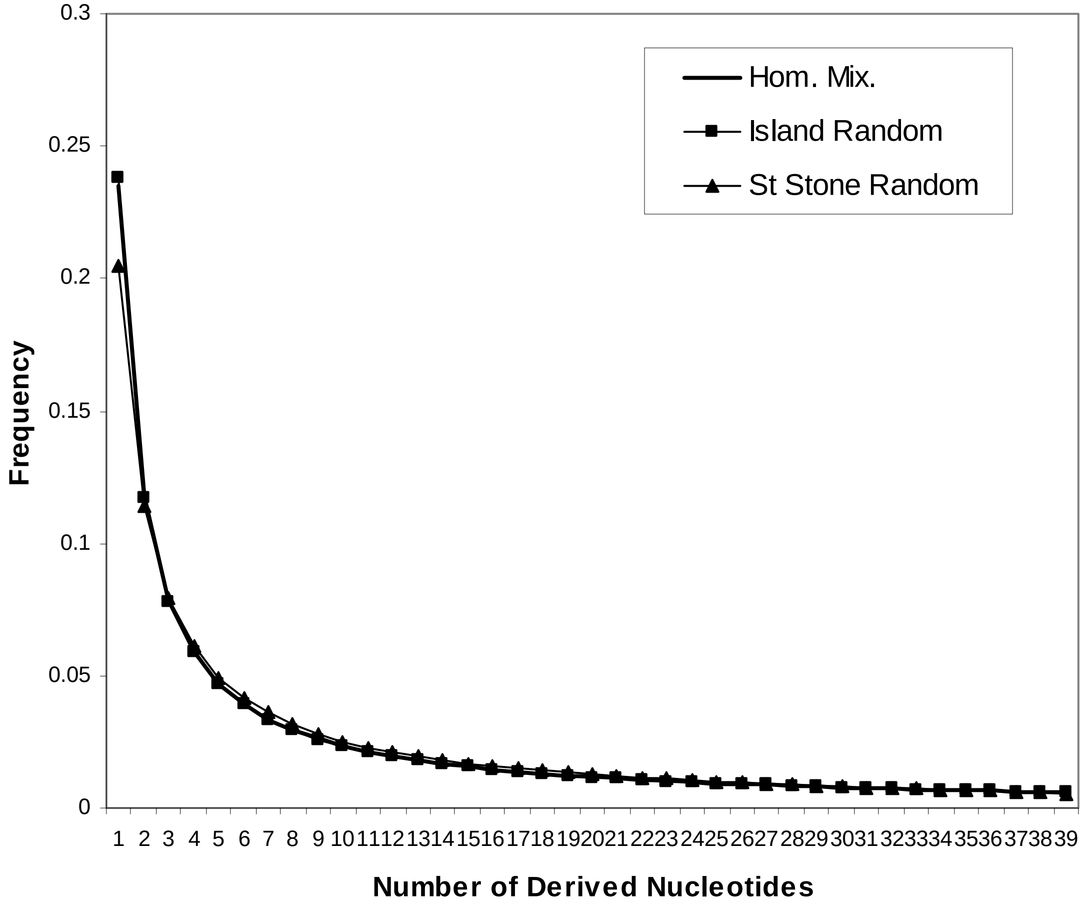


Figure 5, panel a, Site frequency spectrum for random samples.

