Genetic Analysis of Human Evolutionary History with Implications for Gene Mapping

David E. Reich
Department of Zoology
University of Oxford

Doctoral Thesis
November 11, 1999

Genetic Analysis of Human Evolutionary History With Implications for Gene Mapping

Submitted for the Degree of D.Phil., Michaelmas 1999, by David E. Reich

(Department of Zoology, St. Catherine's College)

Abstract

Genetic variation contains detailed and quantitative evidence about the history of populations. The historical traces of demographic growth and contraction, as well as the history of human disease, have left traces on the patterns of modern variation and can be studied by sampling present-day populations. However, the data sets that are necessary in order to take full advantage of this living archaeological record have not been available until recently. The quality and quantity of data have increased dramatically during the past decade because of the identification of polymorphisms, including SNPs and microsatellites, that are much more amenable to mathematical modeling and efficient genotyping than earlier marker systems. The research in this thesis has been carried out in response to the need to provide new methods of analysis to match the new types of data. Chapter 1 describes multilocus tests of demographic history and their application to real data. Chapter 2 describes how the pattern of linkage disequilibrium around a disease-predisposing mutation can be used to estimate the date of a mutation—that is, the age of the most recent common ancestor of a set of modern samples. Finally, Chapter 3 draws several direct connections between human evolutionary history and medical genetics.

Table of Contents

Introduction

Genetic variation contains detailed and quantitative evidence about the history of populations. The historical traces of demographic growth and contraction, the history of specific subgroups, and the patterns of relationships among populations have all left their traces on the patterns of modern variation, which can be studied by sampling present-day populations. However, the data sets that are necessary in order to take full advantage of this living archaeological record have not been available until recently. The quantity of data has increased dramatically in the past decade because of the availability of genomic technologies for the efficient identification and genotyping of polymorphisms. The quality of data has increased because the markers can be selected at precisely defined genomic locations, and because short-tandem repeat polymorphisms (microsatellites), and single nucleotide polymorphisms (SNPs), are much more amenable to mathematical modeling than earlier marker systems, including allozymes and blood group antigens (which are also more likely to be under the influence of selection). The vastly improved quality of the new data sets thus makes it possible not only to ask the simple questions about population history, but also to get answers to some of the complex and interesting questions that, using previous methods, could not be adequately addressed. Simultaneously, important connections have emerged between the genetic study of human population history for its own sake and the applications of those insights to improve medical genetics.

The research in this thesis has been carried out in response to the need to provide new methods of analysis to match the new types of data. To extract subtle information from the data, the methods should be theoretically innovative and deeply sensitive to experimental issues: to the genetic system, the sampling strategy, and the historical questions at hand. Every analysis described in this thesis therefore used a rather similar analytical approach. Each project began with a hypothesis about how to extract the information of interest. To obtain as thorough as possible an understanding of the genetic system, a coalescent computer simulation was used in order to investigate a wide range of reasonable historical scenarios (simulations were based on a core algorithm by R. HUDSON (1991), modified to conform to a specific genetic system). Finally, statistical tests, sensitive to the parameters of interest, were introduced, and additional simulations were used to investigate the efficacy and behavior of the statistics. Thus, methods of data analysis were derived that were sensitively tailored to the type of practical data sets that were available.

The thesis is a compilation of research papers, four of which are already published (references are given in Appendix 1 and at the beginning of each Chapter). Permission to submit a thesis in this format was granted by the Faculty Board of Biological Sciences at the University of Oxford. The papers are united by a common interest in how genetic variation can be used to study the histories of human populations, and the implications of such population genetic techniques for medical genetics.

Chapter 1 describes multilocus tests of demographic history. Tests of this type are important because they can elucidate the histories of growth, contraction, and

subdivision in natural populations. But they are also interesting from a medical perspective, because populations that have undergone severe recent bottlenecks exhibit linkage disequilibrium around disease genes that can be used for fine-scale genetic mapping (for example, the Finnish population; DE LA CHAPELLE and WRIGHT 1998). The first generation of tests of demography were based on variation at a single locus, usually mitochondrial DNA (mtDNA; ROGERS and HARPENDING 1992). Single-locus data, however, cannot distinguish a signal of expansion due to true population growth, which would appear throughout the genome, from a signal of expansion that is restricted to a single locus, perhaps due to strong historical selection acting on one of the many genes that are fully linked to each other on the mitochondrial DNA locus. The second generation of tests of demographic history therefore used data from multiple unlinked loci (REICH and GOLDSTEIN 1998; KIMMEL and CHAKRABORTY 1998; DIRIENZO *et al.* 1998), which can make the distinction between locus-specific and population-wide expansion as a result of the fact that gene histories are different across loci. Chapter 1 presents two microsatellite-based tests of demography and applies them to data to show evidence for a Paleolithic human population expansion in Africa. The last section of Chapter 1 extends multilocus tests to SNPs, which have an obvious advantage over microsatellites because their mutation process is better understood. An additional, surprising advantage of SNP-based tests of demography is that they have relatively greater power to detect expansions, per locus screened, than tests based on microsatellite markers that are thought, for most other purposes, to contain more information per locus than SNPs (e.g., KRUGLYAK 1997).

Chapter 2 describes how the pattern of breakdown of linkage disequilibrium around a mutation can be used to estimate the date of a mutation—that is, the age of the most recent common ancestor of a set of modern samples. To make this calculation, the full potential of modern data sets is required, and in particular the ability to identify polymorphisms at appropriate and precisely defined distances from the mutation. A number of methods exist for making date estimates (e.g., THOMPSON 1976; RISCH *et al.* 1995; SLATKIN and RANNALA 1997; REICH and GOLDSTEIN 1999). However, what distinguishes the present method is the ease with which it can be applied to data, and the fact that it can produce not only a best estimate, but a confidence interval for the date of a mutation. The method is applied to two examples in Chapter 2: the CCR5-Δ32 mutation, which confers resistance to HIV infection among Caucasians, and mutations in the Factor XI gene, which occur in relatively high frequency among Ashkenazi and Iraqi Jews (STEPHENS *et al.* 1998; GOLDSTEIN *et al.* 1999). The conclusions are of medical as well as historical interest. In particular, the very recent age of the CCR5-Δ32 mutation, in conjunction with the large number of copies of the mutation present in modern Europeans, indicates powerful selection in favor of the mutation in Europeans during the past several thousand years. It therefore seems likely that the CCR5-Δ32 mutation confers resistance to an important disease other than AIDS, implicating the CCR5 chemokine receptor in the etiology of that disease. This provides a direct example of how the study of the history of populations can elucidate biomedical issues, and even provide clues that may eventually lead to disease treatment.

Chapter 3 synthesizes some of these ideas by directly connecting the history of populations to medical genetics. The chapter presents a method for detecting genetic associations in a case-control study in the face of population stratification. Case-control studies usually assume that subjects are drawn from a single, ethnically homogenous population. However, false-positive associations can arise between a disease and markers that are completely unlinked to the disease locus when populations are ethnically heterogeneous. To investigate how various scenarios of stratification affect case-control studies, computer simulations are used, and on the basis of simulation results, a method is proposed for using the level of genetic association observed at the unlinked markers, in the same cases and controls in which the candidate association was found, to calculate an empirical significance level for the detected association.

The second part of Chapter 3 investigates the issue of how a genetic association between linked markers in one population is preserved in other, distantly related populations. The degree to which an association is preserved can be used as a type of genetic distance, with advantages for studying population history compared to traditional distances based on allele frequencies. From a medical perspective, the rate of decay of a genetic association, as a function of the degree of relatedness of populations, is also interesting because it predicts whether a disease association that is discovered in one population is likely to remain as a signature of disease in other populations. While the final part of Chapter 3 is considerably more speculative than previous parts of the thesis—it is as much a projection of future work as a finished

research project—it nevertheless reiterates the direct relevance of population genetics to medical genetics.

References:

DE LA CHAPELLE, A. and F. A. WRIGHT, 1998 Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc. Natl. Acad. Sci. USA **95:** 12416-12423.

DIRIENZO, A. P. DONNELLY, C. TOOMAJIAN, B. SISK, A. HILL, *et al.*, 1998 Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics **148:** 1269-1284.

GOLDSTEIN, D. B., D. E. REICH, N. BRADMAN, S. USHER, U. SELIGSOHN and H. PERETZ, 1999 Age estimates of two common mutations causing Factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. Am. J. Hum. Gen. **64:** 1071-1075.

HUDSON, R. R., 1991 Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) **7:** 1-44 (Oxford University Press, Oxford).

KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS and L. B. JORDE, 1998 Signatures of population expansion in microsatellite repeat data. Genetics **148:** 1921-1930.

KRUGLYAK, L., 1997 The use of a genetic map of biallelic markers in linkage studies. Nature Genetics **17:** 21-24.

REICH, D. E. and D. B. GOLDSTEIN, 1998 Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. U.S.A. **95:** 8119-8123.

REICH, D. E. and D. B. GOLDSTEIN Detecting association in a case control study in the face of population stratification. Submitted to Nature Genetics, 23/8/99.

REICH, D. E., M. W. FELDMAN and D. B. GOLDSTEIN, 1999 Statistical properties of two tests that use multilocus data sets to detect population expansion. Mol. Biol. Evol. **16:** 453-466.

RISCH N., D. DE LEON, L. OZELIUS, P. KRAMER, L. ALMASY, *et al.*, 1995 Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nature Genetics **9:** 153-159.

ROGERS, A. R. and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552-569.

SLATKIN, M. and B. RANNALA, 1997 Estimating the age of alleles by use of intraallelic variability. Am. J. Hum. Gen. **60**: 447-458.

STEPHENS, J.C., D. E. REICH, D. B. GOLDSTEIN, H. D. SHIN, M. W. SMITH, *et al.*, 1998 Dating the origin of the CCR5-Δ32 AIDS resistance allele by the coalescence of haplotypes. Am. J. Hum. Gen., **62:**1507-1515.

THOMPSON, E. A., 1976 Estimation of age and rate of increase of rare variants. Am. J. Hum. Gen. **28:** 442-452.

Chapter 1

Multilocus Tests of Demographic History

Part A: "Genetic evidence for a Paleolithic human population expansion in Africa was published in *Proc. Natl. Acad. Sci. USA*, **95:**8119-8123 (1998). A correction was published in the same journal, **95:**11026a-11026a (1998). The corrections are incorporated into the text. Some changes in the text, as well as the "Addendum," are entirely new.

Part B: "Statistical properties of two tests that use multilocus data sets to detect population expansions" was published in *Mol. Biol. Evol.*, **16:**453-466 (1999), and explores the behavior of the tests presented in Part A.

Part C: "Single nucleotide polymorphisms as tools for studying demographic history," extends the multilocus approach for studying demographic history, developed in Parts A and B, to SNP markers. The manuscript was submitted to *Genetics* on November 4, 1999.

Genetic Evidence for a Paleolithic Human Population Expansion in Africa

David E. Reich and David B. Goldstein

Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK.

ABSTRACT

Human populations have undergone dramatic expansions in size, but other than the growth associated with agriculture, the dates and magnitudes of those expansions have never been resolved. Here, we introduce two new statistical tests for population expansion, which use variation at a number of unlinked genetic markers to study the demographic histories of natural populations. By analyzing genetic variation in various aboriginal populations from throughout the world, we show highly significant evidence for a major human population expansion in Africa, but no evidence for expansion outside of Africa. The inferred African expansion is estimated to have occurred between 44,000 and 570,000 years ago, certainly before the Neolithic expansions, and probably before the splitting of African and non-African populations. In showing a significant difference between African and non-African populations, our analysis supports the unique role of Africa in human evolutionary history, as has been suggested by most other genetic work. In addition, the missing signal in non-African populations may be the result of a population bottleneck associated with the emergence of these populations from Africa, as postulated in the "Out of Africa" model of modern human origins.

—

Genetic approaches to the study of human population expansions previously have focused on variation at a single genetic locus, the “control region” of mitochondrial

DNA (mtDNA) (1). However, in the study of demographic history, single locus studies suffer from pronounced statistical and biological limitations. The statistical problem is that the conclusions rely on only one particular realization of a gene genealogy, the "tree" determining the ancestral relationships among a set of alleles. The biological problem is that there are a large number of functional genes in the mitochondrion (2), and because there is complete linkage, a selective sweep for any one of the genes may lead to a spurious signal of expansion. Genome-wide data sets provide a promising alternative, overcoming most of the statistical and biological limitations inherent in single locus studies. If genome-wide data sets are used, population expansions will be distinguishable from natural selection because expansions affect all loci, whereas selection only affects loci tightly linked to the selected locus.

DATA AND ANALYSIS

The markers we use in our tests are "microsatellites," which were first identified in large-scale gene mapping projects but are increasingly used for inferring population parameters (3). Microsatellites, which exhibit extensive "length" variations, are widely distributed throughout the genome, seem to be selectively neutral, and appear to conform reasonably well to a simple mutation process (stepwise mutation model), whereby mutations change the length by one or occasionally two units (3). On the basis of this mutation model, we have developed two statistical tests to discern whether populations have been constant or growing in size.

Within-Locus *k*-Test for Population Expansion. For a population of constant size, gene genealogies tend to have a single ancient bifurcation, implying that most pairs of alleles are either closely or distantly related, with few in between (4). The distribution of allele lengths, therefore, has discrete peaks that correlate with the descendants of each side of the ancient bifurcation (Fig. 1). For a growing population, in contrast, most of the bifurcations tend to date back to the time of expansion—the genealogical tree is "comblike," and the resultant allele length distribution is more smoothly peaked (Fig. 2).

To differentiate between the ragged, multi-peaked distribution expected for a constant population size, and the smooth, single-peaked distribution expected for an expansion, we construct a statistic, denoted *k*, which is a decreasing function of the fourth central moment of the sample, $1/n\sum_{i=1}^{n}(X_i - \overline{X})^4$, where *n* is the number of chromosomes, $\overline{X}$ is the average allele length, and the X_i values are the individual allele lengths. Because the fourth central moment is related to the kurtosis, which increases with peakedness, the statistic *k* tends to decrease systematically with the degree of peakedness caused by expansion (5) (note that the kurtosis is equal to the normalized fourth central moment plus an overall constant).

To set the parameters of the *k*-statistic empirically, we use computer simulations based on the coalescent algorithm of R. Hudson (6). Genealogies are traced backward in time from the sampled individuals to their most recent common ancestors, and stepwise mutations are distributed along the genealogies according to a random Poisson process. We use the results of the simulations to set the parameters of the *k* statistic so that the probability of a locus being positive when the population is

constant in size is constrained to a narrow range (between 0.515 and 0.55) for sample sizes greater than 10 and for a wide variety of population sizes and mutation rates. In this way, we derive the statistic $k = 2.5*Sig^4 + 0.28*S^2 - 0.95/n - Gam_4$, where S^2 is the sample variance and Sig^4 and Gam_4 are unbiased estimators for the variance squared and fourth central moment, respectively. Note that Sig^4 and Gam_4 were derived specifically for this analysis (5), and their validity was checked by computer simulation.

$$Sig^4 = \frac{(n^2 - 3n + 3)}{n(n-1)(n-2)(n-3)}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2 - \frac{1}{(n-2)(n-3)}\sum_{i=1}^{n}(X_i - \bar{X})^4 \qquad [1]$$

$$Gam_4 = \frac{(n^2 - 2n + 3)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(X_i - \bar{X})^4 - \frac{(6n-9)}{n(n-1)(n-2)(n-3)}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2 \qquad [2]$$

To implement the approach, we set the probability of a positive k conservatively at 0.515, and use a simple one-tailed binomial test to determine whether fewer loci were associated with a positive k than would be expected for a constant sized population. Because the expectation of k decreases with increasing kurtosis, such a reduction in the number of positive k values can be interpreted as a sign of population expansion.

Interlocus *g* Test for Population Expansion. The second technique for detecting an expansion focuses on a feature of multi-locus data sets that has no analog in studies of a single gene. When populations are of constant size, the dates of the most ancient bifurcations are subject to considerable variation from locus to locus (Fig. 1). Under conditions of growth, the most ancient bifurcations tend to have similar dates at all loci (Fig. 2). To distinguish between the demographic scenarios,

we note that the characteristic differences associated with demography—so evident in a comparison of gene genealogies in Figs. 1 and 2—also will be reflected in the variance of the variance of the allele length distributions. Specifically, because the variance of an allele length distribution depends mainly on the ages of the few most ancient bifurcations (7), the variance of the variance is expected to be larger for constant sized populations than for growing ones.

To test for this effect statistically, we take advantage of the fact that there is an analytical expectation for the variance of the variance in a constant sized population, $4/3E[V_j] + 1/6E[V_j]$ (8, 9). To estimate this quantity, we substitute $\overline{V}$, the average variance across loci, for $E[V_j]$. To formulate the test explicitly, we consider the ratio, *g*, of the observed value to the expected value.

$$g = \frac{\mathrm{Var}[V_j]}{\frac{4}{3}\overline{V}^2 + \frac{1}{6}\overline{V}} \qquad [3]$$

A sufficiently low value of *g* is taken as a sign of expansion. A useful and interesting feature of this ratio is that, as shown by the computer simulations used to calculate *P* values, its expectation and confidence intervals are essentially independent of $N\mu$ (mutation rate times population size) and nearly independent of sample size (5). For sample size greater than 25, a *g* ratio less than 0.35 is sufficient to reject the null hypothesis. A full look-up table of significant cut-offs is presented elsewhere (5).

Paleolithic Human Population Expansion in Africa. One tetranucleotide and one dinucleotide microsatellite data set, each of 30 unlinked loci, have recently become available, providing information about genetic diversity represented in

hundreds of individuals and several populations around the world (10, 11)*. In the tetranucleotide data, the "within locus" k test shows that only two populations give a significant signal of expansion, and they are both in Africa (San and Sotho-Tswana, both with P values of < 0.01) (Table 1). The interlocus g test applied to these data also suggests a difference between African and non-African populations: the four lowest g values are in Africa and clearly lower than those found elsewhere in the world.

In the dinucleotide data, the within locus k test produces no significant P values, possibly because, as demonstrated by computer simulations, the test loses power for higher values of the mutation rate (5) (Table 2). With the interlocus g test, however, the North-Central African population shows a significant sign of expansion ($P < 0.037$). The significance of the detected expansion increases even further, to $P < 0.006$, when we drop an exceptionally variable locus (D13S122), which has a variance in the world-wide sample of 89.2 compared to a narrower range of 1.0 to 17.2 for the other 29 loci (12). In contrast, non-African populations fail to show signs of expansion when the high-variance locus (D13S122) is dropped, although g-values for these populations are all below 1, suggestive of expansions.

Correction for Variation in the Mutation Rate. We can improve the power of the interlocus g test by taking into account variation in the mutation rate, denoted σ_μ^2, which is certainly substantial for microsatellite loci, and which weakens the interlocus test by increasing the g ratio. We obtain an estimate of the variation in the mutation rate by considering the statistic $\mathrm{Var}[(\delta\mu)^2]/((\delta\mu)^2)^2$, where $(\delta\mu)^2$ is a genetic distance introduced by Goldstein et al. (12), and is defined as the square of the

---

*One of the markers in the "dinucleotide" data set is actually supposed to be a tetranucleotide (11). When we examine the allele length distribution at the locus, however, we observe alleles every 2 rather than 4 base pairs. We therefore choose to treat the marker as dinucleotide.

difference between the mean allele lengths at a locus in two populations. The ratio of expectations is equal to $2(1+\sigma_\mu^2/\mu^2)$ (8), and we use computer simulations to show that the ratio $\mathrm{Var}[(\delta\mu)^2]/((\delta\mu)^2)^2$ does indeed approach 2 for $\sigma_\mu^2 = 0$ and as $(\delta\mu)^2$ becomes large, in the range of what is observed in human populations. We now use the equation $\mathrm{Var}[(\delta\mu)^2]/((\delta\mu)^2)^2 = 2(1+\sigma_\mu^2/\mu^2)$ to estimate σ_μ^2. Note that error in this estimate could arise because our calculations are based on analytical expectations for $(\delta\mu)^2$ and $\mathrm{Var}[(\delta\mu)^2]$, which both derive from an assumption that populations are in mutation-drift equilibrium. However, such inaccuracies in our estimate of σ_μ^2 are likely to be only moderate, since demography has only a small effect on expectations for the $(\delta\mu)^2$ genetic distance (13).

To extract σ_μ^2 from the data, we average $\mathrm{Var}[(\delta\mu)^2]$ and $(\delta\mu)^2$ over all pairwise comparisons of African and non-African populations, and then use the ratio of averages to calculate σ_μ^2. In order to obtain a confidence interval on the resulting estimate, it is necessary to know the effective number of independent observations of the $(\delta\mu)^2$ distance between African and non-African populations. The number of calculations of $(\delta\mu)^2$ that were actually made is likely to be considerably larger than the effective number because of correlation due to shared genealogical history among the populations in our data set. By assessing the shape of the genealogical tree relating the populations, and specifically noting that the total branch length of the tree is likely to be more than twice the branch length of any single Africa/non-Africa comparison, we conclude that we have made at least two independent calculations of $(\delta\mu)^2$ in both data sets. The confidence intervals on the estimate can then be calculated empirically using computer simulations.

From this procedure, we obtain a variance of the mutation rate of $\sigma_\mu^2 = 0.97\mu^2$ for the tetranucleotides (90% CI: $0.20\mu^2$-$3.10\mu^2$), and $\sigma_\mu^2 = 1.30\mu^2$ for the dinucleotides (90% CI: $0.14\mu^2$-$3.13\mu^2$). Incorporating these estimates into our analysis, and assuming that the mutation rate varies according to a truncated Gaussian distribution, the signal of expansion for the North-Central African population (dinucleotide data set) becomes highly significant at $P < 0.0016$, whereas *g* ratios for the non-African populations in both data sets come closer to the expectation for a constant population size. We can also combine the two approaches for correcting the interlocus *g*-test—dropping the high variance locus (D13S122) and adjusting for variance in the mutation rate among the remaining 29 loci ($\sigma_\mu^2 = 0.28\mu^2$)—to obtain a *P* value of < 0.0007 in the North-Central African population. Table 2 shows *P* values that result from this combined approach.

Effects of Inaccuracies in the Stepwise Mutation Model and the Demographic Model. A known inadequacy of the stepwise mutation model is that occasional mutations occur that change allele lengths by more than a single repeat unit. For the within locus test, such multi-step mutations have a conservative effect (5). For the interlocus *g* test, multi-step mutations can be accounted for explicitly by use of an analytical prediction (8), which shows that for reasonable frequencies of these mutations, any effect on *g* will be too slight to affect our primary conclusion that there is a clear difference in the signal of expansion between African and non-African populations (5).

An additional problem with the stepwise mutation model is that it assumes an infinite range of allowable allele sizes, whereas in reality the range is known to be constrained (3). The effect of range constraints is potentially non-conservative, but if

this were a cause for bias, genetic distances between the various humans populations, calculated using the assumptions of the model, would be systematically inconsistent with inferences from other sources. Goldstein *et al.* showed, however, that for the dinucleotide data, inferences about the date of splitting of African and non-African populations are consistent with other estimates (12), indicating that range constraints do not substantially retard genetic differentiation among human populations. For the tetranucleotide data, on the other hand, range constraints may have an influence on human population differentiation. This conjecture is supported by the lower measure of genetic population differentiation (F_{ST} value) observed in tetranucleotide relative to dinucleotide microsatellite data sets (10, 11).

Real human populations are not perfectly isolated from one another and are internally structured. This could produce a deceptive signal of expansion if appropriate populations are not selected for testing. Indeed, structuring appears to be a serious problem for the within locus test; the signal of expansion is consistently stronger in populations clumped into continental and whole world samples than in populations that are considered separately. The interlocus *g* test, in contrast, seems relatively insensitive to clumping schemes, indicating that slight deviations from the correct demographic model are unlikely to produce a false-positive signal of expansion.

Estimating a Date For the Expansion. The observed values of *g* and the average variance across loci put constraints on the possible dates of the detected expansion. In estimating the date, we define an expansion time, as well as its associated pre-expansion population size and factor of expansion, to be "allowable" if computer simulations using these three parameters generate 90% confidence intervals

that include the observed values of *g* and the average variance across loci. With N as the pre-expansion population size, we consider 60 values of Nμ between 0.05 and 25, 50 values of the expansion time from 0 to 10N generations in the past, and factors of sudden expansion ranging from 3 to 100. Applying this procedure to the data, we calculate allowed dates for 29 of the dinucleotide loci typed in the North Central African population, neglecting the anomalously high variance locus as before, and incorporating the estimated variation in the mutation rate, which in this case is $0.28\mu^2$. With an average dinucleotide mutation rate that has been estimated at 5.6×10^{-4} per generation (14), and a generation time of 25 years, we are able to make the following inferences.

The maximum pre-expansion population size for the North Central African population is 5,900, the lower bound for the post-expansion population size is 7,500, and the allowed dates are between 44,000 and 570,000 years ago—certainly predating the advent of agriculture (15). Crude estimates of the maximum likelihood surface for the date, based on computer simulations, indicate that the distribution is bimodal, and thus that a point estimate may not be very informative (5). The positions of the peaks for the various factors of expansion, however, constrain the maximum likelihood estimate to between 132,000 and 325,000 years, consistent with the expansion having occurred around or before the split of modern human populations in Africa (estimated to have occurred 75,000-287,000 years ago using the dinucleotide data, and dated to similar times using other data (12, 16)). Note that the method of allowed dates seems to be robust in the sense that it produces similar ranges of dates for widely varying expansion factors; however, it must be remembered that the real pattern of growth is likely to have been considerably more complicated, involving repeated periods of

expansion and possibly even contractions, and it is not clear how these complications would affect our inferences.

DISCUSSION

We have shown that a signal of population expansion in the Paleolithic appears in Africa but not elsewhere in the world. We observe the signal in two different data sets and in two separate statistical tests. Our strongest piece of evidence, a significant signal of expansion in the North-Central African population using the interlocus *g*-test, appears to be conservative to most deviations from the biological and demographic assumptions.

In light of the robustness of the detected signal, it may seem surprising that the within- and between-locus tests do not always agree on a significant result in the same population. For example, the signals of expansion in the North-Central African population, San, and Sotho-Tswana populations, are not replicated in both tests. This is not unexpected, however, and indeed is related to one of the strengths of the tests. As the tests are based on different principles, they are differently sensitive to deviations from the biological and demographic models that could mask the signal of expansion. In addition, as we discovered using a variety of parameter combinations in our computer simulations, the time of maximum sensitivity of the within locus test is about three times as recent as that of the interlocus test (5). Thus, by combining the two approaches, we can obtain more statistical resolution, for a broader range of demographic and historical parameters, than by using either test alone. By comparing

the results of the tests, we may even obtain new information about population history that could not be derived without such a comparison, a unique strength of combining the two approaches.

Whatever the reason for the lack of a signal of expansion outside Africa, our analysis, like many other genetic analyses, assigns a unique role to Africa in human evolution. If, as seems likely, the expansion we have detected in Africa predates the split of modern groups, it follows that non-Africans must have once carried the signal of expansion—a surprising fact because this signal of expansion is not now observed among these groups. Given the properties of the *g* statistic, however, the simplest way to erase the signal of growth would be a population bottleneck that could have occurred in the history of non-African populations. In a bottleneck that is sufficiently severe and long term, genetic drift causes the variances at individual loci to wander, raising the variance of the variance across loci and obscuring the signal of expansion (5). The hypothesis of a bottleneck also is appealing because it explains why no new signals of expansion developed among non-African populations—for example in the Americas, where colonization must have been associated with population growth. Assuming that the ancient bottleneck did not reduce the variances at all loci to zero, any new signals of expansion that could have arisen might have been obscured by residual variance of the variance; in order words, by elevated values of *g* inherited from before the time of the bottleneck.

If a bottleneck is indeed responsible for the high value of *g* outside of Africa, a compelling possibility is that it occurred during the emergence of the first anatomically modern human groups from that continent. Our analysis not only adds further support to the "out of Africa" theory (17) but indicates an approach for

characterizing the demographic nature of the emergence itself, putting constraints on the time frame and severity of the associated bottleneck (see *Appendix*). With only 30 loci, for example, we can conclude that the effective population size must at some point have dropped below 6,200; we may expect that analysis of larger data sets will reveal new details concerning this critical period of human history.

APPENDIX

Locus variances and genetic diversity tend to be significantly higher in Africa than elsewhere (15). By explaining this as the effects of variation lost during a severe bottleneck, the same bottleneck that we suggest has erased the signal of expansion outside of Africa, we can put a maximum on the effective population size during the bottleneck's narrowest point. The dynamic for the change in variance from generation to generation is $\Delta V = [(2N-1)\mu - V_0]/2N$, where V_0 is the current population variance (18). The requirement for variance to decrease is that ΔV should be negative, and thus that $N < V_0/\mu + ½$. Assuming that the pre-bottleneck variance is less than the current North Central African variance, a reasonable assumption since the variance of the North Central African loci should have been growing since the expansion, the effective population size during the bottleneck's narrowest point is constrained to less than 6,200.

1. Rogers, A. R. & Harpending, H. (1992) *Mol. Biol. Evol.* **9,** 552-569.
2. Anderson, S., Debruijn, M. H. L., Coulson, A.R., Eperon, I. C., Sanger, F. & Youg, I. G. (1982) *J. Mol. Biol.* **156**, 683-717.
3. Goldstein, D. B. & Pollock, D. D. (1997) *J. Heredity* **88,** 335-342.
4. Donnelly, P. (1996) *Ciba Foundation Symposia* **197**, 25-40.
5. Reich, D. E., Feldman, M. W. & Goldstein, D. B. *Mol. Biol. Evol.* **16,** 453-466.
6. Hudson, R. R. (1990) In *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) 1-44 (Oxford University Press, Oxford, 1991).
7. Slatkin, M. (1995) *Genetics* **139**, 457-462.
8. Zhivotovksy, L. A. & Feldman, M. W. (1995) *Proc. Nat. Acad. Sci. USA* **92**, 11549-11552.
9. Roe, A. (1992) Ph.D. Thesis, University of London.
10. Jorde, L. B., Bamshad, M. J., Watkins, W. S., Zenger, R., Fraley, A. E., Krakowiak, P. A., Carpenter, K. D., Soodyall, H., Jenkins, T. & Rogers, A. R. (1995) *Am. J. Hum. Genet.* **57,** 523-538.
11. Bowcock, A. M., Ruiz Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. (1994) *Nature* (London) **368,** 455-457.
12. Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6723-6727.
13. Takezaki, N. & Nei, M. (1996) *Genetics* **144**, 389-399.
14. Weber, J. & Wong, C. (1993) *Hum. Mol. Genet.* **2,** 1123-1128.
15. Cavalli-Sforza, L. L., Menozzi, P., Piazza, A. (1996) *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton).
16. Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. & Takahata, N. (1993) *Proc. Natl. Acad. Sci. USA* **92**, 532-536 (1993).
17. Stringer, C. B. and Andrews, P. (1988) *Science* **239,** 1263-1268.
18. Moran, P. A. P. (1975) *Theor. Popul. Biol.* **8,** 318-330 (1975).

FIGURE 1

Three examples of gene genealogies for constant-sized populations of 5,000 chromosomes from which 30 chromosomes are sampled randomly. Associated allele length distributions are shown below for each example and are obtained by distributing stepwise mutations along the genealogy with an average frequency of 0.00056 per generation. The numbers at the bottom of each genealogy show the change in allele length from the ancestral chromosome to each of the sampled chromosomes.

FIGURE 2

Three examples of gene genealogies for 30 chromosomes sampled randomly from a population that currently includes 50,000. All three genealogies correspond to constant-sized populations that underwent sudden, 100-fold expansions 7,000 generations in the past. Associated allele length distributions are shown for each figure and are obtained by distributing stepwise mutations along the genealogy with an average frequency of 0.00056 per generation. The numbers at the bottom of each genealogy show the change in allele length from the ancestral chromosome to each of the sampled chromosomes.

TABLE 1

Population	Average Variance	Number of Chromosomes	Within Locus (k) Test		Interlocus (g) Test*	
			Positive/Total	P value	g-ratio	P value
San	5.0	20-30	8/30	$P < 0.005$	1.07	$P < 0.49$
Sotho-Tswana	5.2	28-38	9/30	$P < 0.014$	1.00	$P < 0.50$
Tsonga	5.4	24-28	13/30	$P < 0.24$	0.87	$P < 0.32$
Biaka Pygmy	6.3	10	16/28	$P < 0.78$	1.11	$P < 0.47$
Nguni	5.9	26-28	13/30	$P < 0.24$	1.81	$P < 0.84$
Mbuti Pygmy	4.6	10	15/26	$P < 0.80$	1.89	$P < 0.80$
Malay	3.4	10-12	15/23	$P < 0.94$	1.29	$P < 0.64$
Japanese	4.1	30-38	13/30	$P < 0.24$	1.32	$P < 0.68$
Chinese	4.5	30-34	11/30	$P < 0.07$	1.67	$P < 0.82$
Cambodian	5.4	20-24	17/30	$P < 0.77$	2.17	$P < 0.88$
Vietnamese	5.0	12-18	12/30	$P < 0.14$	2.47	$P < 0.89$
North European	5.4	130-140	12/30	$P < 0.14$	1.82	$P < 0.84$
French	5.5	36-40	14/30	$P < 0.36$	1.93	$P < 0.84$

* The g-values are biased high because of variation in the mutation rate. This also causes the P values to be biased high, and hence we only list adjusted P values that are obtained after correcting for variation in the mutation rate.

Table 1: Analysis of human microsatellite diversity data using 30 tetranucleotide repeat microsatellites. P values for the within locus test are calculated from a binomial distribution, with probability of a positive k taken conservatively at 0.515. P values for the interlocus g test are calculated from a distribution that is empirically generated using 1000 computer simulations. For the inputs to the coalescent simulations, we use the actual number of loci and samples in our data and estimate $N\mu$ (the mutation rate times population size) as half the variance. Note that inaccuracy in this estimate of $N\mu$ is not a problem since g has the useful property of being independent of $N\mu$ (5).

TABLE 2

Population	Average Variance	Number of Chroms.	Within Locus (*k*) Test		Interlocus (*g*) Test*		
			Positive/Tot.	*P* value	g-ratio	*P* value	adj. *P* val.
N. Cent. African	7.7	32-38	15/30	$P < 0.51$	0.32	$P < .037$	$P < .0007$
Zairian	9.1	16-20	15/30	$P < 0.51$	1.80	—	$P < 0.23$
European	9.0	34-58	17/30	$P < 0.77$	3.88	—	$P < 0.11$
Amerind	8.1	36-60	13/28	$P < 0.24$	3.08	—	$P < 0.40$
East Asian	10.0	32-60	18/30	$P < 0.87$	2.92	—	$P < 0.20$
Sahulland	9.0	22-40	19/26	$P < 0.93$	3.33	—	$P < 0.10$
Melanesian	8.7	14-20	16/23	$P < 0.72$	2.53	—	$P < 0.43$

* The g-values for the interlocus test are biased high due to variation in the mutation rate. This in turn causes the *P* values to be biased high, so we only list *P* values that are significant despite the bias. We also provide a list of adjusted *P* values that emerge when we drop the exceptionally variable locus D13S122 and correct for variation in the mutation rate.

Table 2: Analysis of human diversity data from 30 dinucleotide repeat microsatellites. The populations are clustered according to the convention established in earlier literature: "North-Central African" includes Lisongo and Central African Republic Pygmies; "European" includes North Europeans and Italians; "Amerind" includes Karitania, Surui, and Mayan; "East Asian" includes Cambodian, Chinese, and Japanese; and "Sahulland" includes Australian and New Guinean (12).

FIGURE 1

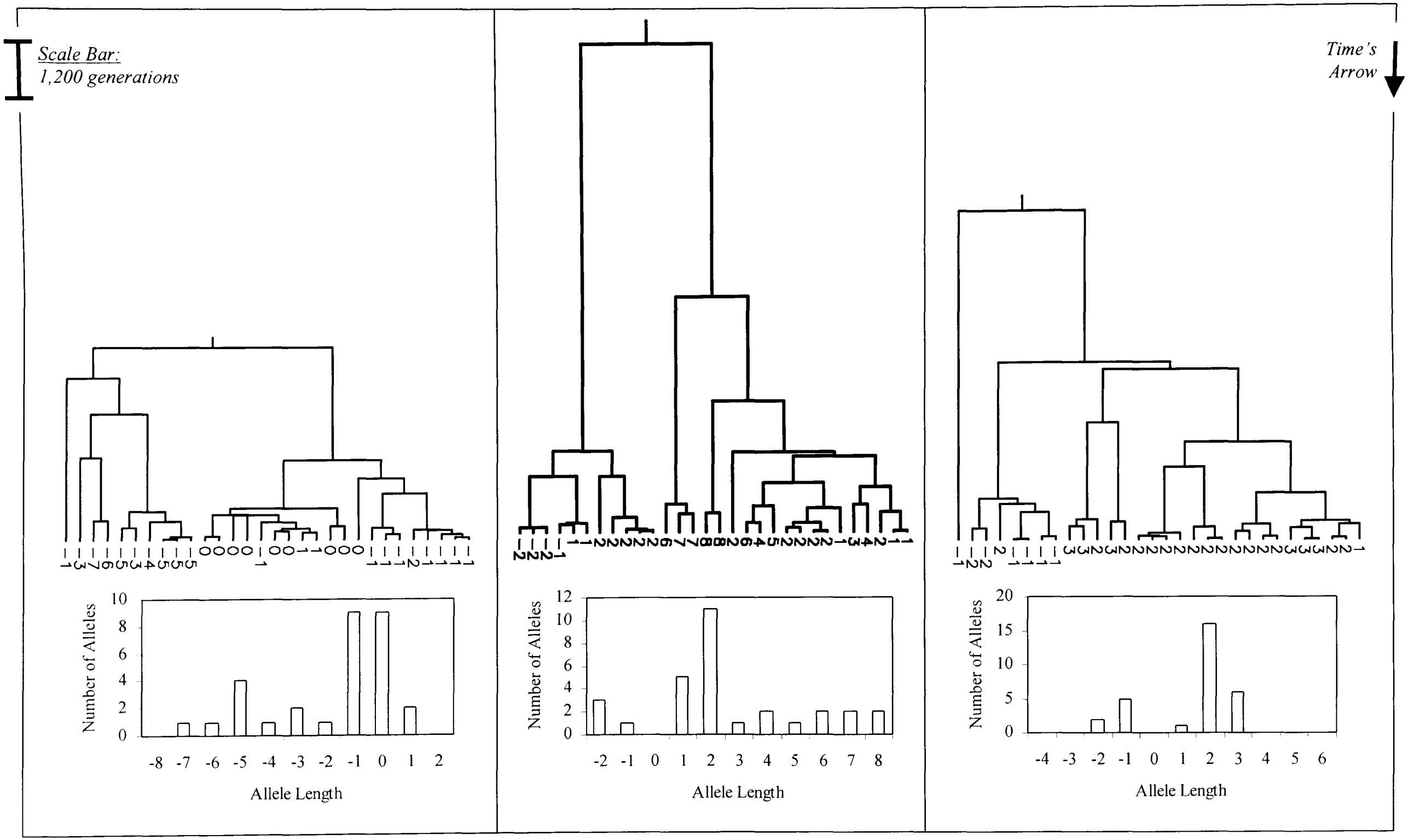
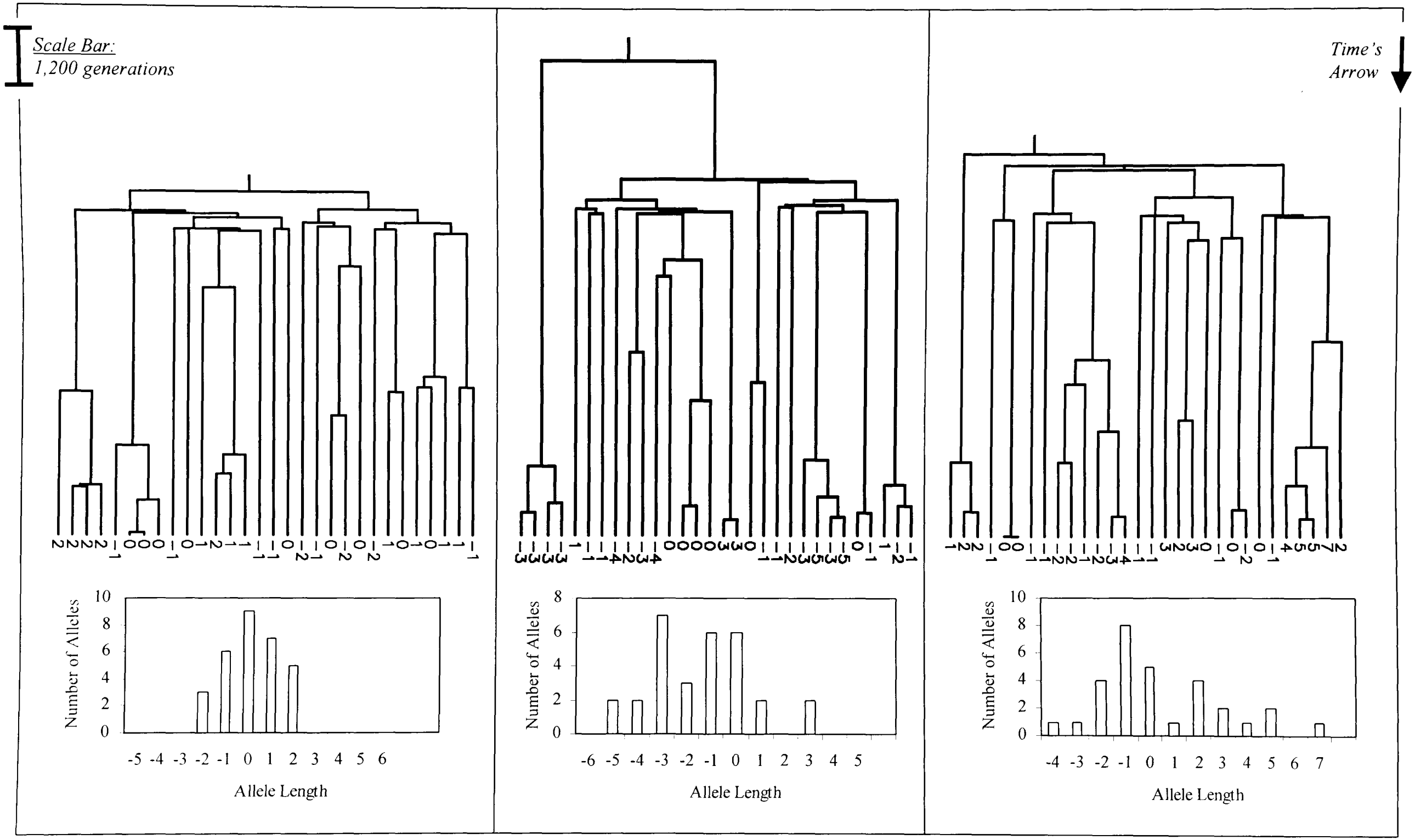

Scale Bar:
1,200 generations
Time's
Arrow
Number of Alleles
Allele Length
Number of Alleles
Allele Length
Number of Alleles
Allele Length

FIGURE 2


Scale Bar:
1,200 generations
Time's
Arrow
Number of Alleles
Allele Length
Number of Alleles
Allele Length
Number of Alleles
Allele Length

Addendum to "Genetic Evidence for a Paleolithic Human Population Expansion in Africa":

Since the writing of "Genetic evidence for a Paleolithic human population expansion in Africa," two papers have appeared that also make inferences about demographic history using variation at multiple unlinked microsatellite markers. These papers deserve special mention here.

Kimmel et al. (1998) used a ratio of two estimators of heterozygosity to study demographic history. Since their test does not assess the variability of gene genealogies across loci, it is analogous to the *k* test described in Chapter I, Part A: a within-locus test. When applied to a data set that includes 60 unlinked tetranucleotide microsatellites, typed in sub-Saharan Africans, Europeans, and Asians, the test suggests that a bottleneck rather than an expansion occurred in all these populations, with the weakest signal of a bottleneck in Africa. However, it is possible that the signal of a bottleneck observed in the data is not due to a real demographic effect, but is rather due to multistep mutations at some loci as well as variation in the mutation rate and mutation process across loci, possibilities that were not taken into account in the design of the tests. Kimmel et al.'s result of a stronger signal of expansion in African than in other populations, however, is consistent with the result obtained here (i.e., in Chapter 1, Part A of this thesis).

DiRienzo et al. (1998) test for expansion using variability of the variance across loci, which is essentially an interlocus approach analogous to the *g* test described here.

Twenty microsatellites were typed in Sardinians, Kaingang (an Amerindian population), and Luo (from Ethiopia). DiRienzo et al. do not correct their statistic for variation in the mutation rate across loci; however, they do show how to incorporate information about variation in the mutation process across loci, which they obtain by studying the patterns of mutations at the studied microsatellites in colorectal cancer tumors. DiRienzo et al. thereby find suggestions of expansion in the Kaingang and Luo, but the signal is weaker in Sardinians, suggesting a less dramatic history of expansion and possibly even a bottleneck in this population. Note that the strong suggestion of expansion in the Kaingang (Amerindians) is surprising in light of the analysis presented here, which suggests a history of a bottleneck common to all non-African populations. Further work is necessary in order to reconcile the analyses.

References:

A. DiRienzo, P. Donnelly, C. Toomajian, B. Sisk, A. Hill, M.L. Petzl-Erler, G.K. Haines and D.H. Barch (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic history. Genetics **148:**1269-1284.

M. Kimmel, R. Chakraborty, J.P. King, M. Bamshad, W.S. Watkin and L.B. Jorde (1998) Signatures of population expansion in microsatellite repeat data. Genetics **148:**1921-1930.

Statistical Properties of Two Tests That Use Multilocus Data Sets to Detect Population Expansions

David E. Reich*, Marcus W. Feldman†, and David B. Goldstein*

* *Department of Zoology, University of Oxford, United Kingdom*
(and)
† *Department of Biological Sciences, Stanford University*

Abstract

We describe two methods for detecting population expansions based on variation at unlinked microsatellite loci. The tests were first used in a study of human demographic history that showed evidence for a Paleolithic human population expansion in Africa. Here, we provide a simple recipe for applying the tests to other data sets, and describe the power of the tests as a function of the sample size, number of loci, mutation rate, diploid population size N_0, and time since expansion. An important property of the tests is that as long as the population doubles at least once every $0.1N_0$ generations, and the overall factor of expansion is sufficiently large, the signal of growth will be nearly identical to one generated by a sudden and massive expansion. This greatly simplifies the mathematical modeling necessary to evaluate the test results but also means that many patterns of growth will be indistinguishable using the tests. A second conclusion from our analysis is that the tests show different sensitivities to specific deviations from the biological and demographic models. Hence, more information can be garnered from the two tests taken together than from either alone.

Introduction

We describe two tests that use patterns of diversity at unlinked microsatellite loci to study demographic history (Reich and Goldstein 1998). Previous tests for population expansions, applied mainly to DNA sequence variation at the control region of mitochondria (Rogers and Harpending 1992), have yielded ambiguous results because they utilize information from only a single-locus. The problem with single-locus studies is that the effect of a selective sweep on a linked gene cannot be distinguished from the effect of a population expansion. Multilocus data, however, can distinguish between demographic effects and selection because population growth produces a signal at all loci, while selection only influences linked loci.

The basis of our approach is that gene genealogies—the trees of ancestral relationships among sampled alleles—reflect a population's history of growth. For a constant-sized population, a gene genealogical tree tends to be dominated by only a few ancient bifurcations, and mutations occurring in the oldest parts of the tree tend to divide the alleles into a few distinct clusters (Donnelly 1996). On the other hand, for an expanding population, allele types do not tend to be clustered (Donnelly 1996). By taking advantage of the systematic differences in gene genealogies for the two scenarios, and by assuming a simple stepwise mutation model for microsatellites whereby allele lengths change by a single unit and the direction of change is unbiased, we have developed within-locus and interlocus tests of demographic history (Reich and Goldstein 1998).

Two Tests for Population Expansion

Within-locus k test

The within-locus test (Reich and Goldstein 1998) is based on the assumption that microsatellite loci mutate according to a simple stepwise model, as well as on expectations about the systematic differences in the shapes of allele-length distributions for constant-sized and expanding populations. In the case of a constant-sized population, the typical allele-length distribution at a locus is expected to have several modes, corresponding to the small number of ancient bifurcations in the gene genealogy at that locus, while for an expanding population, the distribution is expected to have a single mode and to be more peaked (Reich and Goldstein 1998).

Peakedness is classically quantified using the kurtosis, γ_4/σ^4-3, where σ^4 is the variance squared, and γ_4 is the fourth central moment of a frequency distribution. The

kurtosis is expected to be positive if a distribution is more peaked than Gaussian and is expected to be negative otherwise. Reich and Goldstein (1998) defined a statistic, denoted *k*, which measures peakedness and is centered at 0 for the allele length distributions expected from a constant-sized population. The combination of statistical moments was selected by trial and error using coalescent-based simulations (APPENDIX 2), and the value of *k* tends to decrease with the greater peakedeness and the higher value of the kurtosis that occurs after an expansion.

Specifically, *k* is a linear combination of unbiased estimators—S^2 for the variance, Sig^4 for the variance squared, and Gam_4 for the fourth central moment (APPENDIX 1)—and an adjustment for the sample size, *n*. We used computer simulations (APPENDIX 2) to assign weights to these terms, and adjusted the weights empirically so that for a constant-sized population, *k* is positive about 50% of the time for as wide as possible a range of sample sizes and values of $N_0\nu$. (N_0 is the pre-expansion diploid effective population size, and ν is the mutation rate.)

$$k = 2.5*Sig^4 + 0.28*S^2 - 0.95/n - Gam_4 \quad (1)$$

The relationship between *k* and the kurtosis is evident in equation (1). Specifically, when the equation is divided through by $(-Sig^4)$, the first and last terms form an estimator for γ_4/σ^4-2.5, and thus the equation has a similar form to the kurtosis. The additional terms of equation (1), although not directly related to the kurtosis, are included to ensure that the probability of a positive *k* is nearly constant for as wide as possible a range of sample sizes and values of $N_0\nu$. As shown in Figure 1, the probability that *k* is positive is fully constrained between 0.515 and 0.55 for sample sizes of at least 10 and $N_0\nu > 0.5$. To

make sure that $N_0\nu$ is at least 0.5, we use the relation that for a population of constant-size, $E[Var] = 2N_0\nu$ (Zhivotovsky and Feldman 1995), implying that if the variance of an allele length distribution at a locus (*Var*) is greater than 1, it is likely to meet the conditions on $N_0\nu$. Finally, to implement the test, we limit ourselves to loci for which the sample size is at least 10 and the variances of allele-length distributions are at least 1 and count the proportion of loci that give positive *k* values. In order to assess significance levels, we use a binomial distribution with the number of trials equal to the number of loci, and the probability of a positive *k* set conservatively at its lower boundary of 0.515.

Interlocus g *test*

Variance in the widths of allele-length distributions across loci is usually lower in an expanding population than in a constant-sized population (Reich and Goldstein 1998). Thus, by measuring the variance of the allele-length distribution at each locus (V_j), and considering the variance of these variances across loci ($Var[V_j]$), we obtain a statistic that can be compared with its theoretical expectation to test for size constancy. An interlocus test on this basis can be applied to any type of genetic marker. However, what makes our test specific for microsatellites is the incorporation of the analytical expectation of $Var[V_j]$ for the specific case of the stepwise mutation model.

To implement this approach, we use the analytical result for $E[Var[V_j]]$ in the case when microsatellite mutations are single-step, and mutation rates are the same across loci and there is no selection (Roe 1992, pp. 137-203; Zhivotovsky and Feldman 1995):

$$E[Var[V_j]] = (4/3)(E[V_j])^2 + (1/6)E[V_j]. \qquad (2)$$

Here, $E[V_j]$ is the expected value of the variance at a given locus. Kimmel and Chakraborty (1996) have shown this equation to be true even for directionally-biased mutations.

To formulate the interlocus test explicitly, we consider the ratio $Var[Vj]/E[Var[V_j]]$; that is, the ratio of the observed to the predicted variance of the variance. Substituting the average variance across loci ($\overline{V}$) for the expected variance ($E[V_j]$) in equation (2), the test statistic we propose is as follows:

$$g = \frac{Var[V_j]}{\frac{4}{3}\overline{V}^2 + \frac{1}{6}\overline{V}}. \qquad (3)$$

An unusually low value of g may be interpreted as an indication of an expansion.

Significance levels for the interlocus test are obtained empirically, using a modified form of Hudson's (1991) coalescent computer simulation (see APPENDIX 2). A look-up table of 5^{th} percentile cutoffs for a range of numbers of loci and sample sizes—including the $g = 0.333$ cutoff for 30 loci and 40 samples used in many of the examples in this paper—is given in Table 1. Importantly, the behavior of g is nearly independent of the mutation rate and population size for all $N_0\nu > 0.5$ (Figure 2), so that the cutoffs provided in Table 1 are valid for many values of $N_0\nu$ even though they were derived specifically for $N_0\nu=0.63$. However, the interlocus test also applies when $N_0\nu < 0.5$, even though cutoffs are not presented for this case. Whatever the underlying value of $N_0\nu$, loci should never be selectively dropped from the analysis, even if they are monomorphic, because doing so would cause bias in the test results. An exception to this rule might be made in cases of *a*

priori evidence for a lower mutation rate, for example due to an interrupted stretch of repeats in the microsatellites.

An alternative multilocus test of demographic history has recently been introduced for use with microsatellite data (Kimmel et al. 1998). This is based on an "imbalance index," $\beta(t)$, which is a ratio of two estimators for $4N_0\nu$: one using the average variance across loci, and the other using the average heterozygosity. The imbalance index changes in response to population expansions and bottlenecks, because its numerator and denominator are differently affected by departures from size-constancy. However, the intuitive effects of an expansion on $\beta(t)$ are not obvious, as in the case of g. In addition, statistical concerns about $\beta(t)$ remain. For example, due to the statistical difficulties posed by ratios of random variables, Kimmel et al. (1998) construct $\beta(t)$ as the ratio of estimators averaged separately. However, it may be that variation in the mutation rate causes the ratio of the expectations to have a different value from the expectation of the ratios. Despite these potential problems, g and $\beta(t)$ may turn out to be complementary, with different sensitivities to certain aspects of demographic history.

Power of Tests as a Function of Number of Loci, Sample Size and Value of $N_0\nu$

In designing studies of demographic history, it is important to know how many loci, what sample size, and what value of $N_0\nu$, where N_0 now represents the *pre-*expansion size, are necessary for the tests to have a specific level of resolving power. To assess these questions, we consider 100-fold expansions that began N_0 generations ago, and that have an approximately 50% chance of producing unusually low test statistics for 30 loci, a sample size of 40, and an $N_0\nu$ value of 0.88. (This set of conditions is evaluated

for both the within-locus and interlocus tests.) By varying the number of loci, sample size and $N_0\nu$ in turn (with the other two parameters fixed at the above values), the power of the tests can be assessed as a function of the three main demographic parameters.