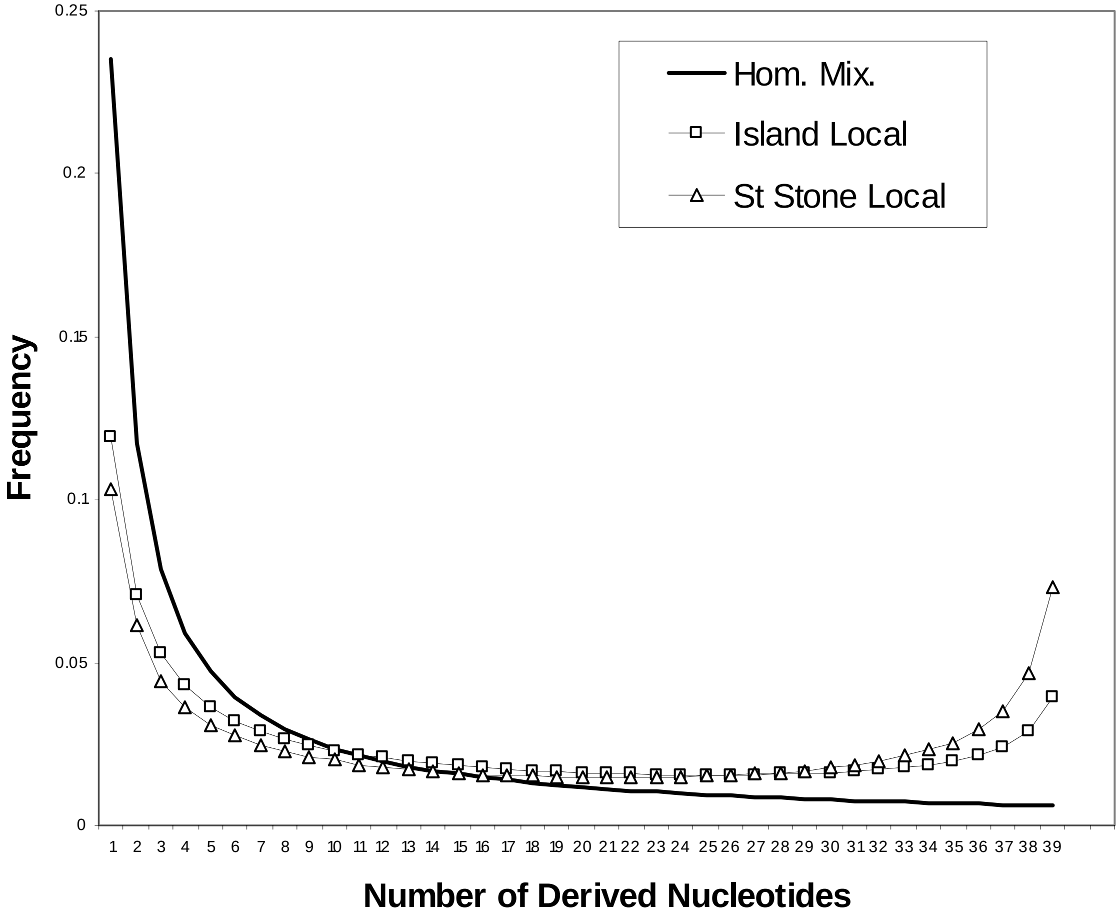


Figure 5, panel b, Site frequency spectrum for local samples *4Nm*=1.