The power of a test is 1 minus the probability that the test fails to reject size constancy for a particular combination of parameters when an expansion actually occurred (that is, one minus the Type II error rate). To estimate this probability, it is first necessary to find the 5^{th} percentile lower cutoffs for the proportion of positive *k* values (within-locus test) and for the value of *g* (interlocus test). The cutoffs for the within-locus test are obtained from the binomial distribution, and the cutoffs for the interlocus test are taken from Table 1. By performing 1,000 computer simulations for each of several combinations of demographic parameters, and then counting the proportion of test statistics that are above the fifth percentile cutoffs when a 100-fold sudden expansion occurred N_0 generations ago, we derive the probability of not rejecting size constancy and use one minus this probability to calculate the power of the tests (Figure 3).

Figure 3A shows that the power of both tests increases with an increasing number of loci. Figure 3B shows that the power of the within-locus test increases with increasing sample size, while the power of the interlocus increases very little once the sample size is above a minimum. The reason for this is that the within-locus test is based on the fourth central moment—best assessed using a large sample size—while the interlocus test is based on the variance, which can be estimated with relatively fewer samples. Finally, Figure 3C shows that the power of both tests is nearly independent of $N_0\nu$ as long as $N_0\nu \geq 1.0$, as noted elsewhere. However, the test statistics will not be completely independent of $N_0\nu$ if microsatellites are subject to constraints on the lengths of alleles.

Due to space limitations, only a few combinations of demographic parameters are presented in Figure 3. Three general principles, however, are apparent: (1) increasing the number of loci increases the power of both tests, (2) increasing the sample size increases the power of the within-locus test only, and (3) the magnitude of $N_0\nu$ affects neither test when $N_0\nu > 1.0$.

Sensitivity of Within-Locus and Interlocus Tests to the History of Expansion

Statistical signals of expansion take a considerable amount of time to develop and fade gradually as the genetic variation in a population approaches a new state of mutation-drift equilibrium. The time window of sensitivity of the within-locus and interlocus tests—that is, the range of dates for which growth would be detected at a specified significance level—is now explored for a range of demographic models.

Sudden and gradual expansions

We begin by simulating 100-fold, sudden population growth from $N_0\nu = 0.88$ to $N_f\nu = 88$, where N_f is the post-expansion population size. For each date of expansion, we perform 1,000 simulations—each involving 30 unlinked microsatellite loci and a sample size of 40—and calculate central confidence intervals for the two tests. The dates of expansion to which the tests are maximally sensitive, that is, that produce the most marked reductions of the expected test statistics, are $5.1N_0$ generations ago for the within-locus, and $14.6N_0$ generations ago for the interlocus test (Figure 4). The within-locus test is more sensitive to recent expansions than the interlocus test, a fact that is also reflected in the dates of expansion that the tests can detect with greater than 50% probability:

$0.87N_0$ to $26N_0$ generations ago for the within-locus test, and $1.02N_0$ to $171N_0$ generations ago for the interlocus test (data not shown). Simulations for a range of $N_0\nu$ values other than 0.88 show that as long as $N_0\nu > 0.5$, the range of dates of expansion to which the tests are most sensitive, scaled in units of N_0, is independent of $N_0\nu$.

We now plot the proportion of loci with positive k, and the average value of g, against the time since expansion for 10-fold, 100-fold, and 1,000-fold expansions (Figure 5). The minimum points of the curves in Figure 5, corresponding to the dates of expansion to which the tests have the greatest sensitivity, occur at older times when the factor of expansion is larger, since it takes a longer time for the population to return to mutation-drift equilibrium. A second important property of the tests—which can also be seen in Figure 5—is that the test statistics are independent of the growth factor in the period immediately following expansion. To quantify this observation, we compare the average test statistics for a massive (10,000,000-fold) expansion, to those occurring for smaller factors of growth. Specifically, we assess how long after an expansion the average test statistics stay within 10% of the value expected for a 10,000,000-fold expansion: the answer is $0.7N_0$ generations in the case of a 10-fold expansion, $2.3N_0$ generations in the case of a 100-fold expansion, and $20N_0$ generations in the case of a 1,000-fold expansion. We conclude that if the factor of growth is large enough, and the time since expansion is less than a value determined by the factor of expansion, the tests have almost no power to distinguish among various factors of sudden growth. This is actually a manifestation of a more general property of the tests: whether growth is sudden or more gradual, the tests have almost no power to distinguish among alternative scenarios of expansion. The behavior is likely to be due mainly to the effects of growth on gene genealogies, and is therefore probably a general feature of genetic tests of demographic history.

We now consider the effect of gradual expansions (as opposed to sudden expansions) on the test statistics. Specifically, we consider the case of exponential growth. Figure 6 shows that when the population size doubles at least once every $0.1N_0$ generations, the expected values of the test statistics are always within 10% of the expectation for a massive (10,000,000-fold), sudden expansion (Figure 6). The same holds true for more complex, non-exponential histories of expansion, as long as doubling occurs at least once every $0.1N_0$ generations. The similar behavior of the test statistics for a range of growth rates—evident in Figure 6—greatly simplifies the study of demographic history using these tests. In particular, for many realistic growth scenarios, we can treat gradual expansions as if they had been sudden. For example, in the case of the Paleolithic human population expansion that was recently detected in Africa (Reich and Goldstein 1998), the doubling need only have occurred at least once every 10,000 years—assuming a pre-expansion population size of at least 4,000 and a generation time of 25 years—for the expansion to have been effectively sudden for the purposes of our tests. Since 10,000 years and 4,000 individuals seem to be a rather minimal requirement for a doubling time and pre-expansion population size in humans, the assumption of a sudden expansion that was made in our earlier analysis (Reich and Goldstein 1998) seems appropriate.

There are also certain cases in which it is not valid to treat expansions as sudden: for example, when the population doubles less frequently than once every $0.1N_0$ generations. In such circumstances, the earliest periods of growth dominate the test statistics, since they influence the times of occurrence of the first few bifurcations in the genealogical trees. These are the critical bifurcations for determining the shapes of allele-length distributions.

Population growth interrupted by a bottleneck

We now examine the response of the test statistics to strong growth interrupted by a bottleneck. For simplicity, we consider a model in which a sudden population expansion, by a factor of F, occurs t_0 generations ago. The population remains constant in size until $t_1+\Delta t$ generations ago, at which point it contracts to N_C individuals for Δt generations. Finally, the population returns to the pre-bottleneck size ($N_f = N_0F$) at time t_1, and remains static until the present. Using coalescent simulations of this model, we estimate average values for the test statistics (and also the average value of the variance across loci) as functions of several demographic parameters. Figure 7 graphs simulation results for the average test statistics and average variance when the parameters F, $N_0\nu$, t_0, and Δt are fixed at 100, 0.88, $3.0N_0$ generations, and $0.04N_0$ generations, respectively, and when the population size during the bottleneck, N_C, and the time when the bottleneck ended, t_1, are varied. As shown, the average test statistics tend to be raised due to a bottleneck; however, when the population contraction during the bottleneck is sufficiently severe, the average value can actually be depressed below what is expected, which explains why the curves are non-monotonic. Another important feature of Figure 7 (A-C) is that the effects of a bottleneck are more pronounced when the bottleneck is relatively recent (i.e. t_1 is small), than when the bottleneck is older.

To understand the behavior of Figure 7, note that a bottleneck increases genetic drift, thereby transforming the starlike genealogies expected for an expansion into the more clustered topologies expected for a constant-sized population. However, the opposite effect can also occur: when a bottleneck is sufficiently severe, genealogies tend to become more "star-like," and a powerful signal of expansion is generated because the typical gene genealogy is reduced to only a single ancestral lineage as old as the

bottleneck, and this "resets" the genetic "clock" in the sense that the average statistics are below what is expected even in the absence of a bottleneck (c.f. the right side of the curves in Figures 7A and 7B). Interestingly, the resetting of the genetic clock actually requires a more severe contraction in the case of the interlocus test than in the case of the within-locus test (c.f. the peak positions in Figures 7A and 7B). For the interlocus test, every one of the sampled loci must be reduced by drift to a single lineage before the genetic clock is reset, while for the within-locus test, the clock is reset locus by locus and the overall requirement for the development of a new signal of expansion is less stringent.

We now consider the effects of varying the parameters that were held fixed in Figure 7. Variation in t_0 has a complex effect on the test statistics; however, the factor of expansion to which the tests are maximally sensitive, as well as the overall shape of curves in Figures 7A and 7B, is independent of t_0. Variation in Δt and its effect on the interlocus test statistic is depicted in Figure 8. Figure 8 shows that when Δt is changed by a given factor, the curves in Figure 7B shift to the left by the same factor and are mostly unchanged in shape after the shift. (Similar effects are observed for the *k* statistic.) Hence, it is possible to extrapolate the quantitative results of Figures 7A and 7B to bottlenecks of durations other than $0.04N_0$ generations. Indeed, if the "severity" of the bottleneck is defined as a simple combination of N_C and Δt—$N_0\Delta t/N_C$—Figures 7A and 7B can be used to make the rough estimate that if $\Delta t(N_0/N_C) < 0.04$ (in units of N_0 generations) for the within-locus test, or $\Delta t(N_0/N_C) < 0.08$ (in units of N_0 generations) for the interlocus test, the test statistics will be essentially unaffected by a bottleneck. When N_C is time-dependent, a similar expression (involving an integral over the duration of the bottleneck), is the critical determinant of the effect on the test statistics (not shown).

Application to human demography

In connection with an expansion that was recently detected in Africa but not outside of Africa, Reich and Goldstein (1998) speculated that the statistical signal of population expansion may have been generated among the ancestors of all modern humans in the Paleolithic, but was then erased in non-African groups due to a population bottleneck that occurred approximately 80,000-100,000 years ago during the emergence of the first humans from Africa. In order to test this idea, we used Rogers and Harpending's (1992) estimation (based on analysis of variability in mitochondrial DNA) of an expansion from $N_0 = 3{,}254$ to $N_f = 547{,}586$ that occurred 4,800 generations ago (c.f. Kimmel et al. (1998)). With an estimated mutation rate for dinucleotide microsatellites of $\nu = 5.6 \times 10^{-4}$ (Weber and Wong 1993), this corresponds to an expansion from $N_0\nu = 1.8$ to $N_f\nu = 306$ occurring $1.48N_0$ generations ago. We then model a bottleneck starting $1.04N_0$ generations ago (85,000 years ago, assuming 25 years per generation), and lasting for $0.04N_0$ generations. Our simulations show that a bottleneck of this type could indeed have obscured an ancient signal of expansion, and would be expected to cause a rise in the average value of g above the 0.05 percentile lower cutoff for significance as long as $\Delta t(N_0/N_C)$ is in the approximate range $0.00025 < \Delta t(N_0/N_C) < 0.0045$ (in units of N_0 generations; see e.g., Figure 7B). This bottleneck is not expected to cause g to rise above the cutoff if the bottleneck occurred more than $1.2N_0$ generations ago. Note that erasure of a signal of population expansion outside of Africa is not necessarily the result of an out-of-Africa bottleneck; it could also be due to more recent bottlenecks, or to population structure (see below) in non-African populations.

Our time-sensitivity results also shed light on the date of the detected human population expansion. In particular, to verify our estimate that the expansion was

Paleolithic, we ask whether the expansion would have been likely to be detected had it occurred in the Mesolithic or Neolithic time periods. As described above, the most recent population expansion that can be detected with greater than 50% probability is $0.87N_0$ generations ago for the within-locus test, and $1.02N_0$ generations ago for the interlocus test. If we now assume that effective African population sizes were at least 50,000 during the Neolithic and Mesolithic, this translates to dates of expansion 1.1 million years ago and 1.3 million years ago respectively (assuming a generation time of 25 years). It therefore seems unlikely that the tests would be able to detect a Mesolithic or Neolithic expansion. We conclude that in the case of human demographic history, the two tests are not expected to be sensitive to any but the most ancient expansions that began from relatively small population sizes. Indeed, before applying these approaches to any data set (not just human data sets), it may be worth making a quick assessment—based on what is known about the history of the species and about the properties of the tests—to anticipate whether the expansions of interest are likely to be detected by the tests, or whether the pre-expansion population size and age of the species rule out detection.

Estimating a Date of Expansion and Other Parameters

To estimate a date for a detected expansion, we assume a model of sudden growth by a fixed factor F, and use computer simulations to find the range of expansion times and pre-expansion $N_0\nu$'s that are consistent with the observed variance and test statistic (Reich and Goldstein 1998). A particular combination of parameters (including a date of expansion) is considered to be "allowed" if the observed values of the variance and test statistics are within the specified confidence intervals determined from computer

simulation. To calculate the allowed range of dates for the expansion, we then take the full set of dates that, when combined with an appropriate value of $N_0\nu$, are consistent with the observed variance and test statistic in the sense that they comprise an allowed parameter combination. These allowed ranges of dates are specific to a particular factor of expansion, F. To find allowed ranges of dates for other factors of expansion and more complex models of growth, it is necessary to perform further analysis.

We now show how the same simulations can be used to construct a likelihood surface for the observed results as a function of the date of expansion and value of $N_0\nu$. By counting the proportion of results—for each combination of parameters—that are within a narrow window around the observed variance and test statistic, we obtain a surface that is generally bimodal, with peaks at positions that vary with the factor of expansion F. To understand this, we consider a graph showing the average value of the test statistic as a function of the time since expansion (Figure 5). By drawing a horizontal line through the graph at the height of the observed value of the within-locus or interlocus test statistic, we understand immediately why the likelihood surface has two peaks, and can roughly identify the positions of the peaks as a function of the time since expansion. Specifically, the two dates when the horizontal line and the curve intersect (that is, the dates for which the expected value of the test statistic is equal to the observed value), correspond roughly to the most likely dates of expansion and the peaks in the likelihood surface. The more recent of the two peaks corresponds to a population that is just beginning to register a signal of expansion, while the older peak corresponds to a population that is returning to equilibrium after an older expansion. Note that the points of intersection do not give the best estimates of the positions of the peaks. To obtain the best estimates, it is necessary to take into account the observed average variance across loci.

We now consider how date estimates change in response to different factors of growth. Figure 5 shows growth factors of 10, 100 and 1,000; in all three curves, the expected value of both test statistics shrinks in the period immediately following the expansion, then rises again to its equilibrium levels when the expansion is sufficiently old. Thus, a typical horizontal line through Figure 5 intersects the curves twice, with the date of the more recent intersection being the same regardless of the factor of expansion and the date of the older intersection varying with the factor of expansion. To take a specific example, consider a horizontal line through Figure 5B at the level $g = 0.4$, and note that it first intersects all three curves at a time lag of $1.2N_0$ generations, but intersects the other part of the curves at time lags of $20N_0$, $200N_0$ and $2{,}000N_0$ generations for F = 10, 100 and 1,000 respectively. The similar values of the curves shortly after expansion are due to the inability of the tests to distinguish between different factors of growth in the case of recent expansions. The different values of the curves at the right side of the figures are due to the fact that the rate of return to mutation-drift equilibrium depends on population size.

The independence of the more recent intersection—and hence the date of the more recent peak in the likelihood surface—from the assumed model of population growth is especially useful when it is possible to use other evidence, for example from the fossil record, to eliminate the possibility of the ancient peak. If we can not eliminate an ancient peak, however, our estimate for the date of expansion will be subject to error depending on our assumptions about growth history. There are also other complications that can affect date estimation. For example, if a population has undergone a series of small expansions over an extended period of time, or if a severe bottleneck has occurred, a date

estimate would be obtained that is not likely to correspond to a specific historical expansion.

Finally, it is possible to estimate other parameters: specifically, the maximum pre-expansion population size, N_0, and the minimum post-expansion population size, N_f. To obtain these values, we assume a variety of factors of expansion, and for each F, use the $N_0\nu$ value associated with the most recent and oldest "allowed dates" to find a maximum pre-expansion population size (N_0), and minimum post-expansion population size (N_f). In principle, the estimates obtained in this way could be different for each factor of expansion. However, we find that the maximum N_0 and minimum N_f (but *not* the minimum N_0 and maximum N_f), are generally extremely robust to changes in the assumed F if it is sufficiently large (Reich and Goldstein, unpublished data).

Complications in the Model

We have assumed single-step and unbiased mutations that are constant in rate across loci, and we have also assumed unstructured populations, an absence of selection, and independent sampling. In what follows, we explore the effects of departures from these assumptions on the test statistics (cf. Tables 2 and 3).

Deviations from the mutation model

Although the stepwise mutation model seems to describe changes in microsatellite allele lengths better than alternative models, it is not completely accurate (Goldstein and Pollock 1997). We begin our study of deviations by considering multi-step mutations. In

order to assess this effect, we assume 30 loci, a sample size of 40 and an $N_0\nu$ value of 0.88, and simulate genealogical trees for constant-sized populations, and with mutations distributed along the branches of the trees according to the method described in APPENDIX 2. Each mutation has an equal chance of increasing or decreasing the length of an allele, and the number of steps is determined by adding 1 to a random draw from a Poisson distribution with parameter λ. The average step size for an event drawn from this distribution is $\bar{s} = \lambda+1$. To determine the probability of obtaining a false-positive signal of expansion (Type I error rate), we then use the fifth percentile cutoffs for the case of a single-step mutation model. Table 2A shows that the probability of a false-positive is elevated above 5% in the case of the within-locus test (a non-conservative effect), and reduced in the case of the interlocus test (a conservative effect).

Second, we assess the effect of range constraints on the test statistics, and specifically a model of strict upper and lower boundaries on the allowed allele lengths (Feldman et al. 1997). The within-locus test statistic is expected to rise for this "reflecting boundaries" model due to the flattening of an allele-length distribution between the boundaries (a conservative effect). In contrast, the interlocus test statistic is expected to fall because the boundaries tend to restrict the variation of variances across loci (a nonconservative effect). For more realistic models of range constraints, involving multi-step mutations and length-dependent mutation rates, the overall effect on the interlocus test may be conservative. Indeed, it is difficult to predict the effect of realistic range constraints on the test statistics.

Finally, we consider asymmetry in the direction of mutations. Our simulations reveal that this has almost no effect on the test statistics, although the within-locus test is

slightly conservative to asymmetric mutations when the deviation is extreme (data not shown).

Variation in the mutation process across loci

We now consider sources of interlocus variation, and their effects on the interlocus test (for obvious reasons, we do not describe effects on the within-locus test). We begin by considering the effect of variation in the mutation rate across loci on the interlocus test. For this purpose, we assume the same set of parameters noted earlier, except that to determine $N_0\nu$ for each locus, we sample from a gamma distribution with a mean of $N_0\nu = 0.88$ and a variance that is between $0.05(N_0\nu)^2$ and $1.6(N_0\nu)^2$. Table 2B shows that the 5th percentile cutoffs for *g* rise dramatically as variation in the mutation rate increases across loci. A method for estimating variation in the mutation rate across loci—which can then be incorporated into adjusted fifth percentile cutoffs for the interlocus test—is described elsewhere (Reich and Goldstein 1998).

A second source of variation across loci is variability in the likelihood that a microsatellite will undergo multi-step mutations. In order to assess this deviation, we use simulations of constant-size populations with 30 loci, a sample size of 40, and $N_0\nu$ of 0.88, and consider multi-step mutations with an average step size of $\bar{s}$ ($\bar{s} = \lambda+1$, as described above). To determine λ at an individual locus, we then take a random draw from a gamma distribution with a mean of 0.2 and a variance ranging from 0 to 0.64 (the average step size across loci is then $E[\bar{s}] = E[\lambda]+1 = 1.2$). As shown in Table 2C, variability in step-size across loci causes a rise in the interlocus test statistic, which is a conservative effect.

Finally, we consider variation in the degree of mutational asymmetry across loci, which our simulations show to have very little effect on the test statistics (and hence we do not present quantitative results for it in the paper).

Deviations due to population structure

To assess the effect of population structure on the tests, we first consider an "island model with migration," in which there are j islands, each with a constant-sized population of N_0/j, and the probability of an individual migrating away from its current island is m per generation (note that migration is assumed to occur with equal probability to each of the other islands). To simulate this model, we use Hudson's (1991) coalescent algorithm, and consider a constant-sized population with 30 loci, a sample size of 40, $N_0\nu = 0.88$, and a migration parameter of $M = 4N_0m$. The geographic distribution of sampled alleles at each locus is assumed to be the same for each locus, and each individual contributes a full complement of alleles (two from each locus) to the data set. Our simulations using three populations (Table 2D) show that both statistics are elevated above the expectation for a panmictic population—a conservative effect that is more pronounced when migration rates are low. The effect occurs because population structure causes genealogical trees to be dominated by a few deep splits and to be more variable from locus to locus (Donnelly 1996). When the number of islands is different from 3, the test statistics are also higher than the expectation for a panmictic population (although the deviation from expectation is less pronounced for larger numbers of islands).

As an alternative model of population structure, we consider a multichotomy, in which a population is assumed to be panmictic until a fixed time in the past (t_{sep}), at which point it suddenly breaks up into j isolated subpopulations, each of size N_0/j. To

simulate a multichotomy, we use the same procedure as for the island model, but this time set the migration rate to be 0 and force all subpopulations to be lumped together into a single panmictic population at time t_{sep}. Table 2E shows the results of the simulations for $t_{sep} = N_0$ generations ago, for 30 loci, a sample size of 40, and $N_0\nu$ of 0.88, and for various numbers of subpopulations that together comprise a constant-sized population. We observe that the within-locus test is conservative (the statistic is elevated above the expectation for a panmictic population) unless the number of subpopulations is 15 or larger. In contrast, the interlocus test statistic is conservative for 2 subpopulations, approximately equal to expectation for 3 subpopulations, and suppressed below expectation (a non-conservative effect) for 4 or more subpopulations. The same qualitative effects are observed for both tests for larger values of t—although the effects are more pronounced when the value of t is larger.

Selection and Pseudoreplication

We now consider two additional deviations from the assumed model: selection and pseudo-replication. Selection is a major problem for single-locus studies because the effect of selection at a given locus can not be distinguished from a population expansion that would affect all loci. If many loci are available, however, it is possible to make such distinctions. In particular, for the within-locus test, most loci will reflect the population's history, and it is unlikely that more than a few will contribute to a misleading signal of expansion. In the case of the interlocus test, selection is likely to affect some loci and not others, increasing the variability across loci and thereby causing a rise in g. The conservativeness of the interlocus test in response to selection means that the test can be used to confirm a departure from size constancy that is detected by the within-locus test.

We also consider pseudoreplication, or double-counting of samples, which can result from experimental error or accidental sampling of multiple individuals from the same extended family. To assess the robustness of our two tests to pseudo-replication, we perform simulations of constant-sized populations with 30 loci, a sample size of 40, and $N_0\nu$ of 0.88, and resample the alleles to mimic pseudo-replication. For each allele that is produced in a run (beginning with allele $i = 1$) we assume that the allele is double-counted D_i times, with D_i obtained by adding 1 to a random draw from a Poisson process with parameter κ. To obtain the resampled set of 40 alleles, we then take D_1 alleles of type $i = 1$, D_2 alleles of type $i = 2$, etc., until we obtain the full resampled set of 40 (the average amount of double-counting in this set is about $\overline{D} \approx 1+\kappa$). Table 2F shows that an increase in the amount of pseudo-replication increases the expected values of the test statistics—a conservative effect that applies to both the within-locus and interlocus tests.

Date estimation in the presence of deviation from the assumed model

To describe how estimates for the dates of an expansion change when there are deviations from the assumed model, we return to the illustration of a horizontal line crossing the curves in Figure 5 at the level of the observed test statistic. The intersections of the line with the curves roughly determine the positions of the peaks of the likelihood surface, and if the horizontal line is shifted up (or down) due to a deviation from the assumed model, the maximum likelihood peaks will also shift. For example, if the deviation causes a systematic rise in the test statistics, the estimate for the date of the more recent likelihood peak will be too recent, the estimate for the date of the older likelihood peak will be too old, and the allowed dates of expansion will cover too wide a

range. Opposite effects are expected when the deviations cause a fall in the expected test statistic.

The other parameter that affects date estimation is the rate of rise of the variance after an expansion. A fast rise in the test statistics due to deviations from the assumed model will cause a systematic overestimate of the date of expansion, while a slow rise will result in an underestimate of the date. Table 3 summarizes the effects of each of the deviations discussed above on this parameter.

Conclusions

Our two tests of demographic history are generally conservative to deviations from the assumptions used to design the tests (Table 3). This means that the tests are not likely to generate a false-positive signal of expansion. On the other hand, the conservativeness also weakens the tests, and it is tempting to improve their power by estimating the quantitative extent of the deviations and incorporating the estimates directly into the tests. For example, Di Rienzo et al. (1998) studied colorectal cancer cells displaying microsatellite instability, and by observing the patterns of mutations in the tumors, inferred parameters for a generalized stepwise mutation model. While this approach is very interesting, it also seems problematic because such estimates depend on the assumption that the mutational process in germline cells is the same as in somatic cancer cells where mismatch repair enzymes may be defective. In light of the fact that the microsatellite mutation process is currently so poorly understood, we prefer to perform tests of demographic history based on a simple mutation model rather than on mutation models that are parameterized in more detailed (and possibly non-conservative) ways. In

this case, however, it is always necessary to evaluate the sensitivity of the methods to a variety of departures from these simple assumptions.

We now consider what types of data sets are appropriate for use with the tests. From Figure 3B, a sample size of at least 30 seems appropriate for the within-locus test, and a sample size of at least 15 seems appropriate for the interlocus test. Figure 3A shows that for both tests, at least 25 loci should be included, and of course the tests are more powerful when mutation rates are similar across loci. Finally, samples should be collected from as many isolated populations as possible, since it is possible to learn more from a comparison of populations than from individual populations. An example of the usefulness of a multi-population data set is provided by our study of human data (Reich and Goldstein 1998), in which a significant signal of expansion was detected in some African populations—but no populations outside of Africa—resulting in an inference that a dramatic demographic event must have occurred to separate Africans from non-Africans. This insight would not have emerged had our study focused on African or non-African groups exclusively. Indeed, a multi-population data set can sometimes give an indication of an expansion even when no significant signal is observed. For example, the signal of expansion that was recently detected in Africa was corroborated using a 30 tetranucleotide microsatellite data set for which the g values observed in Africa were consistently lower than g values observed elsewhere in the world—even though no single population gave a significant signal.

We have described the properties of two tests that use genetic data from multiple unlinked loci in order to assess demographic history. By considering the behavior of the tests in response to a number of demographic scenarios, we have shown that the tests are sensitive in different ways to various deviations from the assumed mutation and

demographic models, and that they can be used in conjunction to garner more information about demographic history than could be obtained from either test alone. In addition, the approaches to studying demographic history described in this paper, as well as the results concerning the behavior of the within-locus and interlocus tests in response to different growth models, are not restricted in principle to microsatellite variation. For example, it should be possible to use DNA sequence variation, as well as single nucleotide polymorphisms, to test hypotheses about demographic history. Work with this type of genetic data can be complementary to work with microsatellites because of the different mutation processes that are involved.

Acknowledgments

We would like to thank Peter Donnelly, Jonathan Pritchard and Daniel Reich for their helpful comments and discussions with us during various stages of this project.

APPENDIX 1

Derivation of Statistical Estimators

The within-locus test is based on the *k* statistic shown in equation (1). We construct the statistic empirically, using computer simulations and a linear combination of unbiased estimators for the variance (σ^2), the variance squared (σ^4) and the fourth central moment (γ_4). To estimate the variance (σ^2), we use the usual sample variance,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 , \qquad (6)$$

where the X_i's represent individual allele lengths and $\overline{X}$ represents the average allele length in the sample of n chromosomes. To estimate the variance squared and the fourth central moment, we use the following expressions:

$$Sig^4 = \frac{(n^2 - 3n + 3)}{n(n-1)(n-2)(n-3)}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2 - \frac{1}{(n-2)(n-3)}\sum_{i=1}^{n}(X_i - \overline{X})^4 \qquad (7)$$

$$Gam_4 = \frac{(n^2 - 2n + 3)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(X_i - \overline{X})^4 - \frac{(6n-9)}{n(n-1)(n-2)(n-3)}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2 \qquad (8)$$

In order to derive the estimators (7) and (8), we begin by finding expectations for $E(X_i^3\overline{X})$, $E(X_i^2\overline{X}^2)$ and $E(\overline{X}^4)$. Some simple manipulation produces the following, with $\mu \equiv E(X_i)$.

$$E(X_i^3\overline{X}) = \frac{1}{n}[E(X_i^4) + (n-1)\mu E(X_i^3)] \qquad (9)$$

$$E(X_i^2\overline{X}^2) = \frac{1}{n^2}[E(X_i^4)+2(n-1)\mu E(X_i^3)+(n-1)[E(X_i^2)]^2+(n^2-3n+2)E(X_i^2)] \qquad (10)$$

$$E(\overline{X}^4) = \qquad (11)$$

$$\frac{1}{n^3}[E(X_i^4)+4(n-1)\mu E(X_i^3)+3(n-1)[E(X_i^2)]^2+6(n^2-3n+2)E(X_i^2)+(n^3-6n^2+11n-6)\mu]$$

In order to develop unbiased estimators for σ^4 and γ_4, we now find expectations for the expressions $\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2$ and $\sum_{i=1}^{n}(X_i - \overline{X})^4$. In performing the algebra for these calculations, the identities (9), (10) and (11) are used:

$$\frac{n}{n-1}E\left[\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2\right] = \tag{12}$$

$$(n-1)^2E(X_i^4) - 4(n-1)\mu E(X_i^3) + (n^2-2n+3)[E(X_i^2)]^2 - 2(n^2-5n+6)\mu^2E(X_i^2) + (n^2-5n+6)\mu^4$$

$$\frac{n^2}{n-1}E\left[\sum_{i=1}^{n}(X_i - \overline{X})^4\right] = \tag{13}$$

$$(n^2-3n+3)E(X_i^4) \quad -4(n^2-3n+3)\mu E(X_i^3) \quad +3(2n-3)[E(X_i^2)]^2 \quad +6(n^2-5n+6)\mu^2E(X_i^2) \quad -3(n^2-5n+6)\mu^4$$

Finally, to simplify (7) and (8), we use the identities $\gamma_4 = E(X_i^4) - 4\mu E(X_i^3) + 6\mu E(X_i^2) - 3\mu^4$, and $\sigma^2 = (X_i^2) - \mu^2$, and obtain the results:

$$\frac{n}{n-1}E\left[\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2\right] = (n-1)\gamma_4 + (n^2-2n+3)\sigma^4 \tag{14}$$

$$\frac{n^2}{n-1}E\left[\sum_{i=1}^{n}(X_i - \overline{X})^4\right] = (n^3-3n+3)\gamma_4 + 3(2n-3)\sigma^4 \tag{15}$$

With this simple system of equations, we now find the desired expressions for σ^4 and γ_4:

$$\sigma^4 = \frac{(n^2 - 3n + 3)}{n(n-1)(n-2)(n-3)} E\left[\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2\right] - \frac{1}{(n-2)(n-3)} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^4\right] \quad (16)$$

$$\gamma_4 = \frac{(n^2 - 2n + 3)}{(n-1)(n-2)(n-3)} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^4\right] - \frac{(6n-9)}{n(n-1)(n-2)(n-3)} E\left[\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2\right] \quad (17)$$

Equations (16) and (17) are the basis of the statistical estimators Sig^4 and Gam_4 shown in equations (7) and (8). As desired, $E[Sig^4] = \sigma^4$ and $E[Gam_4] = \gamma_4$.

To verify empirically that these estimators are unbiased, we applied them to uniform and Gaussian distributions generated by computer. In addition, we tested the statistics against the distributions expected from a simple single-step mutation model and used simulations to show that in the case of a constant-sized population, the estimators are independent of sample size and agree with the analytical predictions for σ^4 and γ_4 (Roe 1992; Zhivotovsky and Feldman 1995).

APPENDIX 2

Computer Simulation

Our simulation is based on a coalescent algorithm by Hudson (1991), which was modified to reflect the stepwise mutation model as well as various deviations from the assumed mutation and demographic models. In a coalescent simulation, the genealogical tree is traced backward in time from the sampled individuals to their most recent common ancestor, and a demographic expansion or contraction has the effect of shortening or lengthening of the branches of the tree in proportion to the change in population size (Hudson 1991). Once the genealogical tree is generated, mutations are distributed along

the tree according to a Poisson process. To check the accuracy of our coalescent results, we used a conventional forward simulation—a simple Wright-Fisher model—in which members of a parent generation all have equal probability of producing progeny, and stepwise mutations have a fixed probability of occurring at every generation. Computer code for this simulation (in the C programming language), which takes into account a variety of deviations from the stepwise mutation model, is included on the Goldstein lab web page.

LITERATURE CITED

DI RIENZO, A., P. DONNELLY, C. TOOMAJIAN, B. SISK, A. HILL, M. L. PETZL-ERLER, G. K. HAINES AND D. H. BARCH. 1998. Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics **148:** 1269-1284.

DONNELLY, P. 1996. Interpreting genetic-variability – the effects of shared evolutionary history. Ciba Foundation Symposia **197:** 25-40.

GOLDSTEIN, D. B. AND D. D. POLLOCK. 1997. Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. J. Heredity **88:** 335-342.

HUDSON, R. R. 1991. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) 1-44 (Oxford University Press, Oxford, 1991).

KIMMEL, M. AND R. CHAKRABORTY. 1996. Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor. Popul. Biol. **50:** 345-367.

KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS AND L. B. JORDE. 1998. Signatures of population expansion in microsatellite repeat data. Genetics **148:** 1921-1930.

REICH, D. E. AND D. B. GOLDSTEIN. 1998. Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. U.S.A. **95:** 8119-8123.

ROE. A. 1992. Correlations and interactions in random walks and population genetics, pp. 137-203. University of London, London.

ROGERS, A. R. AND H. HARPENDING. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552-569.

WEBER, J. L. AND C. WONG. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123-1128.

ZHIVOTOVSKY, L. A. AND M. W. FELDMAN. 1995. Microsatellite variability and genetic distances. Proc. Natl. Acad. Sci. USA **92:** 11549-11552.

Table 1

5th Percentile Cutoffs for Interlocus Test as Function of Samp. Size and No. of Loci

	10 Samps.	20 Samps.	40 Samps.	80 Samps.	160 Samps.
5 Loci	0.12	0.10	0.08	0.08	0.08
6 Loci	0.13	0.12	0.11	0.10	0.10
7 Loci	0.16	0.15	0.14	0.13	0.12
8 Loci	0.19	0.16	0.15	0.15	0.14
10 Loci	0.23	0.19	0.17	0.18	0.17
12 Loci	0.26	0.21	0.20	0.19	0.19
14 Loci	0.29	0.24	0.22	0.22	0.22
17 Loci	0.31	0.26	0.25	0.24	0.24
21 Loci	0.35	0.29	0.28	0.27	0.27
25 Loci	0.38	0.33	0.31	0.30	0.29
30 Loci	0.39	0.36	0.33 [a]	0.33	0.32
37 Loci	0.47	0.39	0.37	0.35	0.35
44 Loci	0.48	0.43	0.40	0.37	0.37
53 Loci	0.53	0.46	0.42	0.41	0.41
64 Loci	0.54	0.47	0.45	0.43	0.44
77 Loci	0.56	0.51	0.47	0.46	0.45
92 Loci	0.60	0.54	0.50	0.49	0.48
110 Loci	0.63	0.56	0.54	0.51	0.51
133 Loci	0.67	0.59	0.56	0.53	0.53
159 Loci	0.70	0.61	0.57	0.57	0.56
191 Loci	0.73	0.65	0.61	0.58	0.58
230 Loci	0.76	0.67	0.64	0.62	0.61

NOTE—Each entry in the table is obtained on the basis of 2,000 simulations of a constant-sized population, with $N_0\nu = 2.5$. Since g is nearly independent of $N_0\nu$ for $N_0\nu > 0.5$, the cutoffs apply to a range of values of $N_0\nu$ and not just to 2.5. Interestingly, the cutoffs decrease with increasing sample size; the reason being that the variance at each locus is measured less reliably when the sample size is smaller, and hence the variability of the variances across loci is larger.

[a] The fifth-percentile cutoff of $g = 0.33$ is used in many of the examples in this paper.

Table 2

Quantitative Effects of Deviations from Assumed Model on the 5th Percentile Cutoffs For the Within-locus and Interlocus Tests

(A) Multi-Step Mutations	**Average Step Size $\bar{s}$:**	**1.0**	**1.05**	**1.1**	**1.15**	**1.2**	**1.35**	**1.5**	**2.0**
Within-locus test:	Ratio of observed cutoff to no-deviations cutoff:	1	0.92	0.92	0.92	0.83	0.75	0.75	0.75
	Probability of false rejection of size-constancy:	0.05	0.08	0.12	0.14	0.17	0.25	0.28	0.35
Interlocus test:	Ratio of observed cutoff to no-deviations cutoff:	1.0	1.03	1.07	1.09	1.12	1.15	1.15	1.17
	Probability of false rejection of size-constancy:	0.05	0.04	0.03	0.03	0.03	0.02	0.02	0.02

(B) Interlocus Variability in Mut. Rate	**Variance of $N_0\nu$ in Units of $(N_0\nu)^2$:**	**0.0**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.8**	**1.6**
Interlocus test:	Ratio of observed cutoff to no-deviations cutoff:	1.0	1.06	1.11	1.2	1.36	1.42	1.88	2.59
	Probability of false rejection of size-constancy:	0.05	0.04	0.03	0.01	.007	.005	.002	.000

(C) Interlocus Variability in Step Size	**Variance of $\bar{s}$ (using E[$\bar{s}$] = 1.2):**	**0.0**	**0.01**	**0.02**	**0.04**	**0.08**	**0.16**	**0.32**	**0.64**
Interlocus test:	Ratio of observed cutoff to Var[$\bar{s}$] = 0.0 cutoff:	1.0	1.03	1.05	1.12	1.20	1.29	1.37	1.34
	Probability of false rejection of size-constancy:	0.03	0.02	0.02	0.02	.006	.008	.006	.009

(D) 3-Island Model With Migration	**Migration Rate ($M = 4N_0m$) :**	**12.8**	**6.4**	**3.2**	**1.6**	**0.8**	**0.4**	**0.2**	**0.1**
Within-locus test:	Ratio of observed cutoff to no-deviations cutoff:	1.0	1.0	1.08	1.08	1.17	1.25	1.25	1.33

	Probability of false rejection of size-constancy:	0.05	0.04	0.03	0.03	0.01	.009	.006	.008
Interlocus test:	Ratio of observed cutoff to no-deviations cutoff:	1.0	1.02	1.03	1.08	1.18	1.35	1.61	1.84
	Probability of false rejection of size-constancy:	0.05	0.05	0.04	0.03	0.02	.007	.002	.001

(E) Multichotomy Model	**No. of Pops. (t_{sep} =1.0N_0 gens.):**	**2**	**3**	**4**	**5**	**7**	**10**	**15**	**25**
Within-locus test:	Ratio of observed cutoff to no-deviations cutoff:	1.08	1.17	1.17	1.17	1.08	1.08	1.0	0.92
	Probability of false rejection of size-constancy:	0.02	0.02	0.01	0.01	0.02	0.03	0.05	0.08
Interlocus test:	Ratio of observed cutoff to no-deviations cutoff:	0.96	0.92	0.88	0.86	0.78	0.71	0.65	0.61
	Probability of false rejection of size-constancy:	0.06	0.08	0.09	0.11	0.14	0.2	0.24	0.31

(F) Double-Counting of Samples	**Average Multiplicity of Samples:**	**1**	**1.2**	**1.4**	**1.7**	**2**	**4**	**8**	**16**
Within-locus test:	Ratio of observed cutoff to no-deviations cutoff:	1.0	1.0	1.08	1.08	1.08	1.25	1.42	1.42
	Probability of false rejection of size-constancy:	0.05	0.04	0.03	0.02	0.02	.003	.000	.000
Interlocus test:	Ratio of observed cutoff to no-deviations cutoff:	1.0	1.01	1.03	1.06	1.10	1.26	1.58	2.30
	Probability of false rejection of size-constancy:	0.05	0.05	0.04	0.04	0.03	.01	.001	.000

NOTE—To calculate each entry in Tables 2A-F, we assume a constant-sized population and perform 5,000 coalescent simulations for 30 loci, a sample size of 40, and an average N_0v of 0.88 (deviations are modeled as described in the text). For each combination of parameters, we provide the ratio of the fifth-percentile cutoff calculated for the deviation, divided by the cutoff expected in the case of no deviation, as well as the probability of false rejection of size constancy (probabilities less than 0.05 indicate a conservative test).

Table 3

Qualitative Effects of Deviations from the Assumed Model on the Test Statistics, with Specific Application to Date Estimation

Class of Deviation	Description of Deviation	Effect on the Within-Locus Test Statistic	Effect on Interlocus Test Statistic	Effect on Rate of Rise of the Variance
(A) Deviations from assumed mut. model	(1) Multi-step mutations	(1) Falls	(1) Rises	(1) Quickened
	(2) Range constraints	(2) Depends on model	(2) Depends on model	(2) Slowed
	(3) Directional asymmetry	(3) Slight rise	(3) None	(3) None
(B) Variability across loci	(4) Variable mutation rate	(4) None	(4) Rises	(4) None
	(5) Variable average step size	(5) None	(5) Rises	(5) None
	(6) Var. directional asymmetry	(6) None	(6) None	(6) None
(C) Population structure	(7) Island model with migration	(7) Rises	(7) Rises	(7) None
	(8) Multichotomy model	(8) Rises	(8) Depends on model	(8) None
(D) Miscellaneous	(9) Natural selection	(9) Falls	(9) Rises	(9) Slowed
	(10) Pseudo-replication	(10) Rises	(10) Rises	(10) None

NOTE—The test statistics and the rate of rise of the variance after an expansion both have effects on date estimation. "Rises" indicates a conservative effect on the test statistics, "Falls" indicates a non-conservative effect on the test statistics, and "Depends on model" indicates that the effect on the test statistics depends on the specific type of range constraints that are in place.

FIGURE LEGENDS

FIG. 1—The probability of obtaining a positive k as a function of sample size and $N_0\nu$ under an assumption of constant population size. For each data point (i.e., combination of sample size and $N_0\nu$), we perform 10,000 simulations and calculate the probability of a positive k based on the simulations. The probability fluctuates in a narrow range, between 0.515 and 0.55, for sample sizes of at least 10 and $N_0\nu > 0.5$.

FIG. 2—Ninety percent central confidence intervals for g, and the average value of g over 1,000 simulations, under the assumption of a constant population size. For each value of $N_0\nu$, we use 30 loci and a sample size of 40, and calculate the expectation and confidence interval for g. When $N_0\nu$ is greater than 0.25, g is nearly independent of $N_0\nu$. The near independence of g from $N_0\nu$ also holds for others numbers of loci and sample sizes.

FIG. 3—Confidence intervals for the within-locus and interlocus test statistics as functions of the time since expansion. We perform 10,000 simulations for each date of expansion, and consider 100-fold growth from $N_0\nu = 0.88$ to $N_f\nu = 88$ with 30 loci and a sample size of 40. The within-locus test statistic is expected to be lowest for an expansion that occurred $5.1N_0$ generations ago, and the probability of detecting an expansion using the test is expected to be greater than 50% for expansions that occurred $0.87N_0 - 26N_0$ generations ago (A). The date range is given by the intersection points of the 50^{th} percentile cutoff for a constant-sized population and the curve. In contrast, the interlocus test statistic (g) is expected to be lowest for an expansion that occurred $14.6N_0$

generations ago, and the probability of detecting an expansion is expected to be greater than 50% for expansions that occurred $1.02N_0 - 171N_0$ generations ago (B).

FIG. 4—(A) Power of the within-locus test and the interlocus test as functions of the number of loci, for a sample size of 40, and for population growth from $N_0\nu = 0.88$ to $N_f\nu = 88$. The power is defined as 1 minus the probability of not detecting an expansion at the 5% significance level if one indeed occurred. To obtain this probability, we assume a model of 100-fold growth that occurred $1.0N_0$ generations ago, perform 5,000 simulations for each combination of parameters, and calculate the percentage of results that are below the 5^{th} percentile cutoff for a constant-size population. Graph (B) shows the power of the two tests as functions of sample size, for 30 loci and population growth from $N_0\nu = 0.88$ to $N_f\nu = 88$. Each data point is calculated on the basis of 1,000 simulations. Graph (C) shows the power of the two tests as functions of $N_0\nu$, for 30 loci and a sample size of 40 and with each data point calculated on the basis of 5,000 simulations.

FIG. 5—(Sudden growth) Average values of the within-locus statistic (A), and the interlocus statistic (B), as functions of the time since expansion and the factor of sudden growth. For each date of expansion, we use 30 loci, a sample size of 40, and calculate test statistics based on an average of 250 simulations. The contours correspond to 10-fold, 100-fold and 1,000-fold growth from $N_0\nu = 0.88$. Immediately following the expansion (the left side of the figure), the behavior of both test statistics is nearly independent of the factor of expansion.

FIG. 6—(Exponential growth) Average values of the within-locus statistic (A), and the interlocus statistic (B), as functions of the time since expansion and the rate of

exponential growth. For each date, we use 30 loci, a sample size of 40, and calculate test statistics based on an average of 1,000 simulations. The contours correspond to doubling every $0.1N_0$, $0.03N_0$, $0.01N_0$ and $0.001N_0$ generations (the population size is held constant at N_0 before the expansion begins). For comparison, we show a contour corresponding to a massive, 10,000,000-fold sudden expansion. For all contours shown, the test statistic is always within 10% of what is predicted for a massive, sudden expansion.

FIG. 7—Average values of the within-locus statistic (A), the interlocus statistic (B), and variance of the allele-length distribution (C), for sudden growth interrupted by a bottleneck that lasts $0.04N_0$ generations. Inset shows the time course for this scenario of growth interrupted by a bottleneck. For each combination of parameters, we use 30 loci, a sample size of 40, an $N_0\mu$ value of 0.88, and perform 1,000 simulations of a 100-fold sudden expansion that began $3.0N_0$ generations ago and that was interrupted by a bottleneck ending 0, $0.5N_0$, $1.0N_0$ or $2.0N_0$ generations ago. The average value of the test statistic in the absence of a bottleneck can be extrapolated from the far left side of the curves in (A) and (B). The factor of contraction to which the tests are most sensitive—that is, the position of the peaks of the curves in (A) and (B)—is substantially higher for the within-locus than for the interlocus test. Finally, for severe bottlenecks that ended 0 generations ago, the test statistics behave irregularly because the allele-length distributions have extremely low variance (hence, the contours corresponding to these cases are truncated).

FIG. 8—The effect of the duration of the bottleneck (Δt) on the interlocus test. Using 30 loci, a sample size of 40, and an $N_0\nu$ value of 0.88, and considering a 100-fold

expansion that began $3.0N_0$ generations ago and was interrupted by a bottleneck $1.0N_0$ generations ago, we perform 1,000 simulations for each data point. As shown in the figure, an increase in Δt by a certain factor causes the contours to shift to the left by the same factor. Similar effects occur for the within-locus test.

FIGURE 1

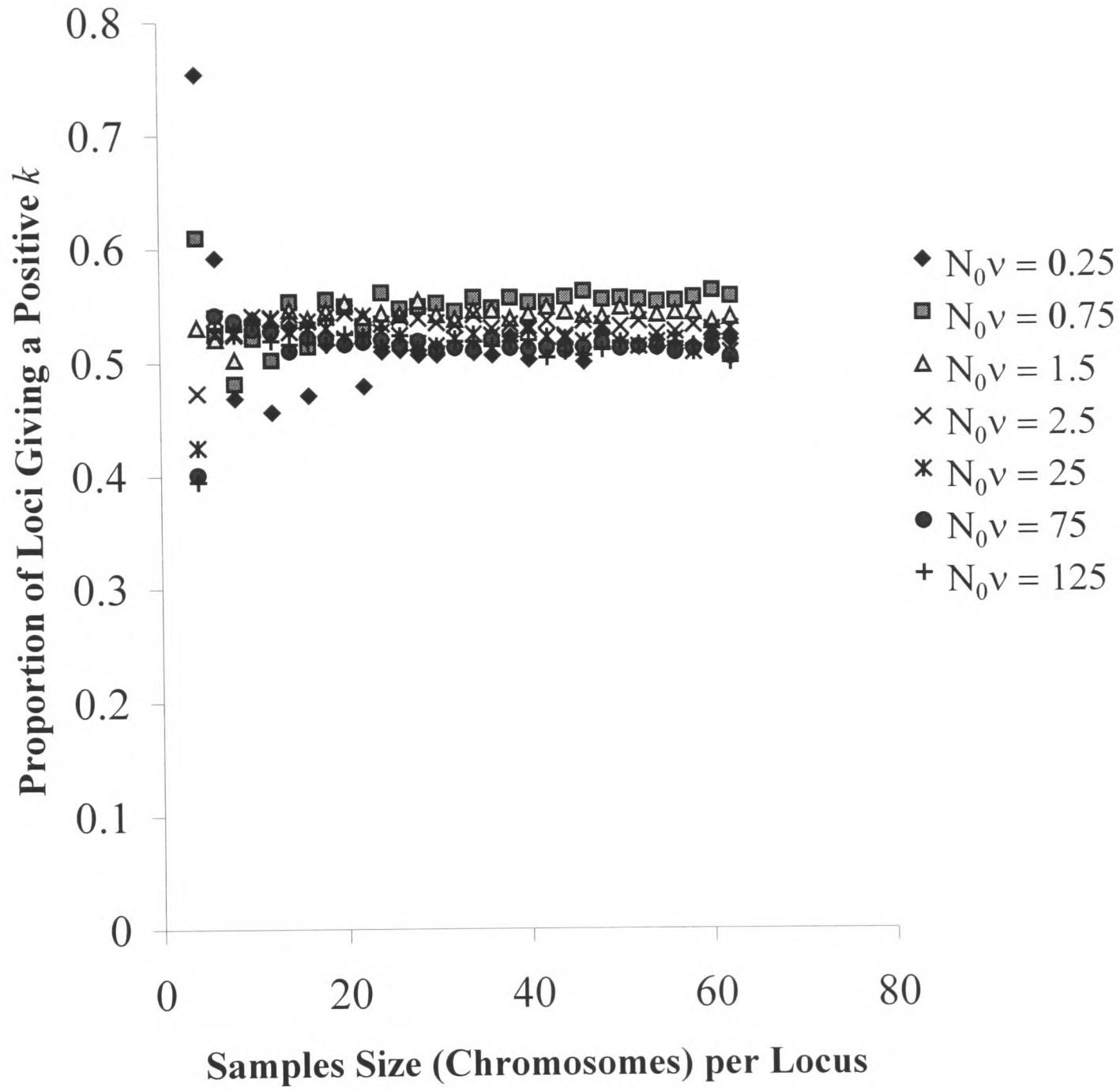

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion of Loci Giving a Positive k
0
20
40
60
80
Samples Size (Chromosomes) per Locus
N0ν = 0.25
N0ν = 0.75
N0ν = 1.5
N0ν = 2.5
N0ν = 25
N0ν = 75
N0ν = 125

FIGURE 2

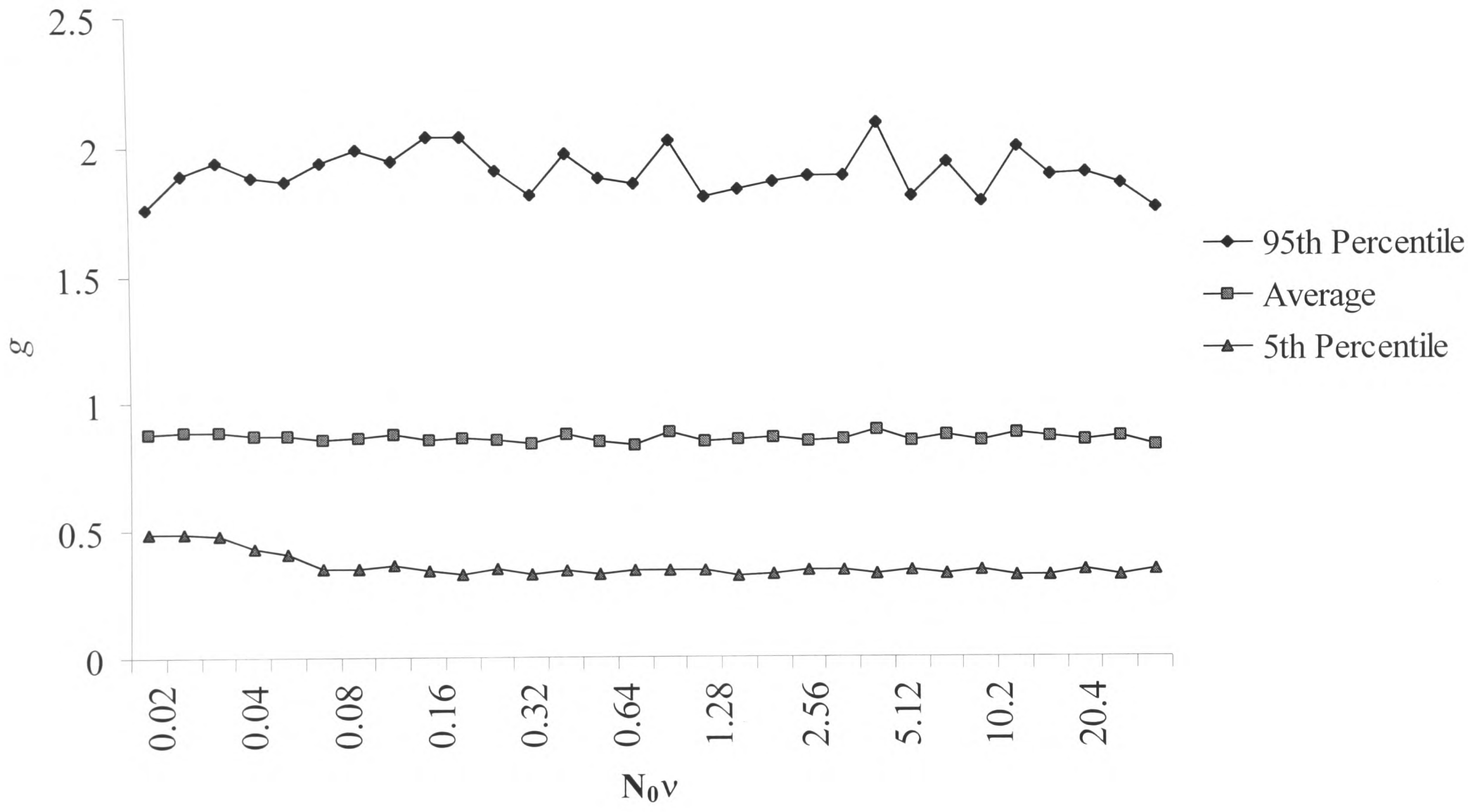

2.5
2
1.5
1
0.5
0
g
0.02
0.04
0.08
0.16
0.32
0.64
1.28
2.56
5.12
10.2
20.4
N0ν
95th Percentile
Average
5th Percentile

FIGURE 3

(1)