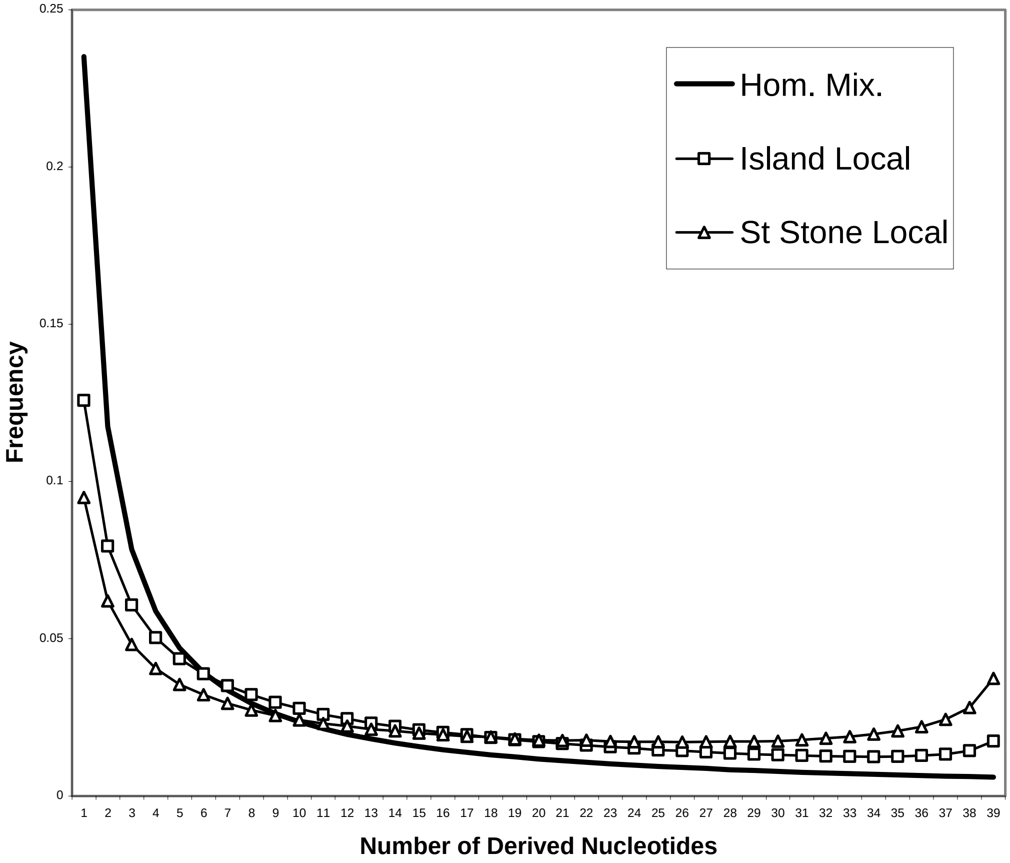


Figure 5, panel c, Site frequency spectrum for local samples *4Nm*=3.

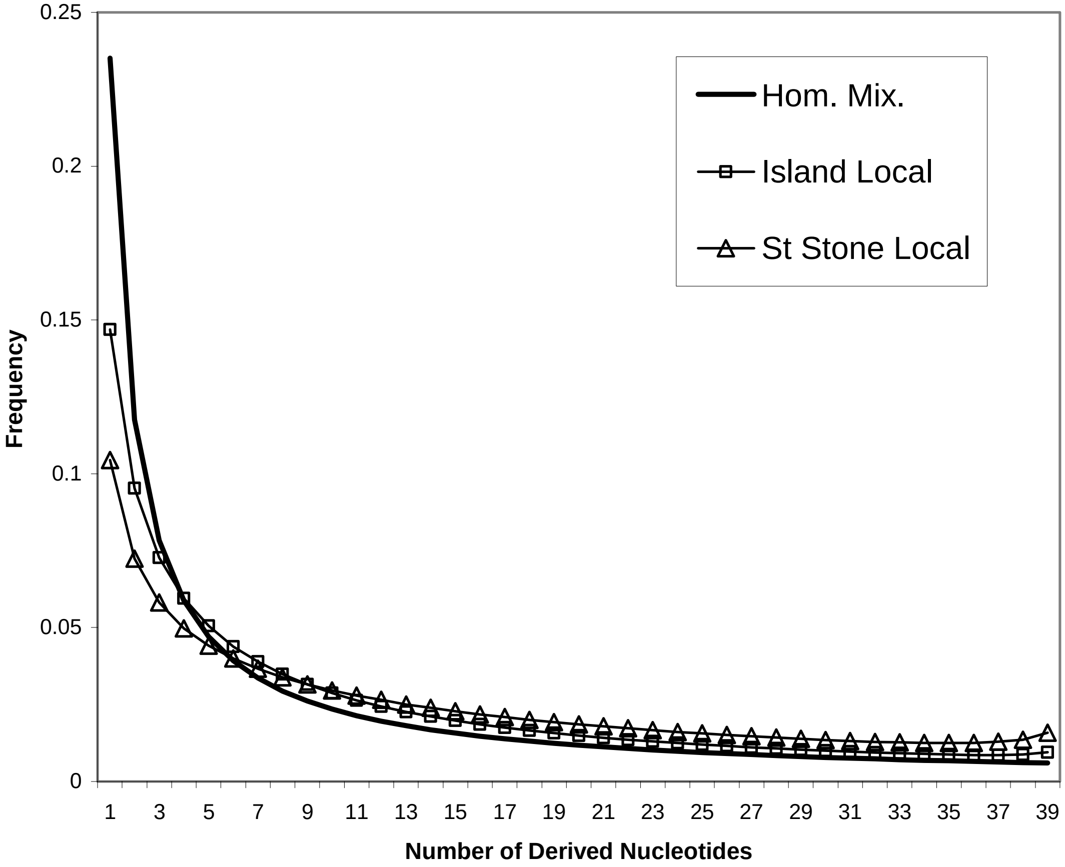


Figure 5, panel d, Site frequency spectrum for local samples *4Nm*=10.