A

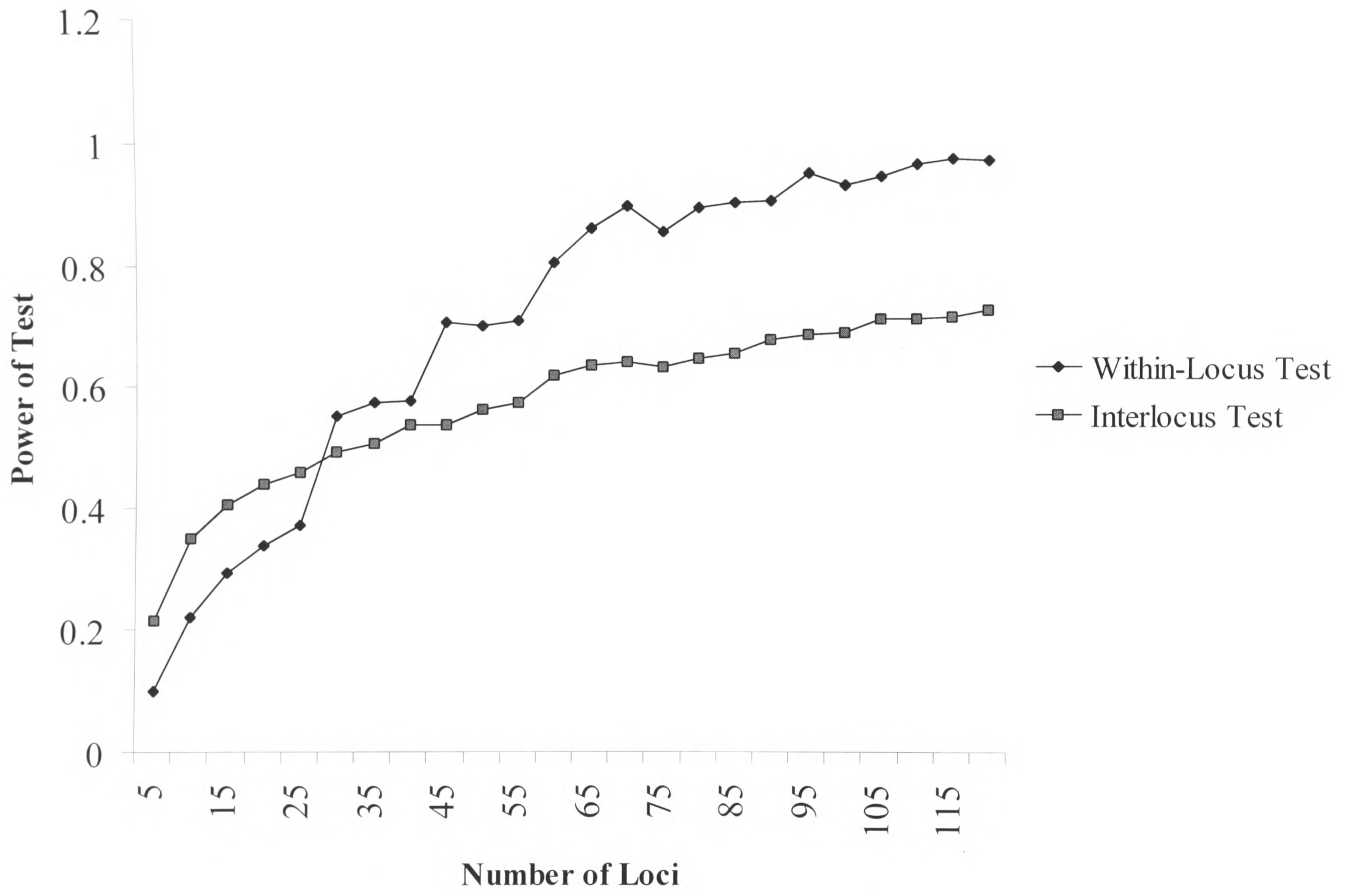

1.2
1
0.8
0.6
0.4
0.2
0
Power of Test
5
15
25
35
45
55
65
75
85
95
105
115
Number of Loci
Within-Locus Test
Interlocus Test

FIGURE 3

(2)

B

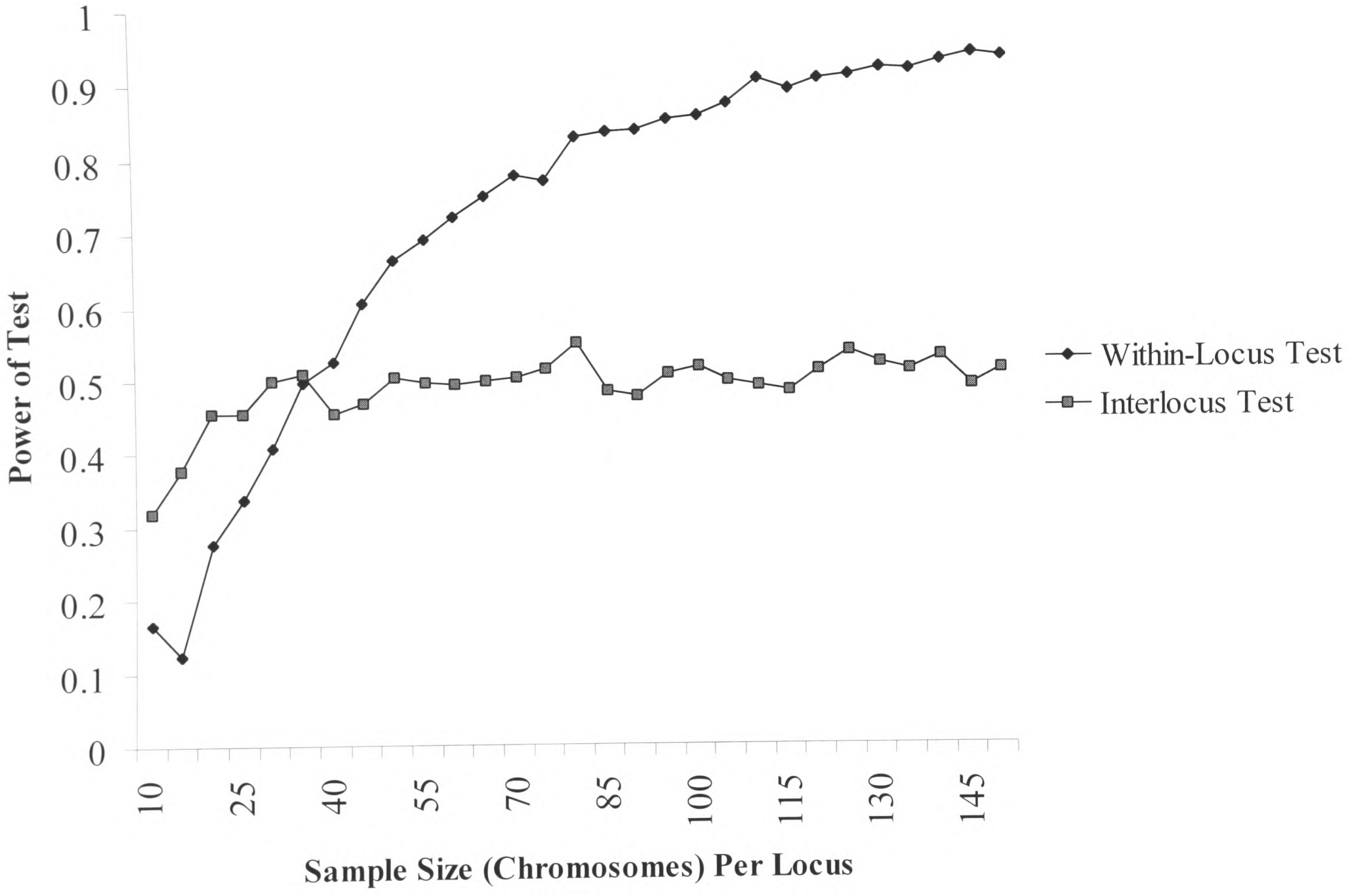

Power of Test
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
25
40
55
70
85
100
115
130
145
Sample Size (Chromosomes) Per Locus
Within-Locus Test
Interlocus Test

FIGURE 3

(3)

C

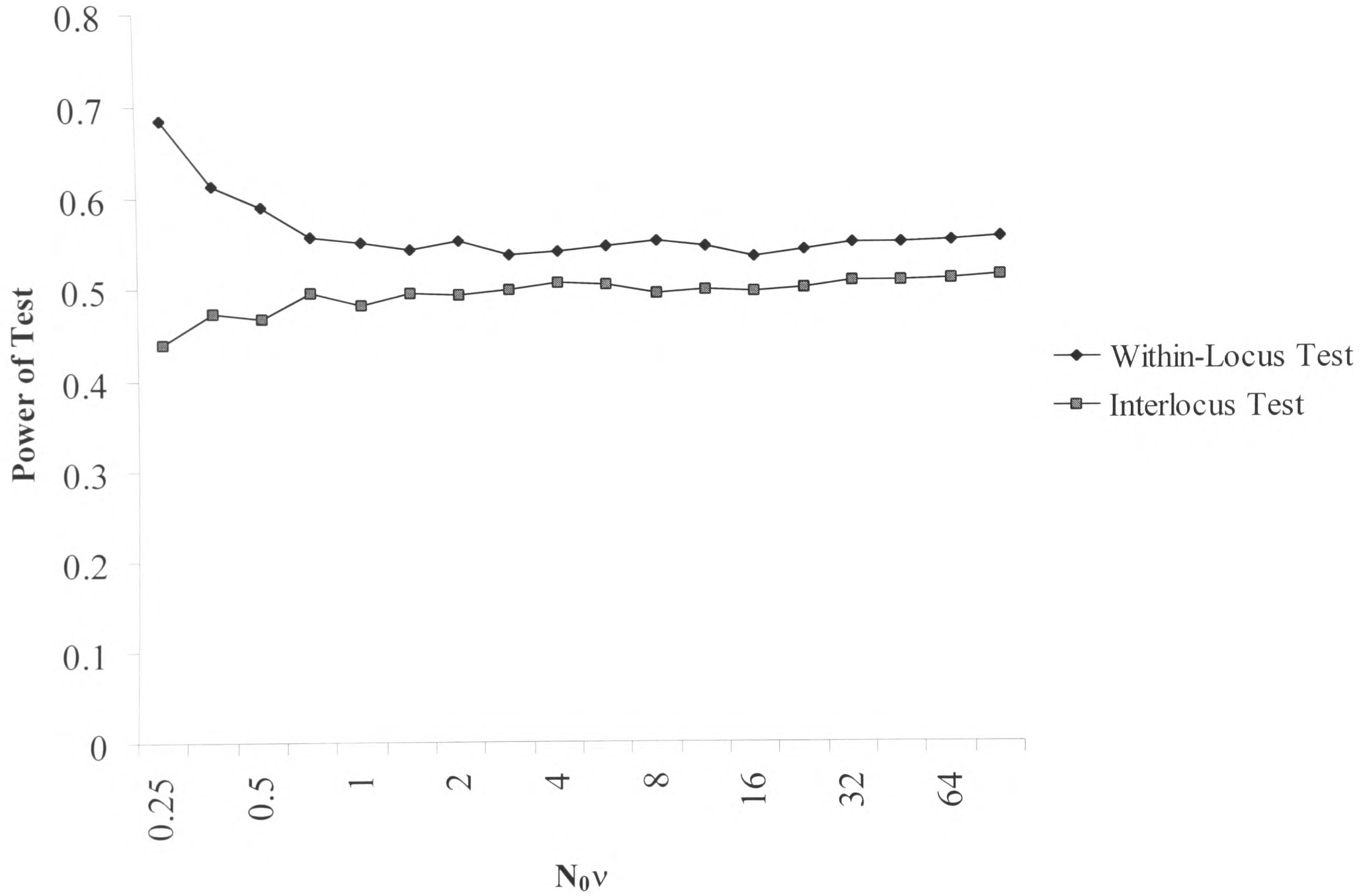
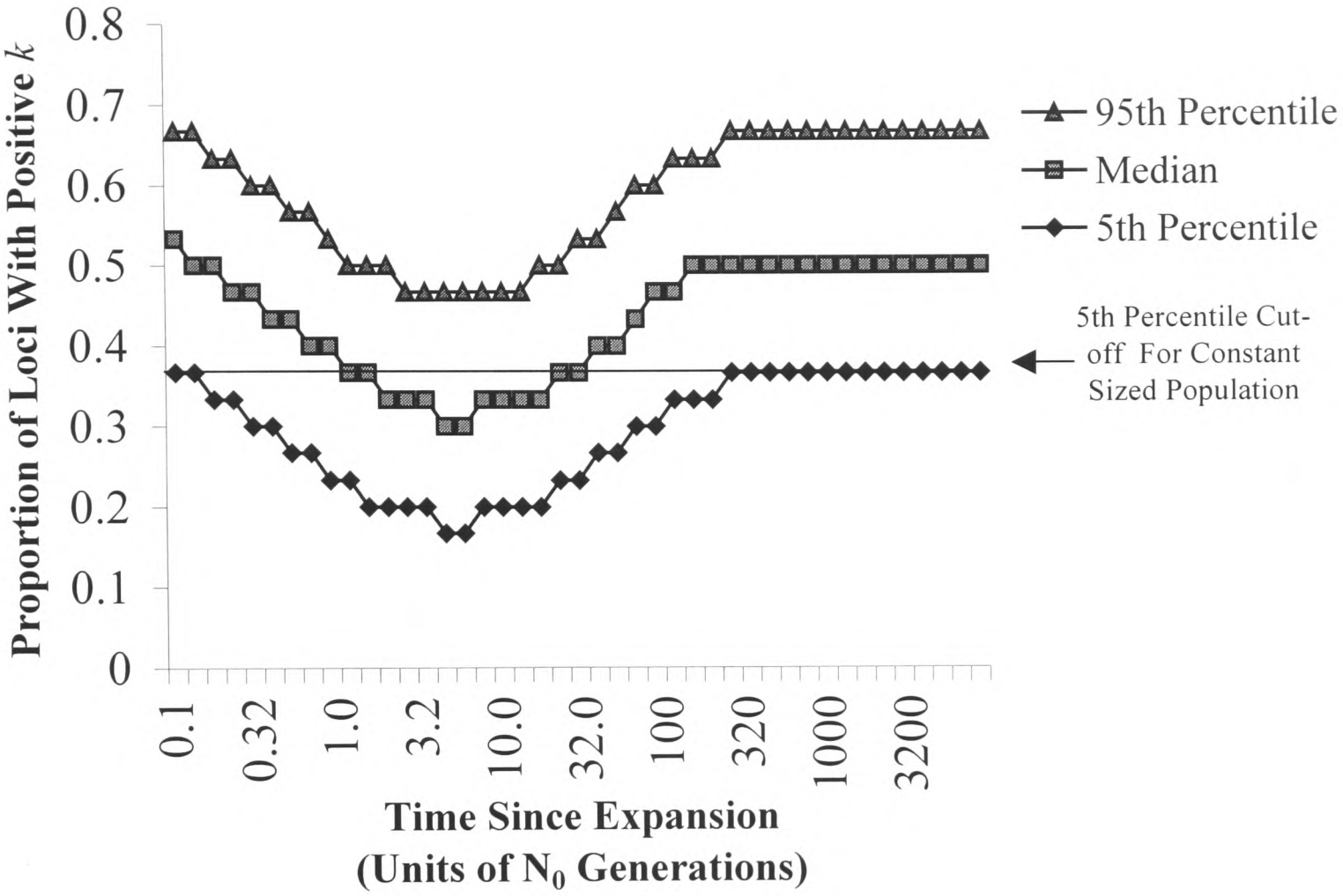

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Power of Test
0.25
0.5
1
2
4
8
16
32
64
$N_0\nu$
Within-Locus Test
Interlocus Test

FIGURE 4

A


Proportion of Loci With Positive k
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
95th Percentile
Median
5th Percentile
5th Percentile Cut-off For Constant Sized Population
0.1
0.32
1.0
3.2
10.0
32.0
100
320
1000
3200
Time Since Expansion
(Units of N0 Generations)

B

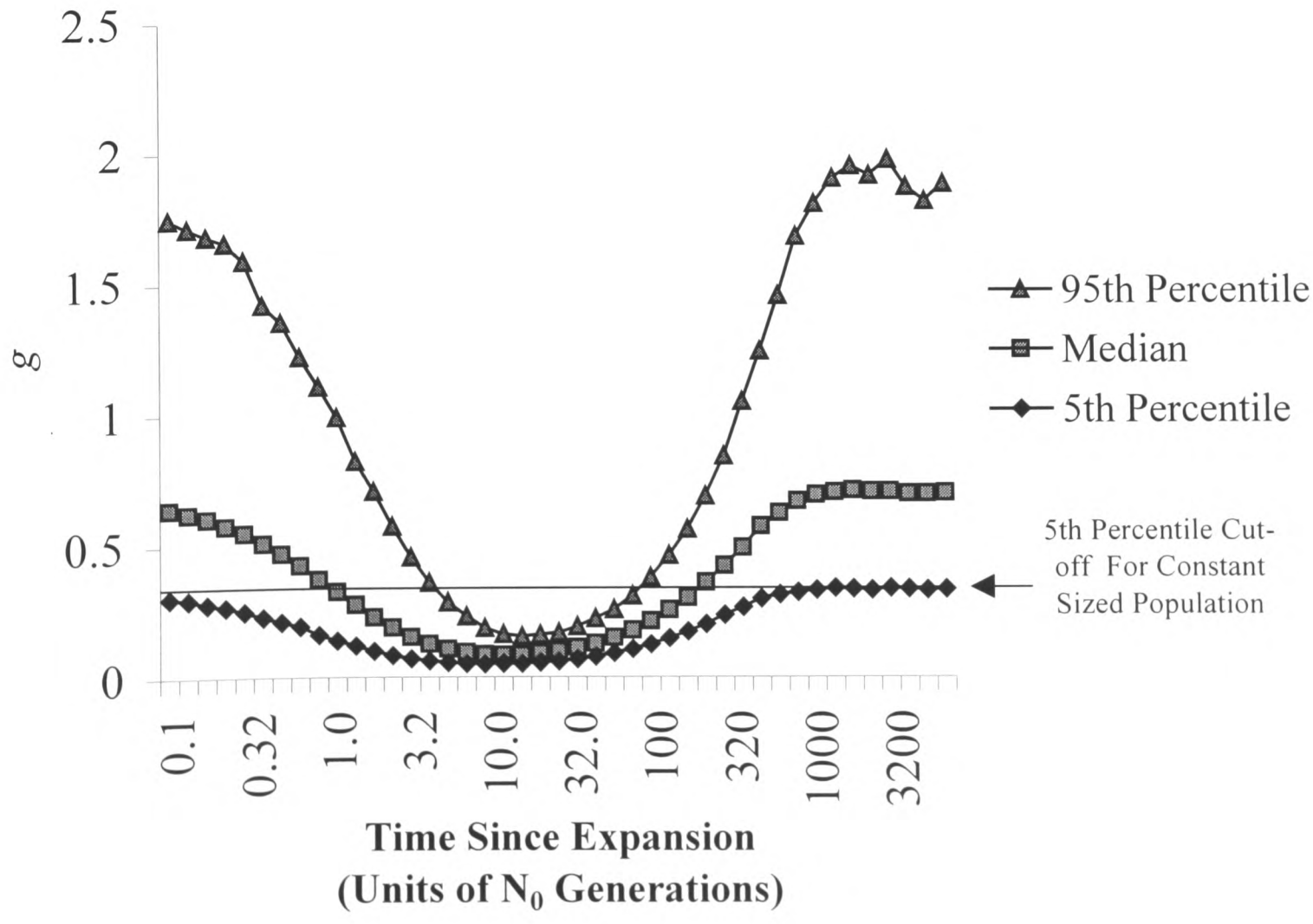

g
2.5
2
1.5
1
0.5
0
95th Percentile
Median
5th Percentile
5th Percentile Cut-off For Constant Sized Population
0.1
0.32
1.0
3.2
10.0
32.0
100
320
1000
3200
Time Since Expansion
(Units of N0 Generations)

FIGURE 5

(1)

A

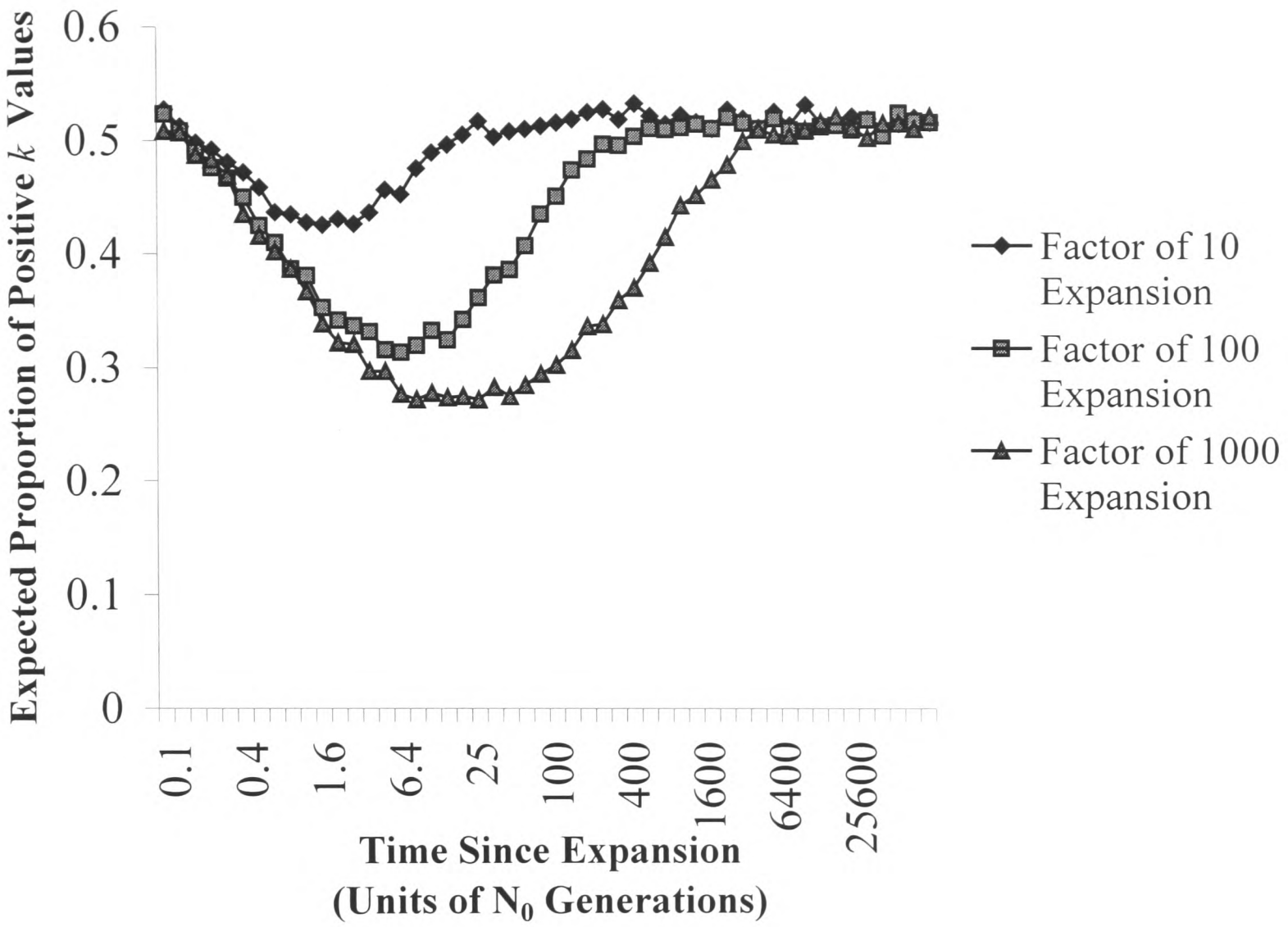

Expected Proportion of Positive k Values
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.4
1.6
6.4
25
100
400
1600
6400
25600
Time Since Expansion
(Units of N0 Generations)
Factor of 10 Expansion
Factor of 100 Expansion
Factor of 1000 Expansion

FIGURE 5

(2)

B

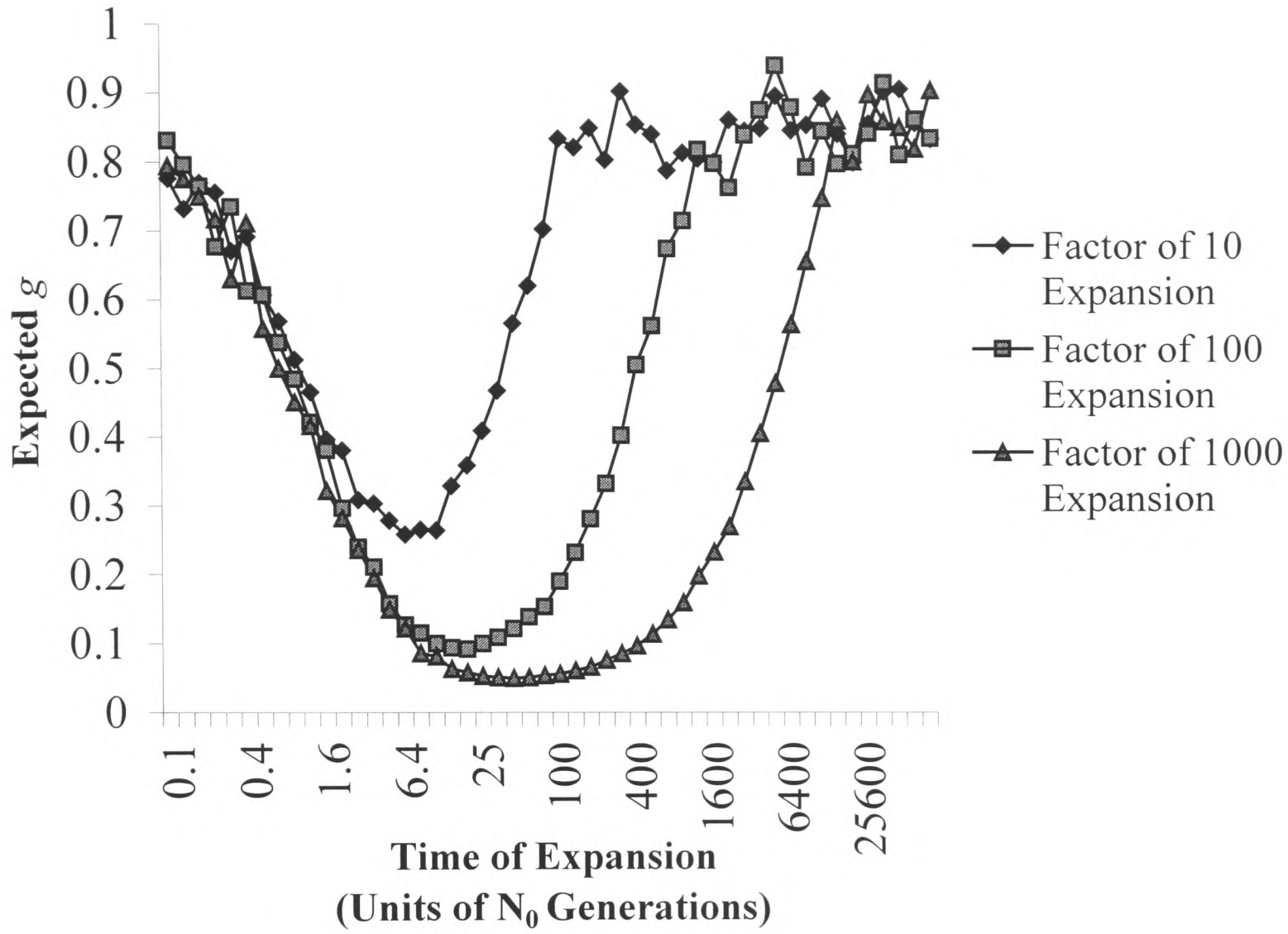
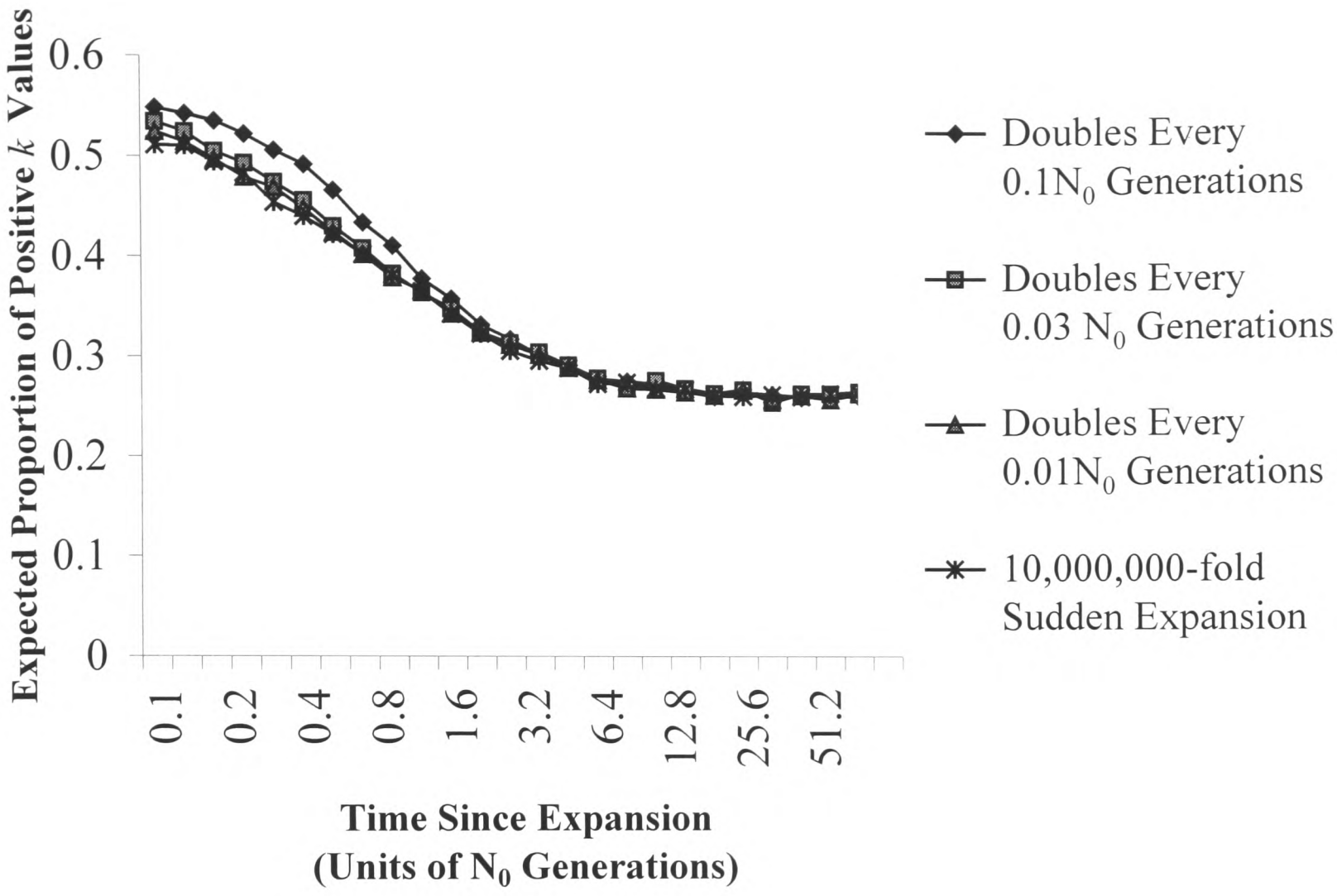

Expected g
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.4
1.6
6.4
25
100
400
1600
6400
25600
Time of Expansion
(Units of N0 Generations)
Factor of 10 Expansion
Factor of 100 Expansion
Factor of 1000 Expansion

FIGURE 6

A


Expected Proportion of Positive k Values
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.4
0.8
1.6
3.2
6.4
12.8
25.6
51.2
Time Since Expansion
(Units of N0 Generations)
Doubles Every 0.1N0 Generations
Doubles Every 0.03 N0 Generations
Doubles Every 0.01N0 Generations
10,000,000-fold Sudden Expansion

B

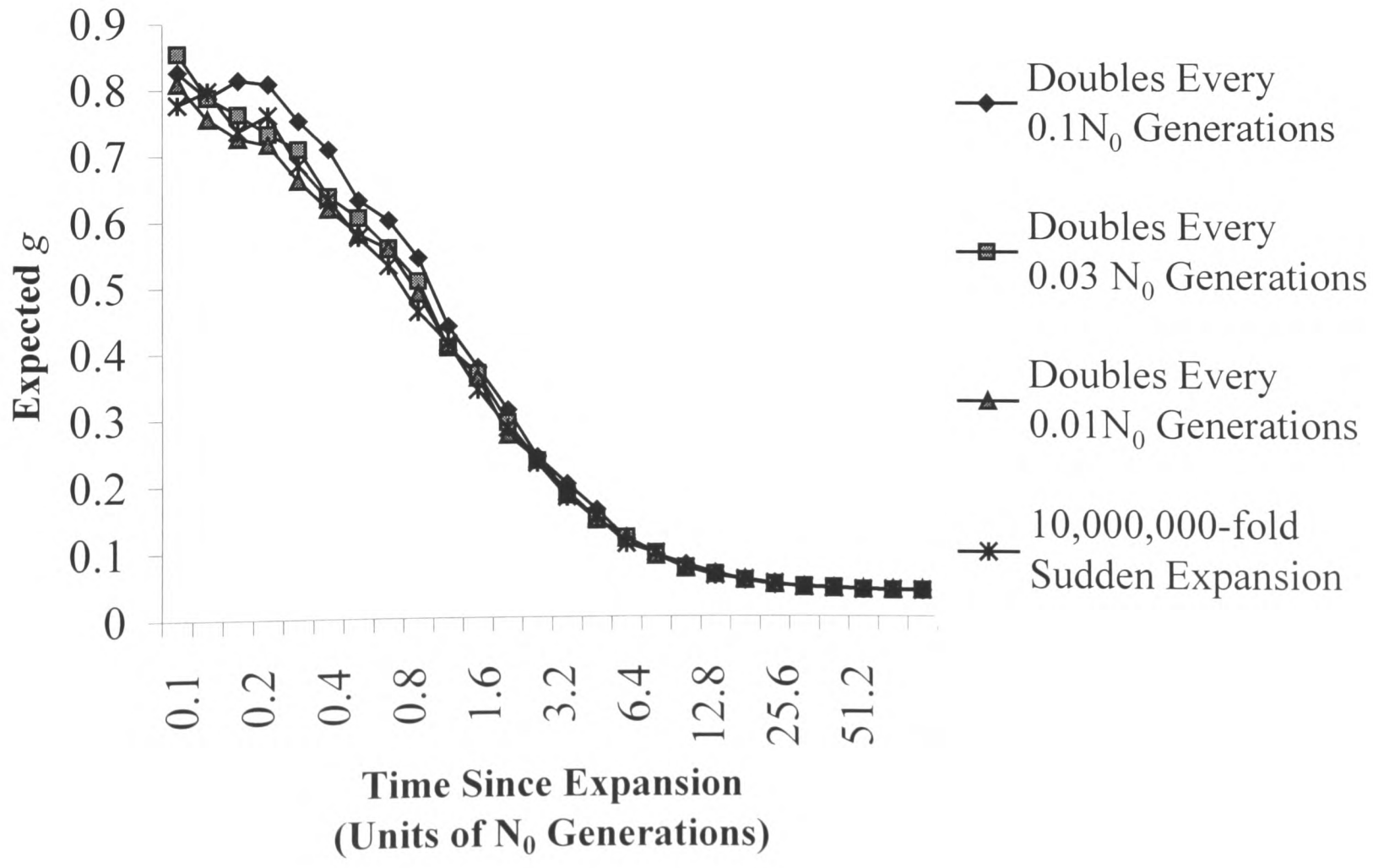
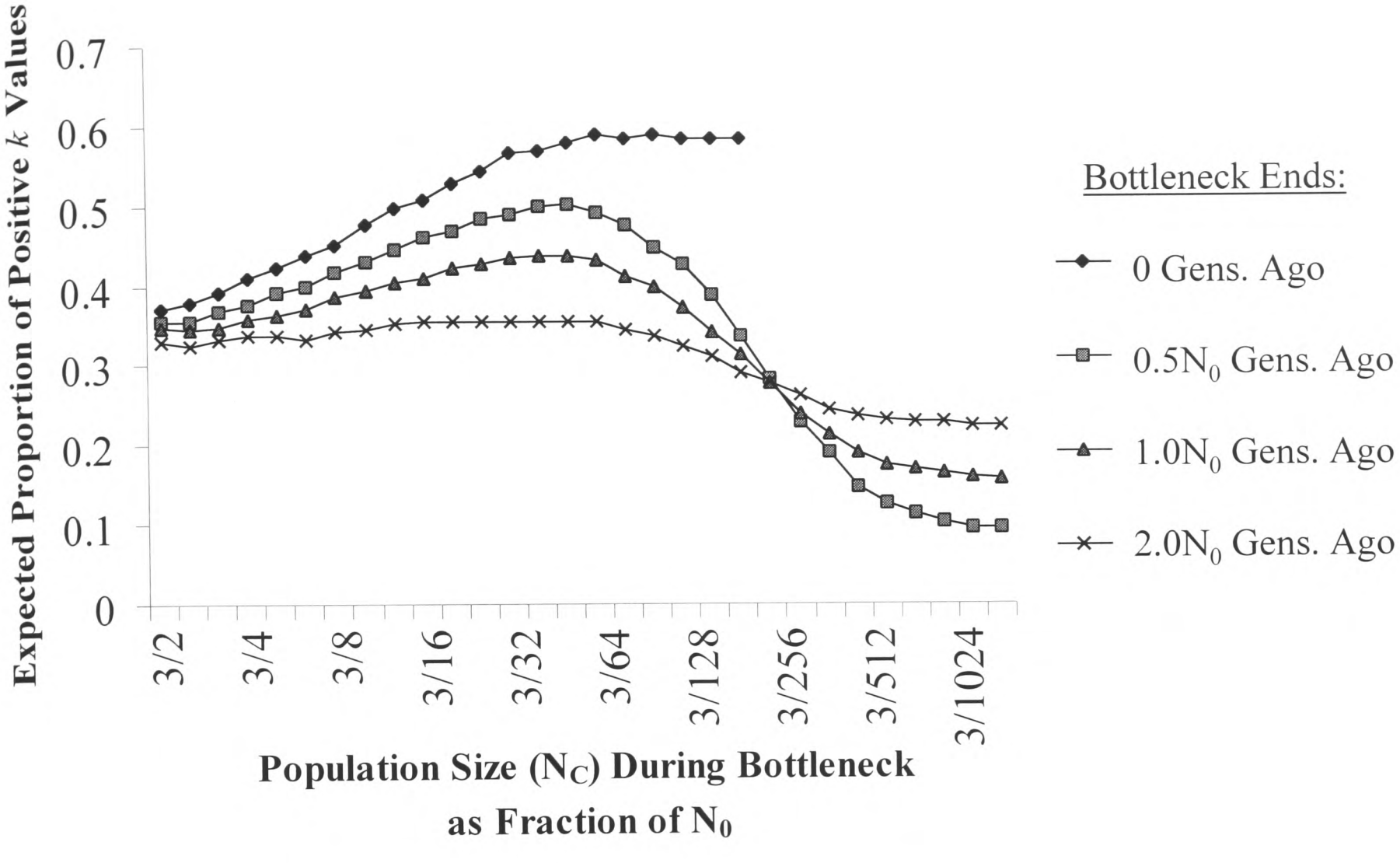

Expected g
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.4
0.8
1.6
3.2
6.4
12.8
25.6
51.2
Time Since Expansion
(Units of N0 Generations)
Doubles Every 0.1N0 Generations
Doubles Every 0.03 N0 Generations
Doubles Every 0.01N0 Generations
10,000,000-fold Sudden Expansion

FIGURE 7

(1)


Expected Proportion of Positive k Values
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3/2
3/4
3/8
3/16
3/32
3/64
3/128
3/256
3/512
3/1024
Population Size (N_C) During Bottleneck as Fraction of N_0
Bottleneck Ends:
0 Gens. Ago
0.5N_0 Gens. Ago
1.0N_0 Gens. Ago
2.0N_0 Gens. Ago

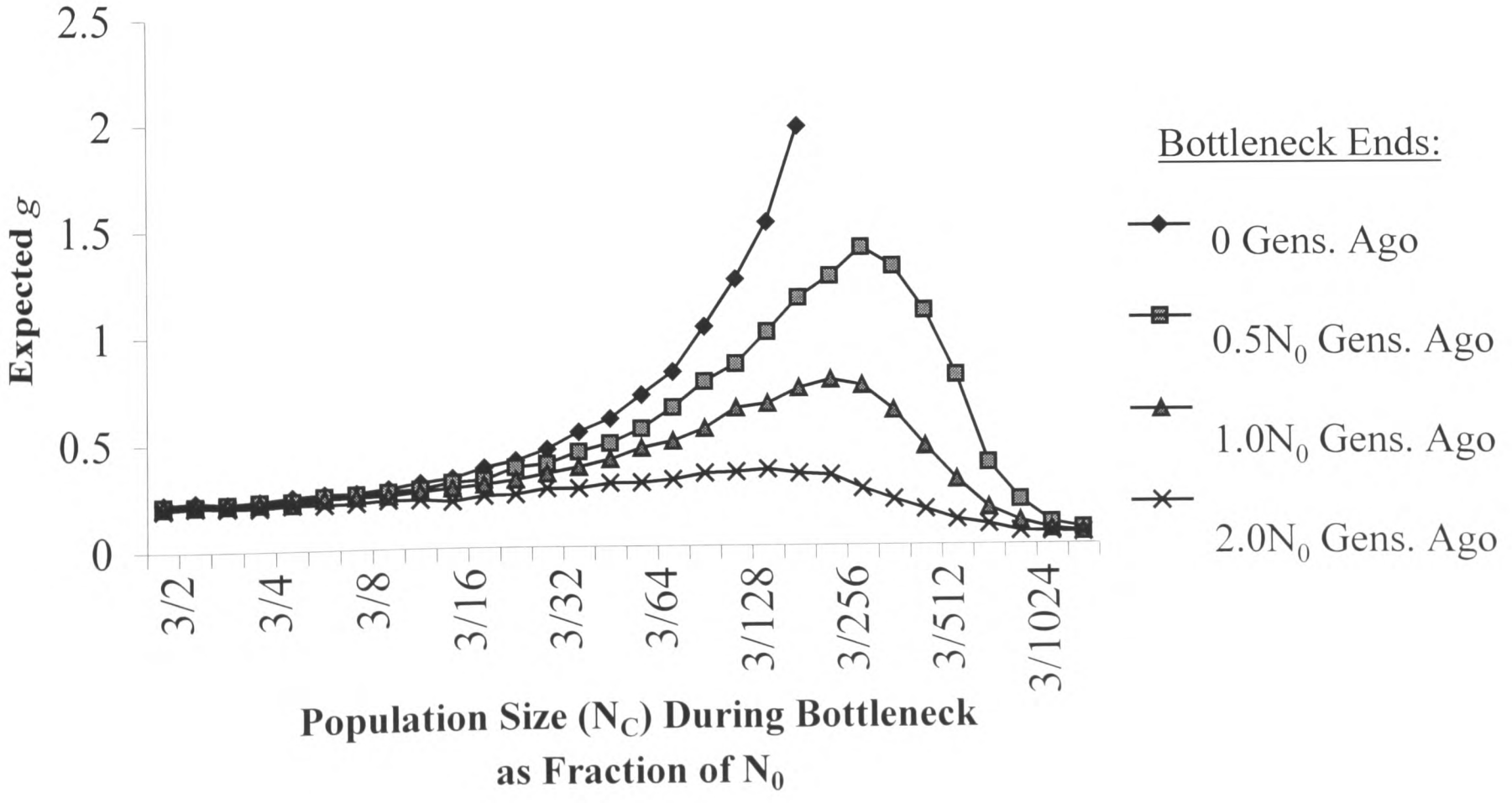
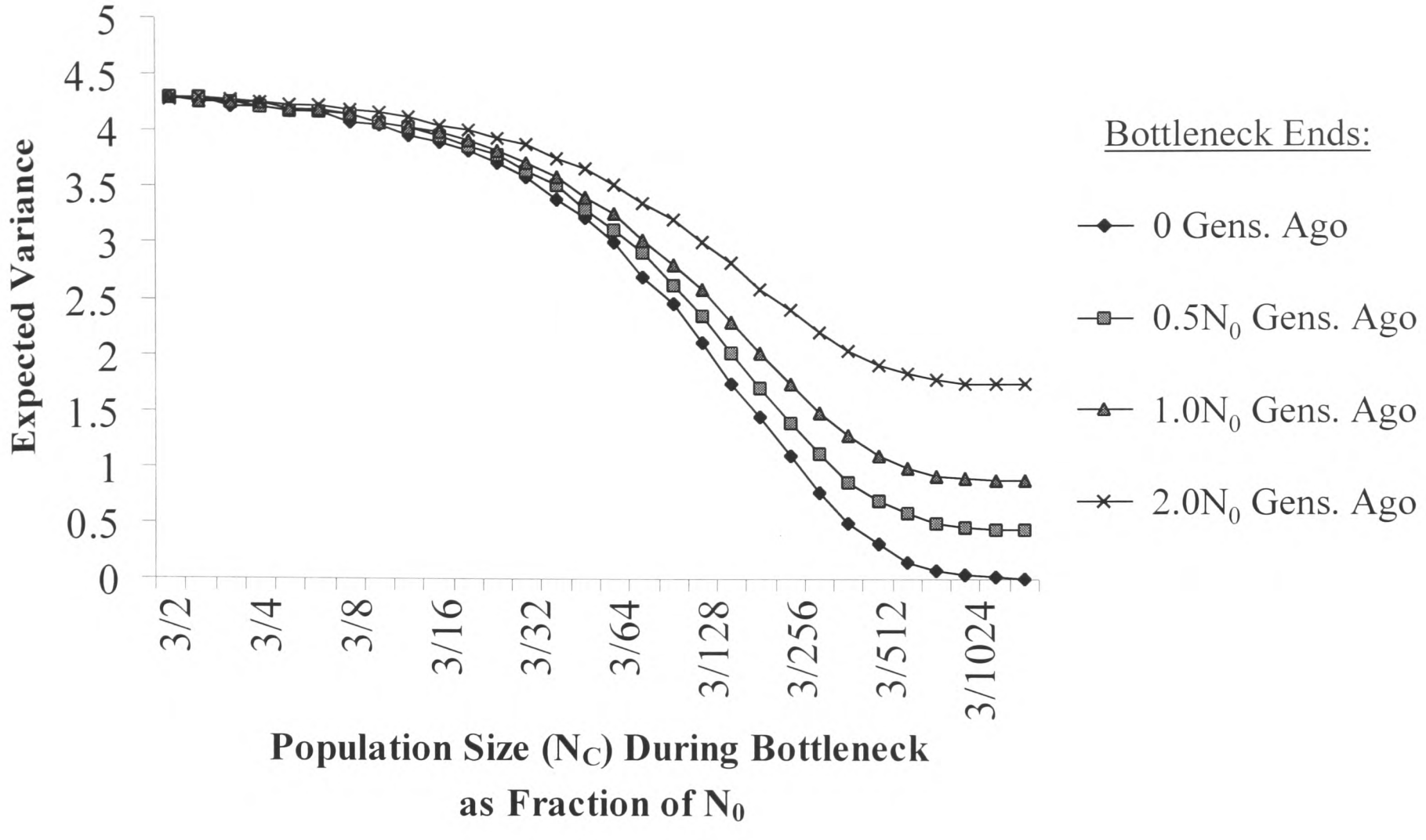

Expected g
2.5
2
1.5
1
0.5
0
3/2
3/4
3/8
3/16
3/32
3/64
3/128
3/256
3/512
3/1024
Population Size (N_C) During Bottleneck as Fraction of N_0
Bottleneck Ends:
0 Gens. Ago
0.5N_0 Gens. Ago
1.0N_0 Gens. Ago
2.0N_0 Gens. Ago

FIGURE 7 **(2)**



NSET) Model of Growth Interrupted by Bottleneck:

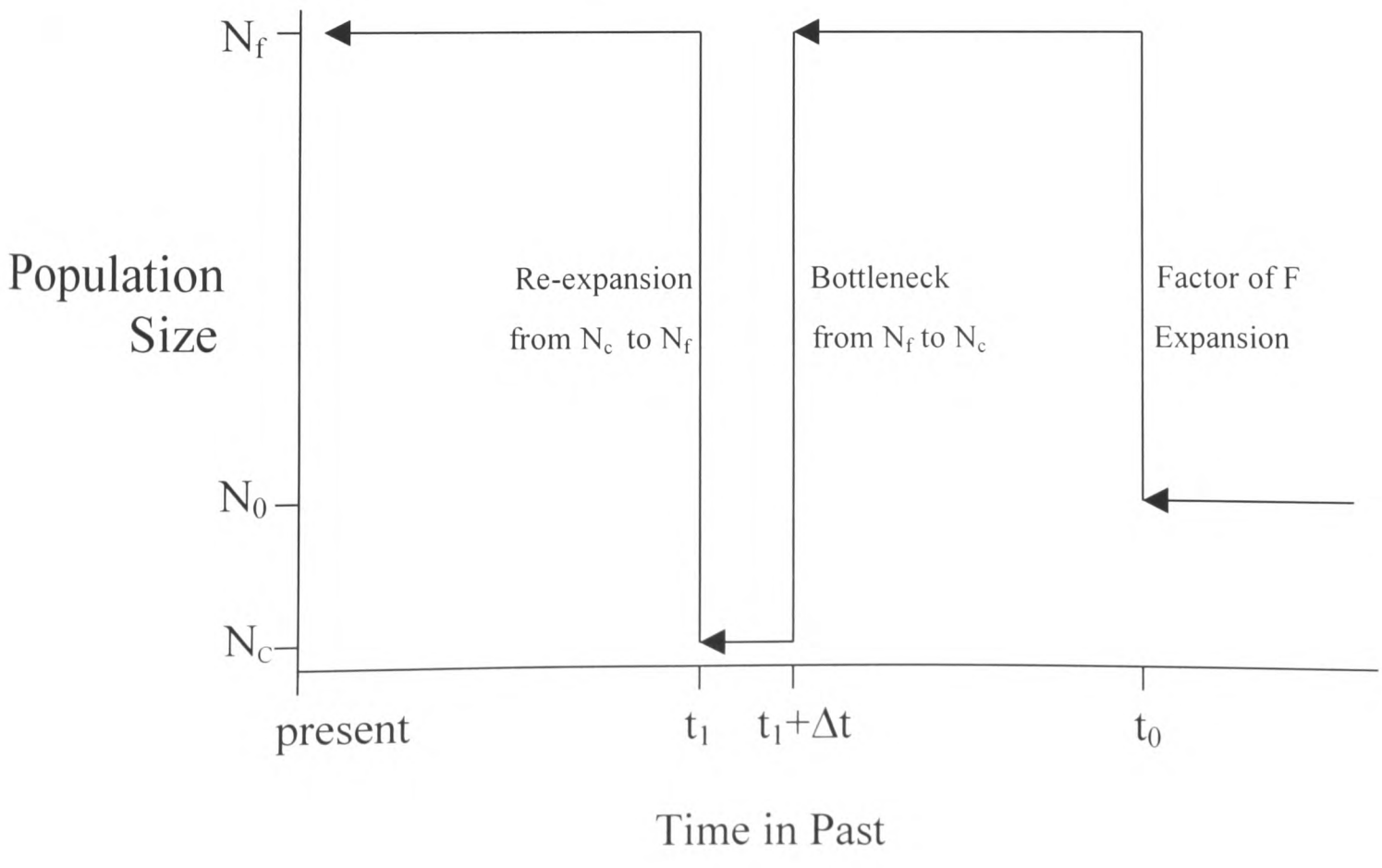
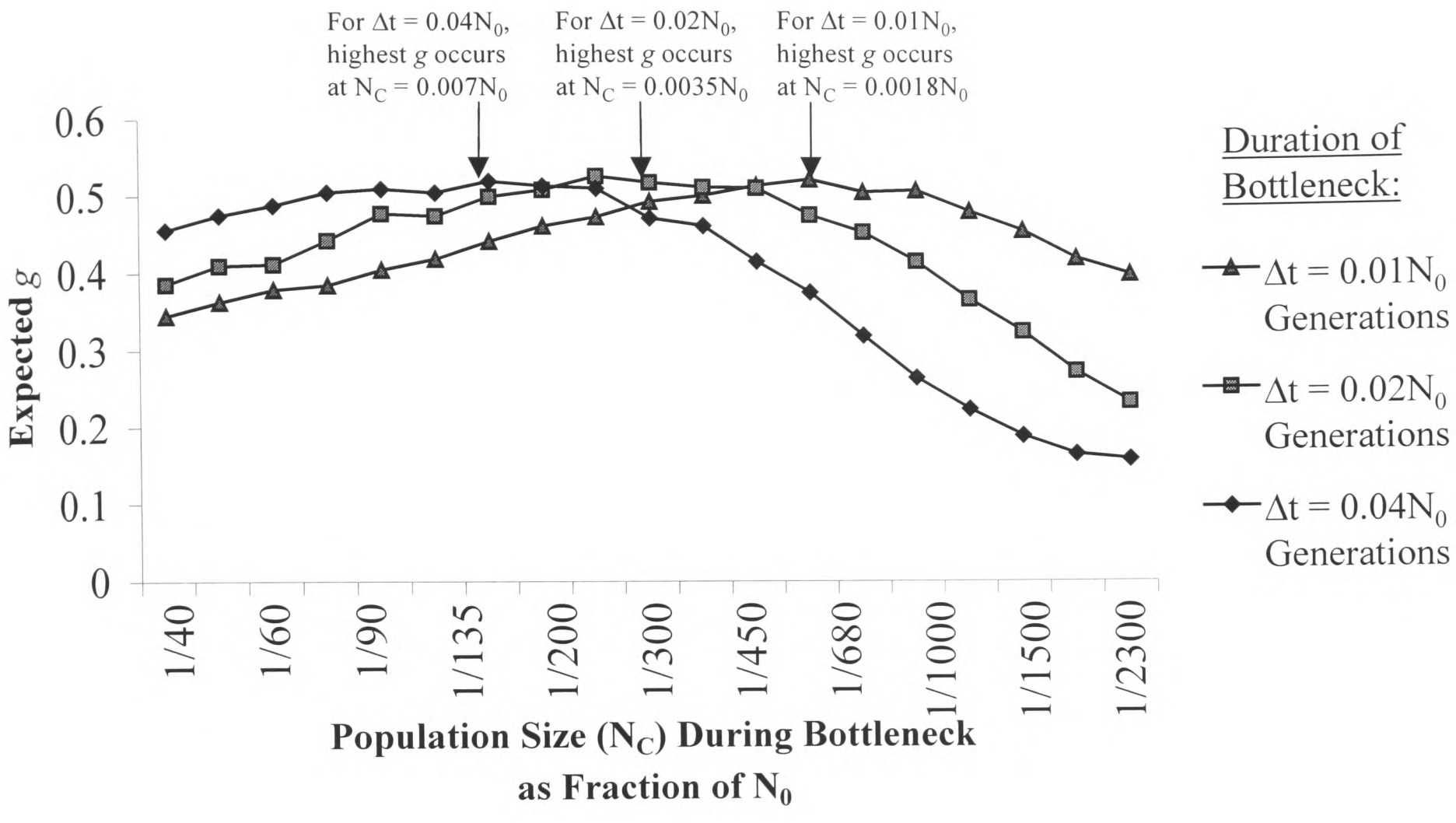


FIGURE 8


For Δt = 0.04N0, highest g occurs at NC = 0.007N0
For Δt = 0.02N0, highest g occurs at NC = 0.0035N0
For Δt = 0.01N0, highest g occurs at NC = 0.0018N0
Expected g
0.6
0.5
0.4
0.3
0.2
0.1
0
1/40
1/60
1/90
1/135
1/200
1/300
1/450
1/680
1/1000
1/1500
1/2300
Population Size (NC) During Bottleneck as Fraction of N0
Duration of Bottleneck:
Δt = 0.01N0 Generations
Δt = 0.02N0 Generations
Δt = 0.04N0 Generations

Single Nucleotide Polymorphisms as Tools for Studying Demographic History

David E. Reich

Department of Zoology, University of Oxford, and
The Whitehead Institute/MIT Center for Genome Research

ABSTRACT

Single nucleotide polymorphisms (SNPs) can provide information about demographic history via their allele frequencies. A simple test for population growth based on SNPs is more powerful, per locus screened, than an analogous microsatellite-based test. Date estimates for expansions detected using SNPs are unbiased by the magnitude of the expansions. When applied to an example data set, demographic analysis finds no evidence for an expansion among Caucasians, and in fact suggests a historical bottleneck in this group. However, the result will need to be replicated in a more carefully collected data set that is less subject to bias.