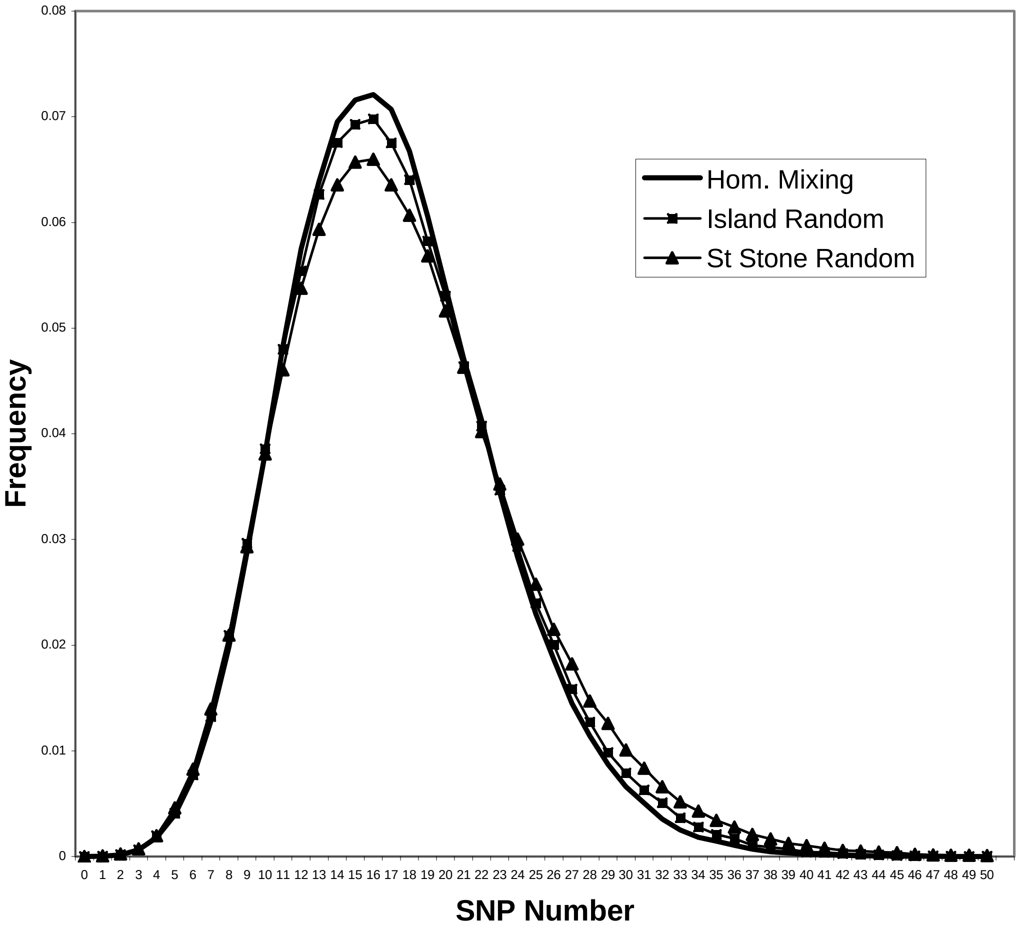


Figure 6, panel a, random samples.

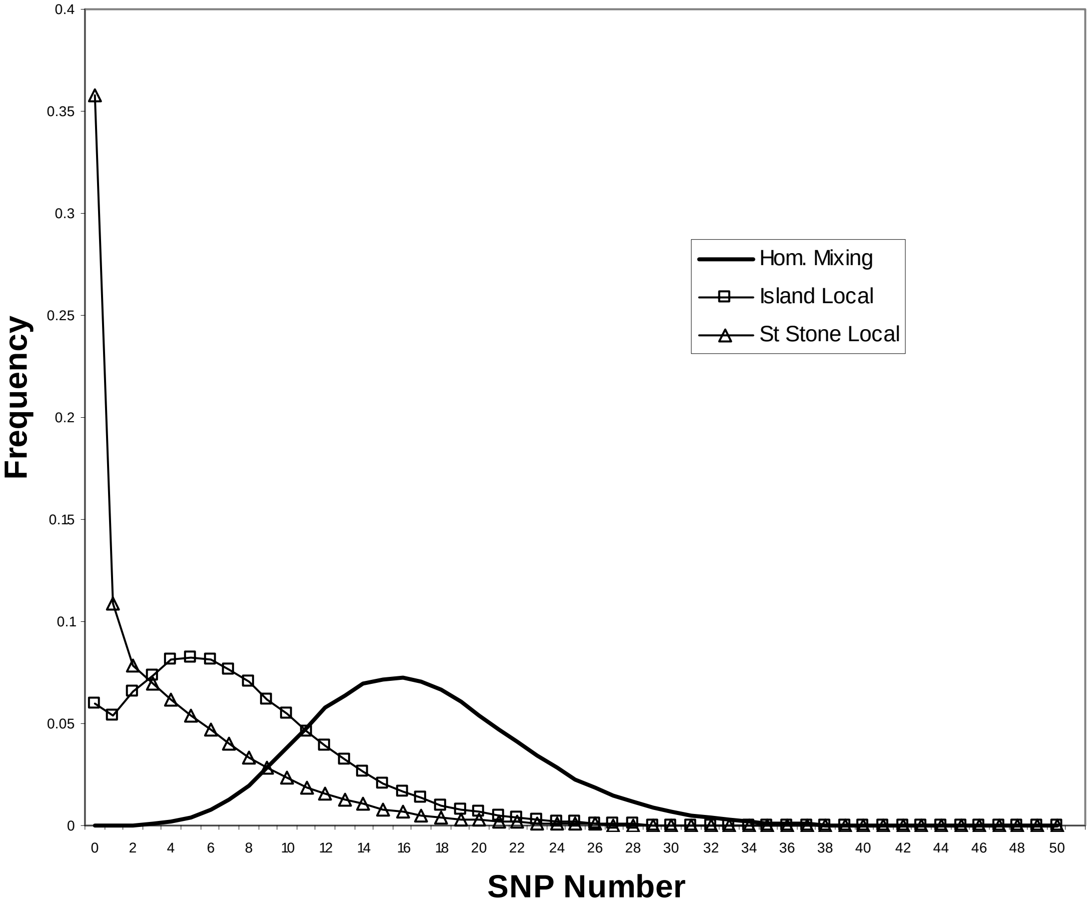


Figure 6, panel b, *4Nm*=1.