Genetic variation has emerged as a powerful tool for studying the histories of growth and contraction in natural populations. However, the tests of demography that have been proposed so far have all been problematic. The first generation of tests focused only on mitochondrial DNA (ROGERS and HARPENDING 1992), and with this type of data it is impossible to determine whether a detected signal of expansion is due to true demographic growth or to a selective sweep on one of the many genes linked to the mtDNA locus. The second generation of tests corrected for this by using multiple unlinked loci (DIRIENZO *et al.* 1998; KIMMEL *et al.* 1998; REICH and GOLDSTEIN 1998). However, the markers on which such tests are based, microsatellites, mutate according to a process that is not completely understood, and deviations from assumptions can affect demographic inferences (REICH *et al.* 1999). SNPs form an appealing alternative to microsatellites because they can also be collected from multiple sites in the genome, and

yet their mutation process is much better understood: most are thought to derive from a single historical mutation event.

DETECTION OF EXPANSIONS AND BOTTLENECKS

Test for Population Expansion: Assume that a SNP has been identified in a sample of n chromosomes, and that it has been found on a DNA fragment that is so short that there is a small probability of more than one polymorphism having been observed on it. This scenario is considered because it results in a simple analytical formula for the allele frequency distribution. Later in the paper, the scenario of longer DNA fragments is discussed. For the short fragment scenario, and for a freely mixing and constant-sized population that has not been subject to substantial selection, the probability that the less common allele is observed j times is:

$$P(j) \propto \frac{\delta_j}{j(n-j)}, \tag{1}$$

where $\delta_j = 1$ if $j < n\text{-}j$, and $\delta_j = .5$ if $j = n\text{-}j$. Expression 1 is a limiting case of the Ewens sampling formula (EWENS 1972; algebra not shown).

A test for growth should take advantage of the fact that population expansion alters the gene tree in a characteristic way, causing a disproportionate amount of the total branch-length of the tree to occur at the tips of the tree (DONNELLY 1996). A mutation occurring on such a tree will have fewer descendents than are expected in a constant-sized

population. Hence, an abnormally low allele frequency at a SNP is an indication of population expansion. A number of test statistics can measure low allele frequency, and this paper focuses on the simplest possible one, the mean frequency across unlinked loci. Significance cutoffs for a test based on the mean allele frequency can be obtained by using a computer to sample repeatedly from the distribution specified in Equation 1. A coalescent computer simulation (HUDSON 1991) that calculates cutoffs is given in Appendix 2 of this thesis.

Power of the test: The power to detect an expansion using the test based on the mean frequency is greater than that of the analogous approach for microsatellites (Figure 1) (REICH *et al.* 1999). Initially, this may seem surprising, since microsatellites are generally thought to contain more information per locus than SNPs (KRUGLYAK 1997). However, the within-locus k test for microsatellites makes much less efficient use of the information at each locus than the SNP test. The microsatellite test calculates a relatively crude summary statistic at each locus to assess whether the observed allele length distribution is more typical of a constant-sized or expanding population. By contrast, the SNP test derives its power by assessing the number of descendents of a mutant chromosome, a type of information that is specifically degraded in a microsatellite (but retained in a SNP) as alleles revert, through new mutation, to the ancestral type.

However, there is another type of information about demographic history that can still be studied most effectively using microsatellites. The variability in gene histories across loci tends to decrease systematically in the case of an expansion, an effect that can be measured much more precisely using the interlocus g test for microsatellites than by any SNP-based approach (REICH and GOLDSTEIN 1998).

Finally, in terms of the effect of sample size on the SNP-based tests, the power to detect an expansion is already substantial for greater than 8 samples, and nearly maximal for 20 or more samples (simulation results not shown). From the point of view of statistical power, small sample sizes are sufficient because the mean is a very simple summary statistic. It nevertheless remains desirable to use larger sample sizes because this increases confidence that the sample is representative of the population as a whole.

SNP-based tests are qualitatively similar to microsatellite-based tests: SNP-based tests for expansion behave quite similarly to microsatellite-based tests, due to underlying similarity of the gene genealogies. To obtain a qualitative picture of the behavior of the tests of demographic history, for a range of demographic scenarios, it is appropriate to refer to a detailed exploration of the within-locus *k* test for microsatellites (REICH *et al.* 1999). That study shows that the effect of expansion on a test is nearly independent of the factor of expansion as long as the rate of growth is at least moderate—which makes sense in light of the fact that the underlying gene trees are expected to be the same (REICH *et al.* 1999). The bewildering array of growth models that are in principle possible can therefore be assessed in terms of a much simpler range of scenarios. This increases the robustness and facilitates the interpretation of the tests, but at the same time reduces their power to detect some important demographic events. For example, while the tests are sensitive to ancient expansions from small population sizes, it is almost impossible to use them to detect recent expansions (REICH *et al.* 1999), such as the dramatic human expansion that is thought to have occurred during the past 10,000 years along with the invention of agriculture (CAVALLI-SFORZA *et al.* 1993).

Estimating a Date of Expansion: To estimate the date of a detected expansion, it is necessary to take account of the full shape of the allele frequency distribution. While a likelihood approach should be the most powerful way to perform this analysis, a more intuitive approach captures a great deal of information. The allele frequency distribution in the case of an expanding population will have unusually many low-frequency alleles and unusually few high-frequency alleles. In the intermediate range, there is one frequency, $\tilde{j}$, for which the probability in the case of an expansion is the same as the probability for a constant sized population. Note that this "frequency" is not a frequency in the usual sense, because it can be intermediate between two integer values. Figure 2 shows that for a sudden expansion, when the population grows instantaneously from a constant size *N/F* to a constant size *N*, this critical frequency, $\tilde{j}$, is independent of the factor of expansion *F* (see Appendix for an explanation). Moreover, the critical value, $\tilde{j}$, changes as a function of the time of expansion. Hence, the empirically observed $\tilde{j}$ can be used to directly predict the date of the expansion in a way that is independent of the factor of expansion. A computer simulation that generates curves appropriate for making extrapolations is provided in Appendix 2 of this thesis.

To obtain the date of expansion in units of generations, it is appropriate to make use of the observed heterozygosity in the population, *H*, as well as an estimate of the average mutation rate per generation per base pair. A standard estimator for heterozygosity is $\hat{\pi} = 2/\ell \sum_{i=1}^{q} p_i(1-p_i)$, where ℓ is the number of nucleotides screened for polymorphisms, q is the number of SNPs identified, and p_i is the frequency of the i^{th} SNP (HARTL and CLARK 1989). For a constant-sized population, the expected

heterozygosity is E[H] = $4N_0\mu/(4N_0\mu+1)$ (HARTL and CLARK 1989), but subsequent to a dramatic expansion, heterozygosity should increase at a rate of nearly 2μ per generation, as new mutations occur and the increase in genetic variability is not retarded by genetic drift. Hence, after a sudden and massive expansion, E[H] $\approx 2\mu t+4N_0\mu/(4N_0\mu+1)$ (Figure 3). Setting $\hat{\pi}$ equal to E[H], and employing estimates of t/N_f, $F=N_f/N_0$ and μ, it is possible to use this expression to obtain a rough estimate of the time t when the expansion began. Note that these inferences hold strictly true only for sudden, "stepwise" growth. However, the approximation of gradual growth as a sudden and massive expansion is likely to be appropriate even for moderately paced expansions (REICH *et al.* 1999).

Testing for a Bottleneck Using SNP Data: The allele frequency distribution contains information about events other than expansions. For example, a bottleneck will have two main effects on a gene genealogy. If the bottleneck is moderate in severity, multiple chromosomes that existed before the bottleneck will have modern descendents. A disproportionate fraction of the branch-length of the gene tree will then occur in the oldest, pre-bottleneck segments of the tree (DONNELLY 1996), and there will be an unusually high probability that the chromosome on which the mutation occurred has many modern descendents. On the other hand, a bottleneck can be so severe that only one ancestral lineage gave rise to the entire post-bottleneck population. A starlike genealogy is then expected, resulting in an allele frequency that is indistinguishable from what would be expected after an expansion. Thus, if a bottleneck is dramatic but not extremely severe, the mean allele frequency will be raised, which can be used as a test for a bottleneck or a contraction.

Higher order moments of the allele frequency distribution can also be used to make inferences about population history, including bottlenecks. The variance divided by the square of the mean (i.e., the "normalized variance," which is related to the second moment of the allele frequency distribution) tends to be elevated in the case of a bottleneck, an effect that is somewhat analogous to the rise of the g statistic in the case of the interlocus test for microsatellites (REICH *et al.* 1999). Although this paper focuses on the mean frequency, a test based on the normalized variance is highly complementary to one based on the mean frequency, since their maximum sensitivities are to bottlenecks of quite different severity.

CONFOUNDING EFFECTS

DNA Fragment Length: The probability of observing a polymorphism in a given DNA fragment is $1\text{-}e^{-\mu\lambda\tau}$, that is, 1 minus the zero-order term of a Poisson series, where μ is the mutation rate per base pair per generation, λ is the length of the DNA fragment, and τ is the total branch-length of the gene tree connecting the sampled chromosomes (τ represents the total opportunity, in units of time, for SNP-producing mutations to occur and be observed at a locus). Since gene trees with large τ's look different than those with low τ's (and have different expected allele frequencies), it is important to explore how varying the length of DNA fragments shifts the balance between these two classes of gene trees and thereby affects the mean allele frequency. Coalescent computer simulations, following an algorithm of R. HUDSON (1991), were carried out for a sample size of 40 and a variety of demographies. These simulations show that for a constant-sized population,

the mean copy number of the less common allele is 6.4 for $1\text{-}e^{-\mu\lambda\tau} << 1$, and 6.2 for $1\text{-}e^{-\mu\lambda\tau} \approx 1$, a small but non-conservative effect on the test statistic. For the case of a bottleneck, variation in gene histories across loci tends to be greater than in a constant-sized population (DONNELLY 1996), and hence the difference in expected allele frequency between loci with large τ's and those with small τ's is even greater than for a constant-sized population. To establish significance cutoffs, however, all that matters is the situation for a constant-sized population. A simulation that produces significance cutoffs, and that does so while taking into account the average value of $1\text{-}e^{-\mu\lambda\tau}$ across loci, is provided in Appendix 2 of this thesis.

Natural Selection: If a SNP is near a locus that has undergone a selective sweep, it may contribute to a false-positive statistical signal of expansion, even though there has been no growth in the population as a whole. To minimize this effect, it is best to choose SNPs that are as far away as possible from genes. In addition, to increase confidence in a signal of population expansion, it is useful to compare the results of the SNP test for expansion, which is non-conservative, to a test that responds conservatively to selection, such as the interlocus *g* test for microsatellites (REICH *et al.* 1999).

Population Subdivision: Population subdivision can occur either as a result of historical admixture, or due to the lumping of differentiated populations in the same sample. Many alleles are in measurable frequency only in specific populations, and hence when populations are lumped together, the frequencies of these "private" alleles are diluted and the overall mean frequency goes down. In principle, it should be possible to make inferences about the history of subdivision using the SNP allele frequency distribution. However, this paper focuses on the simpler case of homogenous populations.

Cryptic Relatedness: If multiple samples are taken from the same family or even from the same individual, the mean allele frequency will be elevated due to multiple observations of the same alleles. To detect multiple sampling, it is sufficient to count the number of shared alleles between all possible pairs of individuals. Related individuals will share a strikingly large number of alleles, and can be rejected from an analysis on this basis.

Ascertainment Bias: Tests of demographic history are ultimately affected by the strategy used to identify polymorphisms, since different approaches will yield different allele frequency distributions. Indeed, the need to identify polymorphisms in a controlled and careful way is much more important for SNP-based than for microsatellite-based studies. One form of ascertainment bias arises when SNPs are first identified as polymorphisms in small samples and then scored for allele frequency in larger samples. This flattens the allele frequency distribution and biases the test of demography. However, if the prescreening strategy is taken into account when calculating significance cutoffs, there will be no bias. Moreover, as will be seen in a later section, the power of the tests of demography can actually increase when a prescreening strategy is employed.

A major requirement for SNP-based tests is that the polymorphism be identified in the study population, or at least in a closely related population. If the SNPs were identified in a different population, the allele frequency distribution in the study population will depend not only on its own demographic history, but also on the relationship between the two populations. This will complicate demographic inferences.

Low Frequency Polymorphisms: Most procedures that screen for SNPs are less than 100% accurate, and low frequency polymorphisms are likely to be missed more

commonly than high frequency ones. The result is a deficiency in the number of low frequency SNPs in a distribution. To reduce this bias, it is possible to drop loci from the analysis that have allele frequencies in the lowest frequency categories (e.g., drop all singleton and doubleton loci), and to adjust significance levels accordingly. The computer simulation in Appendix 2 of this thesis provides an option for dropping low frequency loci and obtaining revised significance levels.

ALTERNATIVE DATA FORMATS

When the Derived (Mutant) Allele is Known: If it is possible to genotype a closely related species at the study loci (for example, a chimpanzee in the case of humans), the allele frequency distribution becomes more informative. The derived (or mutant) allele, rather than the less common allele, can then be scored. Figure 1 shows the improved power to detect an expansion when the mean frequency of the derived allele is used in a test for expansion. Significance levels for this revised test can be obtained using the computer simulation provided in Appendix 2 of this thesis.

Prescreening for SNPs: An alternative way to identify SNPs is to prescreen them in a sample of *m* chromosomes, and then genotype the ones that are found to be polymorphic in additional chromosomes for a total of *n* chromosomes. Alternatively, the allele frequencies of the SNPs that are found to be polymorphic in the prescreening sample of *m* chromosomes can be genotyped in an entirely new sample of *n* chromosomes as a test of demographic history. Either approach saves costs in the preliminary stage of identifying SNPs, since relatively few genotypings are required, and yet large sample

sizes can be obtained. Figure 1 shows that such a method of prescreening actually increases the power to detect an expansion for a fixed sample size, and is simultaneously more efficient from an experimental point of view. Significance cutoffs for a test that uses prescreening can be obtained from the computer simulation in Appendix 2 of this thesis.

APPLICATION TO REAL DATA

Data: The SNP-based test for expansion was applied to a data set that had been generated for another purpose at the Whitehead Institute/MIT Center for Genome Research (unpublished data kindly provided by David Altshuler, Michele Cargill, Eric Lander and Kerstin Lindblad). The data collection was not designed for making demographic inferences, and hence the observed signal of a bottleneck (see below) must be treated with caution. Nevertheless, the data and analysis are presented because they provide guidelines about how to carry out a proper SNP test of demographic history, and for verifying whether a Caucasian bottleneck in fact occurred.

Two individuals from Utah (K1331-1, -13), 3 Amish (K884-2, -15, -16), and 2 individuals from Venezuela (K104-1, -16) were screened for SNPs. The individuals all come came from CEPH pedigrees (Centre d'Etude du Polymorphisme Humain), and the DNA was extracted from immortalized cell lines. Polymorphisms were identified by screening 16,725 sequence tagged sites (STSs) in these 7 samples using DNA resequencing arrays ("DNA chips") (WANG *et al.* 1998). Among the STSs, 2,299 were identified as polymorphic using a clustering algorithm implemented by computer, followed by visual inspection. For validation purposes, a subset of the polymorphic STSs

were sequenced on both strands using an ABI 377, and rescreened using a second computer program that is known to be highly accurate (NICKERSON *et al.* 1997). Thus, 113 autosomal SNPs were identified that could be scored by hand in all 7 cell lines for the purpose of making inferences about demographic history. The number of copies of the less common allele was counted for each SNP; for STSs containing more than one SNP, one was chosen at random for scoring

Results: The observed allele frequency distribution differs substantially from the theoretical expectation (Figure 4A). In particular, the mean frequency is elevated above its expectation ($P < 0.98$), suggesting a bottleneck. An intriguing possibility is that the bottleneck signal is the same as what was suggested by microsatellite-based studies of human demography—an event that occurred during the emergence of the first modern humans out of Africa around 100,000 years ago (REICH and GOLDSTEIN 1998; KIMMEL *et al.* 1998).

Population Subdivision: The suggestion of a bottleneck becomes even stronger when population subdivision is taken into account, since population subdivision lowers the mean allele frequency. To produce a more representative "Caucasian" sample, two Venezuelan DNAs were dropped from the analysis. This leaves 3 Amish and 2 Utah DNAs. In this more homogenous sample, the *P*-value increases to $P < .998$ ($P < .002$ in an explicit test for a bottleneck) (Figure 4B) (Table 2). As the Amish are well known to have undergone a recent bottleneck, it is possible that this history makes some contribution to the elevated mean allele frequency.

Ascertainment Bias: A false-positive signal of a bottleneck could occur if high frequency SNPs were ascertained with more probability than low frequency ones while

screening for SNPs. The extent of this bias would have had to be quite large, however, in order to produce the observed effect. Fully 27 missed SNPs, all with one copy of the less common allele, would have had to be missed in order to raise the mean as far above the expectation for a constant-sized population as was in fact observed.

In order to obtain a qualitative assessment of the extent of bias, and to provide a minimally biased test for demography, the mean frequency was recalculated after dropping all SNPs with frequency 1 from the analysis (singletons), and then dropping all SNPs with frequency 1 and 2 (singletons and doubletons). Table 1 shows that *P*-values indeed decline when low-frequency SNPs are dropped, which suggests that ascertainment bias affects the data. Even when low frequency alleles are dropped, however, the mean frequency remains well above the expectation for a constant-sized population. Thus, while ascertainment bias is likely to be present, the data are sufficiently suggestive of a bottleneck that it seems worthwhile to replicate the SNP test of demography a new Caucasian data set, one that is collected more appropriately for the purposes of the test.

DISCUSSION

SNP studies of demography are appealing because they are nearly immune from uncertainty about the mutation process, and have much more power to detect expansion than comparable tests using microsatellites. However, SNP tests can only be carried out if data sets are collected in a way that does not bias allele frequencies. To take advantage of a SNP-based approach, it is necessary not only to use a large number of individuals (at least 20 chromosomes per population), and many loci (at least 100), but also to screen for

polymorphisms using a method that has an equal chance of missing low-frequency and high-frequency SNPs. If data sets of this sort can be collected, SNPs should become tools of choice for making genomic inferences about demographic history.

ACKNOWLEDGEMENTS

This paper extends work with David Goldstein on microsatellite-based tests of demographic history, and I am grateful for his comments on early drafts of the manuscript. Other aspects benefited from conversations with Ranajit Chakraborty, Michael Charleston and Oliver Pybus. David Altshuler made a valuable contribution by suggesting the use of the example data set, and he and other members of The Whitehead Institute/MIT Center for Genome Research, including Michele Cargill, Eric Lander and Kerstin Lindblad, provided useful comments and were tremendously generous in allowing me to analyze the unpublished data. I was supported by a National Defense Science and Engineering Graduate Fellowship.

APPENDIX

In a suddenly expanded population, there is a unique number of copies of the derived allele, $\tilde{j}$, for which the expected proportion of polymorphisms with that frequency is the same for a constant-sized population as it is for a suddenly expanded population (Figure 2). Note that $\tilde{j}$ may not be an integer value, and hence it is not strictly a frequency even though it will be treated as if it were a true frequency in the discussion that follows.

In order to understand why $\tilde{j}$ is independent of the factor of expansion, define $P(j)$ as the probability that there are j copies of the mutant allele in a sample of size n. To

calculate $P(j)$, it is necessary to know the shape of the genealogical tree connecting the samples at a locus. Designate the k^{th} time segment as the period of time in the past for which there were k ancestors of the modern n samples. L_k is then the probability, which is a function of demographic history, that the mutation occurred in the k^{th} segment, and a_{jk} as the probability that a mutation occurring in the k^{th} segment has j descendents in the modern population, which is a known quantity for a panmictic population. $P(j)$ can then be written as a linear combination of a_{jk}'s and L_k's:

$$P(j) = \sum_{k=2}^{n} a_{jk} L_k \ . \qquad (2)$$

For a sudden, F-fold expansion that occurred at a time t in the past in a freely mixing population, the expected lengths of the segments in the genealogical tree change as a function of F and t, but the a_{jk}'s do not change. Let Λ_k be the value of L_k for a constant-sized population (known analytically), and let $L_{k,F,t}$ be the value of L_k in the case of an expansion. $L_{k,F,t}$ can be derived in terms of Λ_k by noting that if a mutation occurs in the k^{th} segment, the probability that it occurred after the expansion is $p_k(t)$, and the probability that it occurred before the expansion is $1\text{-}p_k(t)$. The expected length of a segment in the gene tree prior to the expansion is equal to the expectation assuming that no expansion occurred, divided by a factor of F (HUDSON 1991). Thus, $L_{k,F,t} = \Lambda_k p_k(t) + \Lambda_k(1\text{-}p_k(t))/F = \Lambda_k(1+(p_k(t)\text{-}1)(F\text{-}1)/F)$, and:

$$P_{F,t}(j) = \frac{\sum_{k=2}^{n} \Lambda_k a_{jk} \left(1 + (p_k(t) - 1)\left(\frac{F-1}{F}\right)\right)}{\sum_{s=2}^{n} \Lambda_s \left(1 + (p_s(t) - 1)\left(\frac{F-1}{F}\right)\right)} \tag{3}$$

where $P_{F,t}(j)$ represents the value of $P(j)$ that is expected in the case of an expansion, and the denominator in Equation 3 is a normalization. For $F \to \infty$ (a massive expansion), and for $F \to 1$ (a constant-sized population), $P_{F,t}(j)$ reaches its asymptotic values, and the degree to which these asymptotes have been reached is $(P_{F,t}(j)-P_{1,t}(j))/(P_{\infty,t}(j)-P_{1,t}(j))$. With some algebra, this leads to the result:

$$\frac{P_{F,t}(j) - P_{1,t}(j)}{P_{\infty,t}(j) - P_{1,t}(j)} \equiv H(F,t) = \frac{\sum_{k=2}^{n} \Lambda_k p_k(t)}{\sum_{k=2}^{n} \Lambda_k \left(F/(F-1) + p_k(t) - 1\right)}, \tag{4}$$

The quantity $H(F,t)$ is independent of j, and yet it describes the degree of approach to the asymptote for all j. Rearranging to obtain $P_{F,t}(j) = H(F,t)(P_{\infty,t}(j)-P_{1,t}(j))+P_{1,t}(j)$, it is clear that for any $\tilde{j}$ satisfying $P_{\infty,t}(\tilde{j}) = P_{1,t}(\tilde{j})$, the equation $P_{F,t}(\tilde{j}) = P_{1,t}(\tilde{j})$ holds regardless of the value of F. Hence, there is a unique frequency, $\tilde{j}$, for which the expected proportion of loci with that frequency is the same for a constant sized population as for a suddenly expanded population. This frequency is functionally related to t, but independent of F as observed in Figure 2. The frequency can be used to estimate a date of expansion that is independent of the factor of expansion F. A computer simulation that generates the $P_{F,t}(j)$ curve for the purpose of extrapolation can be obtained from Appendix 2 of this thesis.

LITERATURE CITED

CAVALLI-SFORZA, L. L., P. MENOZZI and A. PIAZZA, 1993 *The History and Geography of Human Genes*. Princeton University Press, Princeton, New Jersey.

DIRIENZO, A., P. DONNELLY, C. TOOMAJIAN, B. SISK, A. HILL, *et al.*, 1998 Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics **148:** 1269-1284.

DONNELLY, P., 1996 Interpreting genetic variability – the effects of shared evolutionary history. Ciba Foundation Symposia **197:** 25-40.

HUDSON, R. R., 1991 Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) **7:** 1-44 (Oxford University Press, Oxford).

KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS and L. B. JORDE, 1998 Signatures of population expansion in microsatellite repeat data. Genetics **148:** 1921-1930.

KRUGLYAK, L., 1997 The use of a genetic map of biallelic markers in linkage studies. Nature Genetics **17:** 21-24.

LI, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts.

NICKERSON, D. B., V. O. TOBE and S. L. TAYLOR, 1997 PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res **25:** 2745-2751.

REICH, D. E. and D. B. GOLDSTEIN, 1998 Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. U.S.A. **95:** 8119-8123.

REICH, D. E., M. W. FELDMAN and D. B. GOLDSTEIN, 1999 Statistical properties of two tests that use multilocus data sets to detect population expansion. Mol. Biol. Evol. **16:** 453-466.

ROGERS, A. R. and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552-569.

WANG D. G., J.-B. FAN, C.-J. SIAO, A. BERNO, P. YOUNG, *et al.*, 1998 Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science **280:** 1077-1082.

TABLE 1

	14 chromosomes (no. of loci used in test)	**10 chromosomes (no. of loci used in test)**
All frequencies	$P < .98$ (113)	$P < .998$ (103)
SNPs with frequency ≥ 2	$P < .93$ (82)	$P < .96$ (74)
SNPs with frequency ≥ 3	$P < .88$ (63)	$P < .85$ (52)

Table 1 *P*-values for the SNP-based test for expansion. The left column corresponds to 7 individuals (14 chromosomes), and the right column corresponds to 5 individuals (10 chromosomes). Two Venezuelan samples are dropped to produce the data in the right column. *P*-values are calculated on the basis of Expression 1, which applies because DNA fragments are so short that no more than one SNP usually occurs per STS.

FIGURES

Fig. 1 SNP tests based on the mean allele frequency have dramatically more power to detect expansion than the within-locus *k* test for microsatellites. Graphs were all generated for a scenario of 100-fold growth occurring *1N* generations ago and a sample size of 40. The graph for microsatellites was published previously (REICH *et al.* 1999). The graphs for the three SNP-based tests were generated on the basis of 10,000 coalescent computer simulations, and power was assessed as the proportion of replicates below the $P < 0.05$ cutoff for a constant sized population. The basic mean frequency test, and the test based on the mean frequency of the derived allele, are described in the text. For the test with prescreening, $m = 10$ chromosomes are evaluated in order to discover polymorphisms. All polymorphisms are then screened in an additional 30 chromosomes (for a total of $n = 40$ chromosomes) to obtain the allele frequency distribution.

Fig. 2 The effect of a sudden expansion $0.5N$ generations ago on the allele frequency distribution. The sample size is 40, and the pre-expansion population size is given in the legend. The probability of observing a SNP with a given allele frequency is presented as the ratio of the probability assuming that an expansion occurred, to the expected probability for a constant-sized population (derived from Expression 1). Both the expanding and constant-sized population probabilities are obtained by averaging over 1,000,000 replicates. The distribution approaches an asymptote for sufficiently dramatic or sufficiently minor expansions, and there is one special frequency for which the expected value is equal to the observed value, regardless of the factor of expansion.