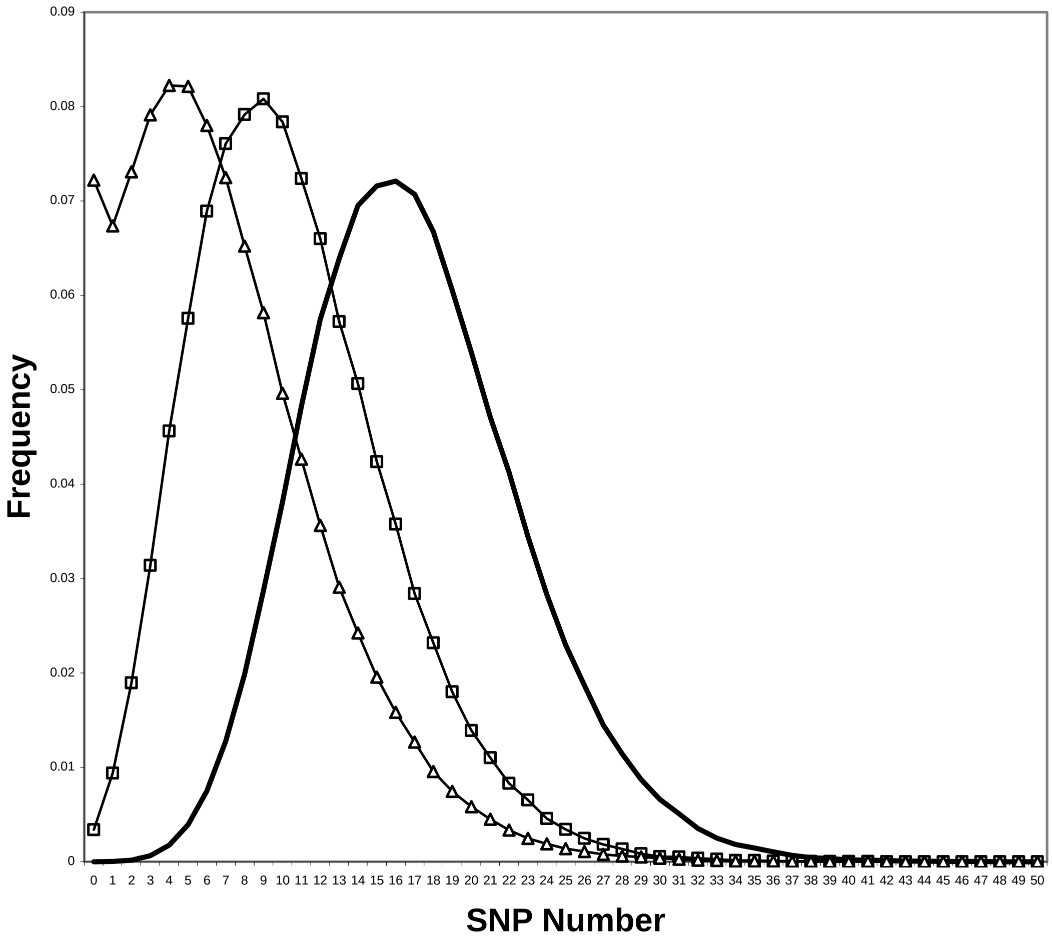


Figure 6, panel c, *4Nm*=3.

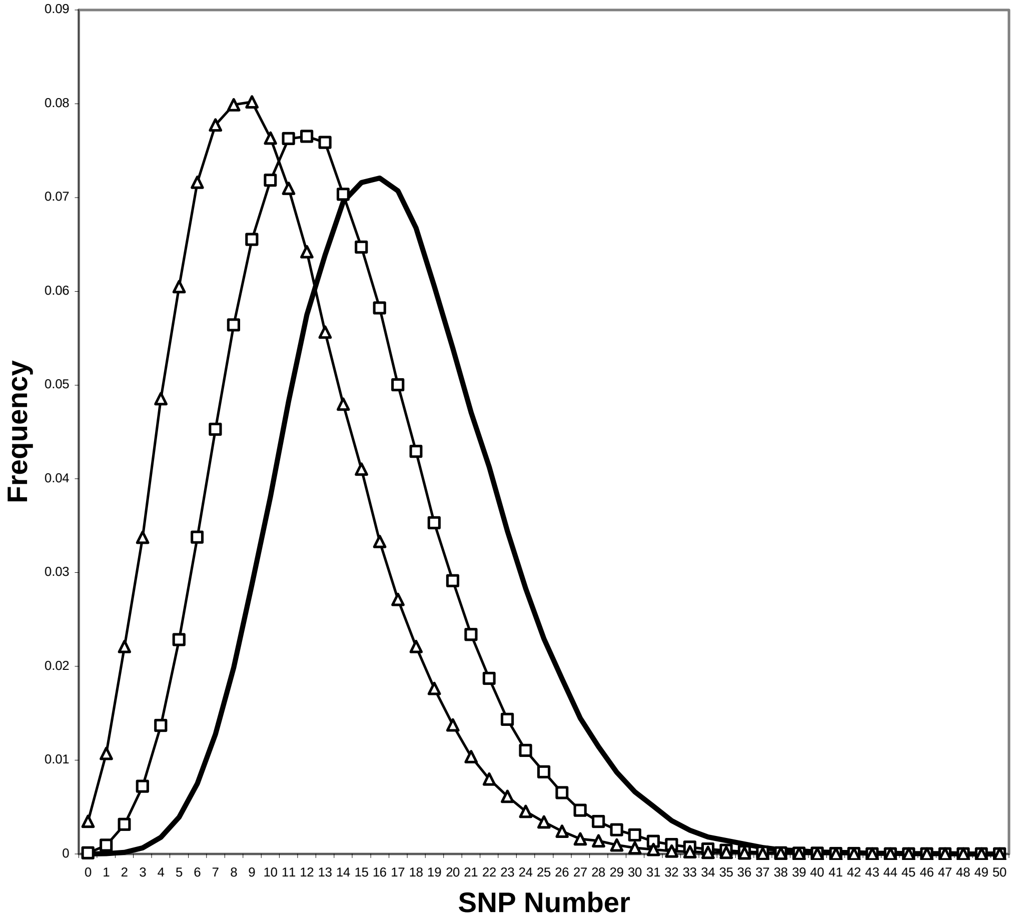


Figure 6, panel d, *4Nm*=10.

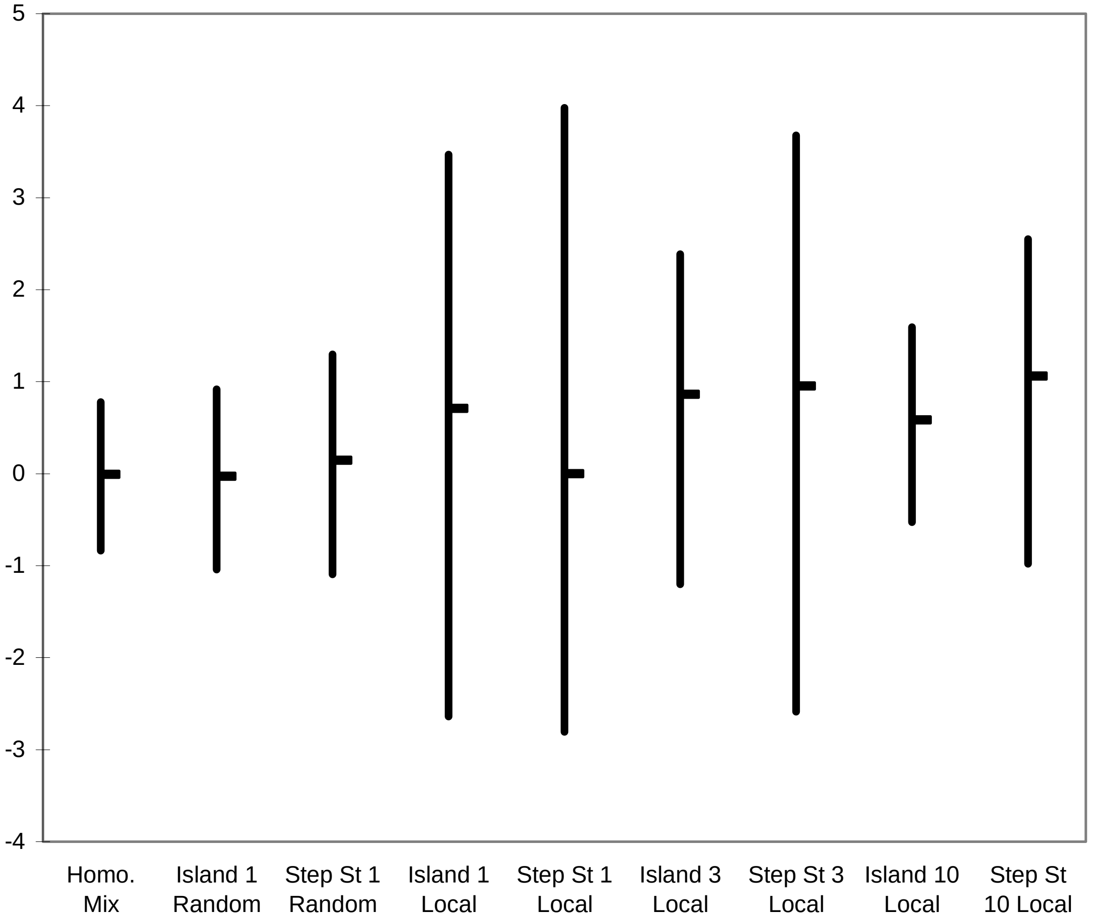


Figure 7, Tajima's *D*, median and 95% interval.

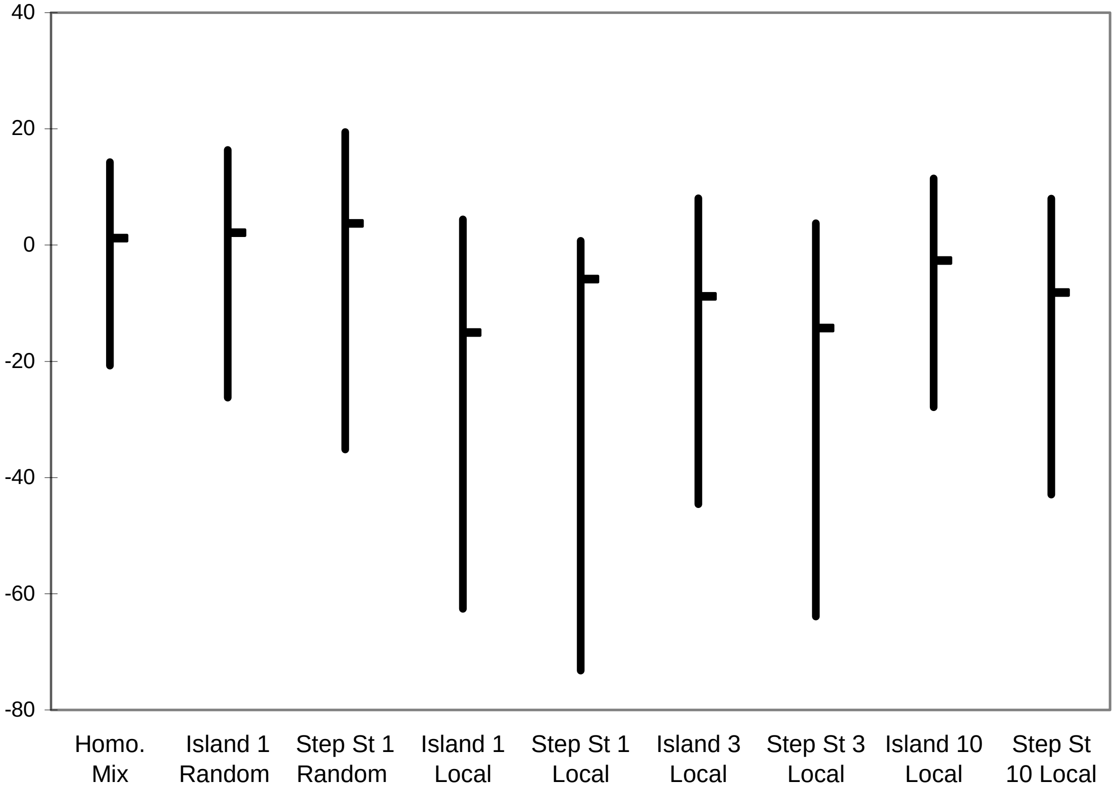


Figure 8. Fay and Wu's *H*, median and 95% interval.

FIGURE CAPTIONS

Figure 1. Plot of 1 over the coalescence rate for our concrete example. L = 10, N = 25, m = 0.1, $\sigma^2 = 0.5$ and β = 0.2. The first phase ends at time $L^2/(4\pi m\sigma^2) = 1591$. After that time the value is $N_e = 9829$. This graph is similar to simulation results presented in Figure 4 of Wilkins (2004), except that he uses a logarithmic time scale which turns the straight line into an exponential.

Figure 2. Sample data sets for (a) a homogeneously mixing population, (b) island model with $4Nm$ = 1, and (c) stepping stone model with $4Nm$=10. Asterisks mark nucleotides that are different from the ancestral state. The numbers at the right indicate how many times each haplotype was observed. Each square in the triangular plot represents r^2 value for a pair of SNPs. r^2 was calculated for all pairs of SNPs and shaded according to the magnitude. The color scale used was a gradation from white to black, with white representing r^2 = 0 and black representing r^2 = 1.

Figure 3. Decay of r^2 for (a) $4Nm$ = 1, (b) $4Nm$ = 3, and (c) $4Nm$ = 10. Here and in all of our figures, the expectation for a homogeneously mixing population is given by a line. Island model results are graphed with square, stepping stone results with triangles, with filled symbols for random samples and hollow symbols for local samples.

Figure 4. Comparison of the distribution of r^2 for a homogeneously mixing population and a stepping stone local sample with $4Nm$ = 10 for distances (a) 10-11 Kb, (b) 30-31 Kb, (c) 50-51 kB

Figure 5. Site frequency spectra for (a) random samples with *4Nm* = 1, and local samples with (b) *4Nm* = 1, (c) *4Nm* = 3, and (d) *4Nm* = 10. Symbols are as described in the caption to Figure 3.

Figure 6. SNP Density under random sampling with *4Nm* = 1, and local samples with (b) *4Nm* = 1, (c) *4Nm* = 3, and (d) *4Nm* = 10. . Symbols are as described in the caption to Figure 3.

Figure 7. Medians and 95% intervals for Tajima's *D* statistic for our spatial structures and sampling schemes.

Figure 8. Medians and 95% intervals for Fay and Wu's *H* for our spatial structures and sampling schemes.