Fig. 3 Expected heterozygosity as a function of time, for a dramatic expansion that occurred *t* generations ago. The expected heterozygosity begins at a constant value determined by mutation-drift equilibrium: $E[H] = 4N_0\mu/(4N_0\mu+1)$. After a dramatic expansion, it increases linearly with time, unimpeded by genetic drift, at a rate equal to about twice the mutation rate. The reason for the factor of 2 is that a new mutation can occur on either of the 2 samples that are being compared.

Fig. 4 Comparison of real data to the expectation for a constant-sized population. Figure (A) compares the observed to the expected distribution for all 7 samples (14 chromosomes). Figure (B) shows the distribution for the more homogenous Caucasian population (two Venezuelans are dropped from the analysis). The expected curve is given by Expression 1, which applies in this case because DNA fragments are so short that no more than 1 SNP usually occurs per STS.

FIGURE 1

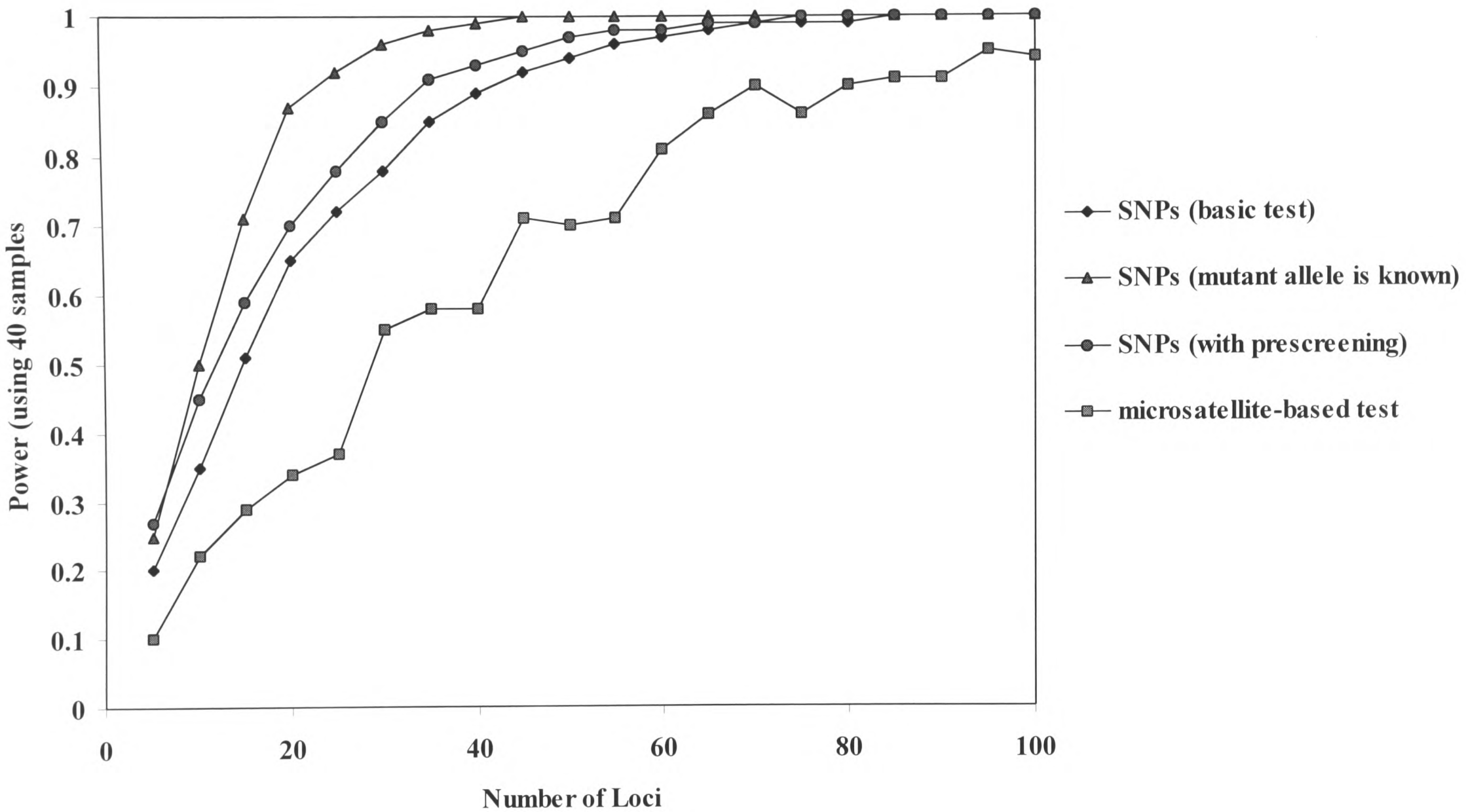

Power (using 40 samples
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
Number of Loci
SNPs (basic test)
SNPs (mutant allele is known)
SNPs (with prescreening)
microsatellite-based test

FIGURE 2

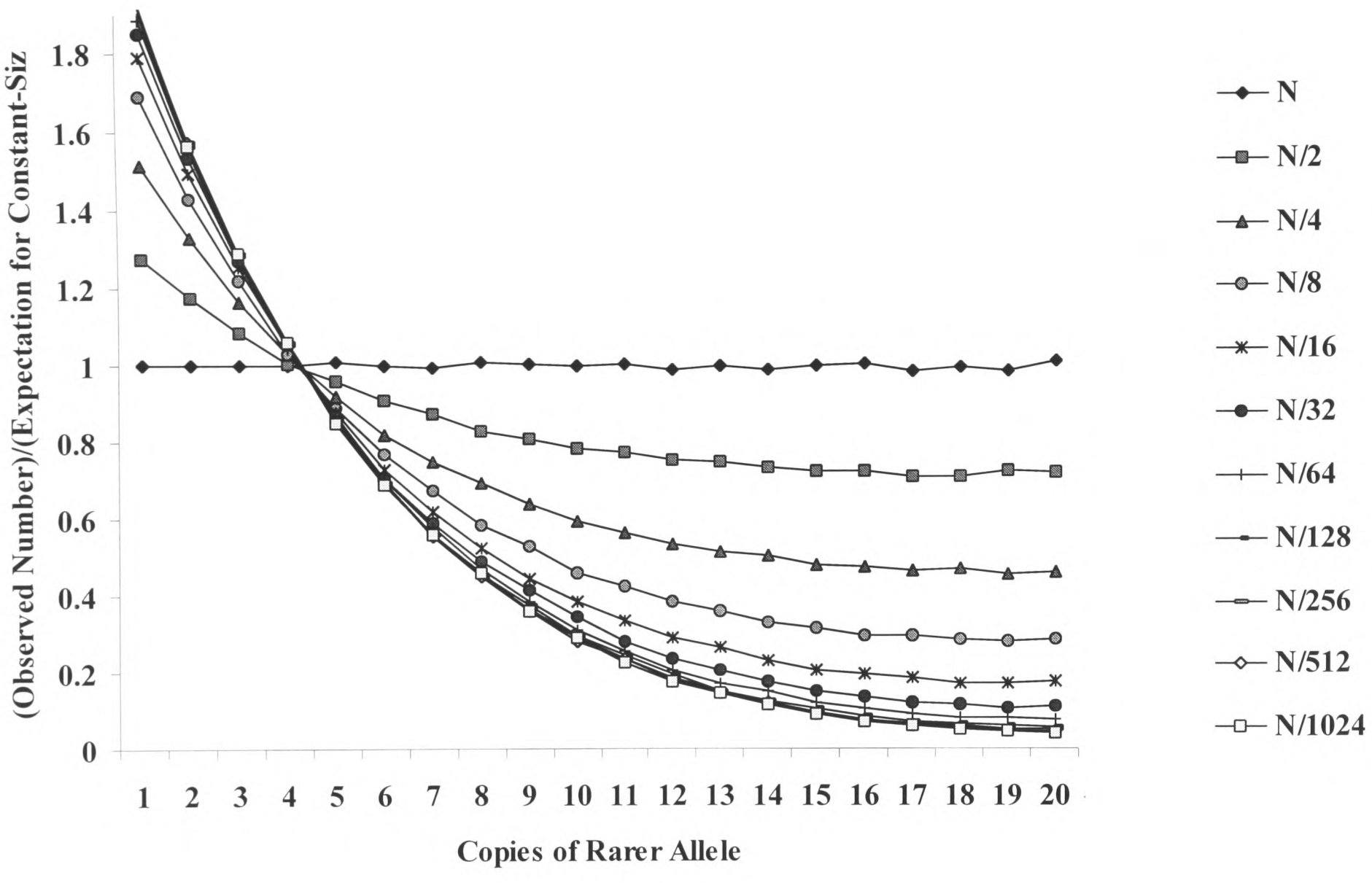

(Observed Number)/(Expectation for Constant-Siz
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Copies of Rarer Allele
N
N/2
N/4
N/8
N/16
N/32
N/64
N/128
N/256
N/512
N/1024

FIGURE 3

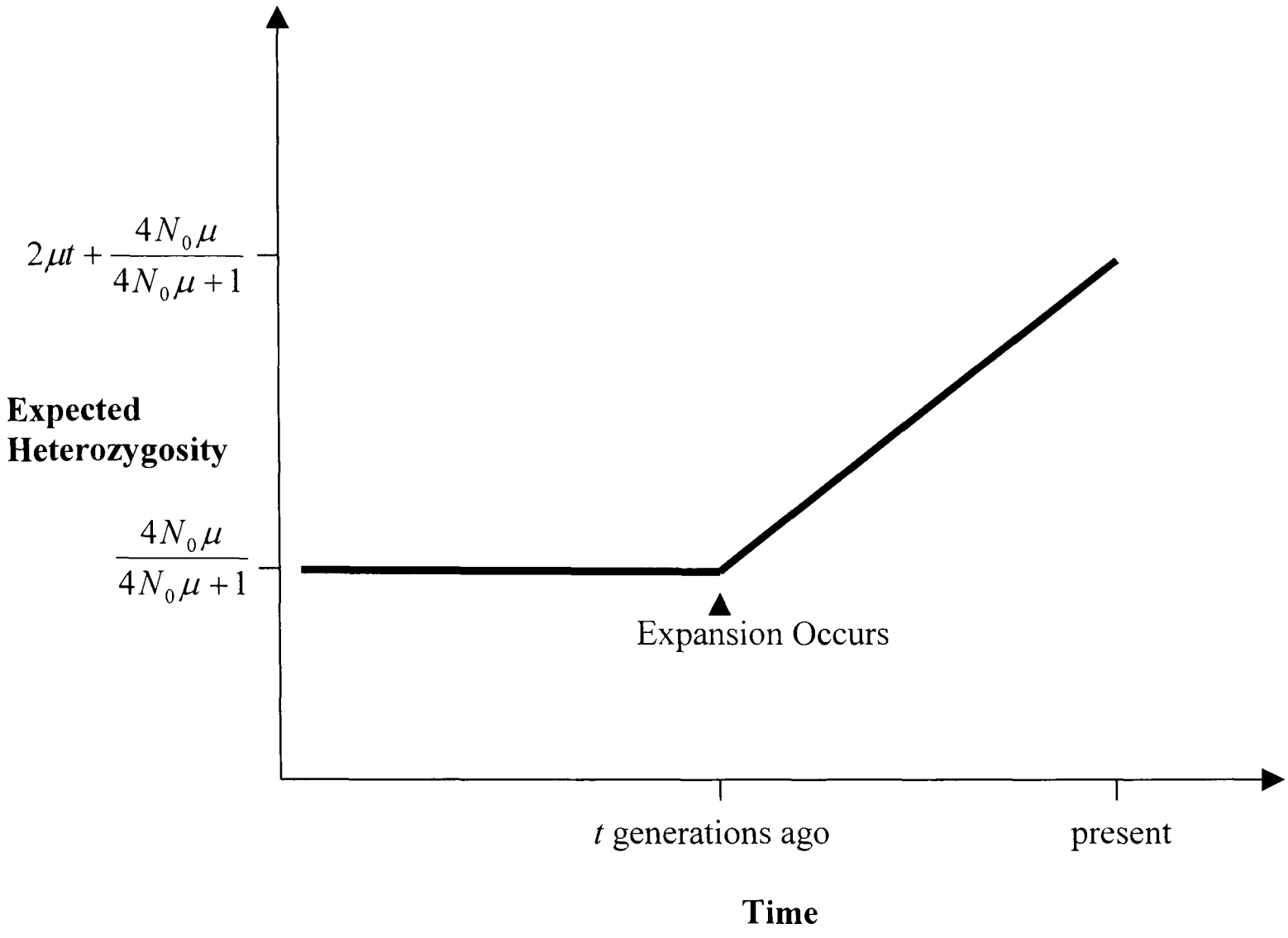

$2\mu t + \frac{4N_0\mu}{4N_0\mu+1}$
Expected Heterozygosity
$\frac{4N_0\mu}{4N_0\mu+1}$
Expansion Occurs
t generations ago
present
Time

FIGURE 4 **(1)**

(A) 7 Individuals

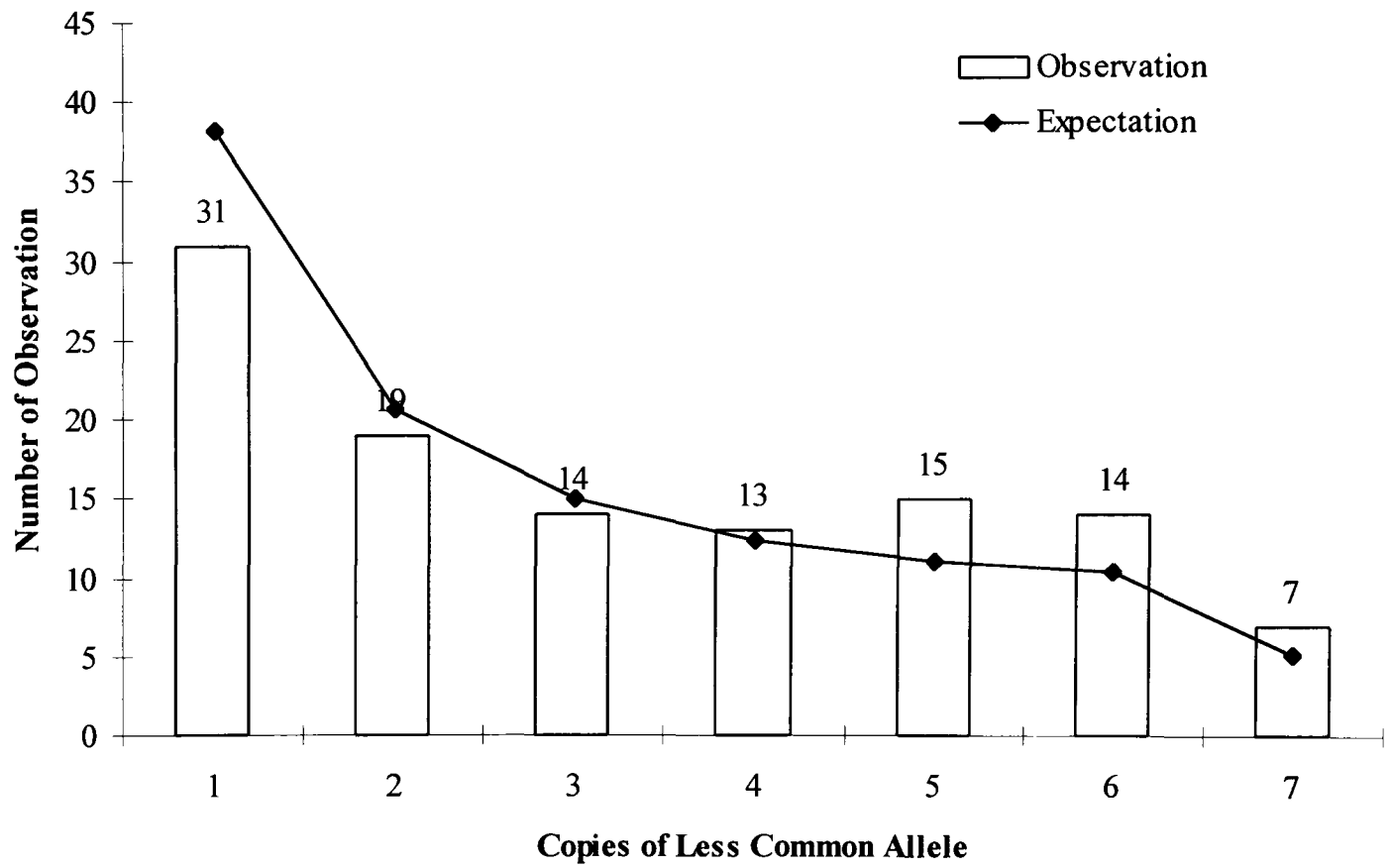
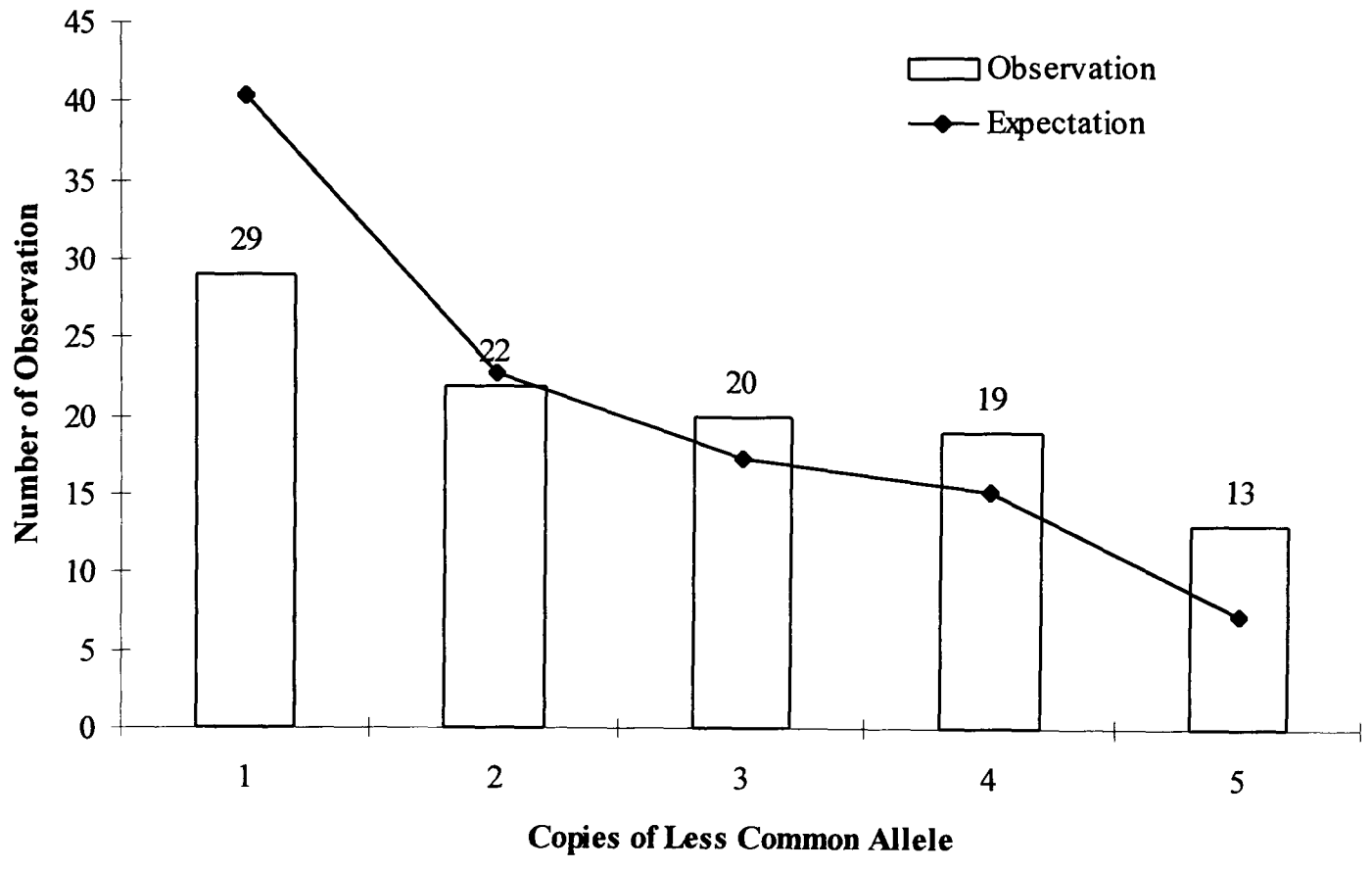


FIGURE 4 **(2)**

(B) 5 Individuals (more homogenous sample)



Chapter 2

Use of Linkage Disequilibrium to Date the Most Recent Common Ancestors of Disease Mutations in Populations

Part A: "Estimating the ages of the most recent common ancestors of mutations using variation at linked markers, and application to the CCR5-Δ32 AIDS-resistance allele" was published as a book chapter in *Microsatellites: Evolution and Applications*, ed. D.B. Goldstein and C. Schlötterer, Oxford Univ. Press, 129-138 (1999). Some changes in the present text, in particular Table 1 and Figure 1, are new. This paper provides a detailed description of the data analysis performed in "Dating the origin of the CCR5-Δ32 AIDS resistance allele by the coalescence of haplotypes," which was published in *Am. J. Hum. Genet.*, **62:**1507-1515 (1998).

Part B: "Estimating the ages of the most recent common ancestors of two common mutations causing Factor XI deficiency in Jews" is based on the "Data and Analysis" section of a paper that was published in *Am. J. Hum. Genet.*, **62:**1507-1515 (1998); David Goldstein was the first author, while I was the second author. Although I did the primary analysis for the paper, I did not write it, so I have not included the entire text. The document has been substantially reworked for this thesis, and in particular, the title, introduction, and conclusions are new. The full author list is D.B. Goldstein, D.E. Reich, N. Bradman, S. Usher, U. Seligsohn and H. Peretz.

Estimating the Ages of the Most Recent Ancestors of Mutations Using Variation at Linked Markers

David E. Reich and David B. Goldstein

Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK.

(A) Introduction

We present a general method for estimating the dates of most recent common ancestors of mutations in a sample (MRCAs) using variation at linked microsatellite markers. Risch et al (1995) take a similar approach to estimating the age of the MRCA of the mutation causing idiopathic torsion dystonia among Ashkenazic Jews, but they do not describe how to produce a confidence interval for the date. Here, we not only obtain a confidence interval for the date by assessing the degree of correlation among samples, but also describe how to use a Markov transition matrix approach to take full account of the complexities of the recombination process. Finally, we show how the method has been applied to a specific example: estimation of a date for the MRCA of a mutation that confers resistance to HIV-1 infection (Stephens et al. 1998).

Date estimates for MRCAs are possible because of the non-random association of alleles (i.e., linkage disequilibrium) that is generated whenever a new mutation occurs. The immediate descendents of a mutant chromosome will be monomorphic for a set of markers linked to the locus of interest. Over time, however, as recombination and mutation undo the linkage disequilibrium, the pattern of variation among mutant chromosomes will gradually reflect the pattern of variation in the population as a whole. By making a quantitative assessment of the extent to which the disequilibrium has been undone, and using known rates of mutation and recombination, we can estimate an age for the most recent common ancestor of mutant chromosomes. This is different from the date of occurrence of the mutation itself, which is generally older than the MRCA. The age of the mutation cannot be

estimated directly using linkage disequilibrium because linkage disequilibrium only begins to break down once there are multiple ancestors of the modern sample; that is, subsequent to the occurrence of the MRCA.

(B) Estimating the Age of a MRCA When Almost All Chromosomes Have the Ancestral Haplotype

To estimate the date of a MRCA when almost all mutant chromosomes are of a single type, we employ a two-pronged strategy. First, we assume that the common haplotype is the ancestral haplotype, a questionable assumption if the genealogical tree of relationships among individuals includes only a few ancient lineages, and in particular, if an early mutation or recombination event occurred on a lineage that was ancestral to the majority of current chromosomes. To determine the ancestral haplotype unequivocally, we use markers that are relatively close to the gene locus of interest. We then use the frequency (r) of mutation and recombination events that have the potential to unlink some chromosomes from the ancestral haplotype to find the most likely number of generations that have passed since the ancestral mutant chromosome.

To obtain the most likely date of the MRCA, we begin by considering a particular lineage of the genealogy, the chain of ancestors linking a present-day haplotype to the haplotype at coalescence. The probability that a haplotype remains ancestral during the time tracing back to the most recent common ancestor is given by the depth of the genealogy, G, and the frequency r of mutation and recombination

$$p = e^{-Gr} \tag{1}$$

Here, p is just the zero term in a Poisson series with parameter Gr.

To find p, we note that for a dramatically expanded population, one for which all lineages are essentially independent, an unbiased estimate of p is the proportion of observed haplotypes that are ancestral (Stephens et al. 1998). A surprising fact is that this statement is true even for constant-sized populations in which many lineages are highly correlated in the sense that pairs of alleles share extensive periods of co-ancestry during the time tracing back to the most recent common ancestor of the sample. The reason why the age estimate is independent of topology is that as long as mutations at the marker loci have no selective effect, the correlations in the tree amount to a process of pseudo-replication of lineages. This process will affect the variance of our estimate of p (see below); however, because the lineages that are replicated are selected independent of allelic state, the proportion of ancestral haplotypes will not be systematically affected.

Finally, to obtain G in terms of the estimate of p, we transform equation (1).

$$G = -\ln(p)/r \tag{2}$$

As discussed previously, this holds true whatever the shape of the genealogical tree.

(C) A Comprehensive Approach for Estimating the Most Recent Common Ancestor of a Mutation in a Sample

The previous method produces an appropriate estimate for the age of the MRCA when the large majority of observed chromosomes carry the ancestral haplotype. However, when enough mutant chromosomes have become unlinked from the ancestral haplotype, the date estimate must account not only for the rate of loss of the ancestral haplotype by mutation or recombination, but also for regeneration of the ancestral haplotype among chromosomes that have already lost it (Risch et al 1995). By including this process in our analysis, the estimated date of mutation becomes systematically older than predicted by Equation (2).

To provide a complete description for a system in which a single locus is typed, we use a Markov transition matrix **K**. Note that Risch et al (1995) have used an alternative approach to the same problem, involving differential equations. However, we have chosen to use the transition matrix approach instead because we find it to be very flexible, and because it allows us to easily incorporate mutation and recombination events into the same evolutionary process. Specifically, the entries in the Markov matrix give the probabilities, per generation, that any one haplotype will transform into any other. To calculate **K**, we then take a weighted sum of matrices corresponding to recombination (**R**), mutation (**M**), and no event occurring (**I**)

$$\mathbf{K} = c\mathbf{R} + \mu\mathbf{M} + (1\text{-}c\text{-}\mu)\mathbf{I} \quad (3)$$

where c is the frequency of recombination, μ is the frequency of mutation, and $1\text{-}c\text{-}\mu$ is the frequency of no event occurring. We now consider a single lineage tracing its ancestry back to the original mutation, and by multiplying **K** by the state vector generation by generation, evaluate the probability that after n generations, the mutation will have lost its linkage to the ancestral haplotype. This is exactly analogous to the method described in section (B), except here we take into account regeneration of the ancestral haplotype as well as the rate of loss of that haplotype.

Consider the case in which only a single microsatellite marker has been typed. For this case, the state of the system can be represented as a vector (q $1\text{-}q$), with the first entry the probability that the allele is of the ancestral type and the second the probability that it is not. The matrices **R** and **M**, and hence the Markov transition matrix, can then be derived straightforwardly from the distribution of alleles in non-mutant chromosomes. We begin with the recombination matrix (**R**). After a recombination event, the probability that the allele will end up ancestral, regardless of the initial state, can be estimated as the proportion of alleles in the population that have the ancestral haplotype (a). The probability that the allele will be non-ancestral type is $1\text{-}a$.

$$\mathbf{R} = \begin{bmatrix} a & a \\ 1-a & 1-a \end{bmatrix} \quad (4)$$

We now calculate the mutation matrix (**M**). According to the stepwise mutation model for microsatellites (Goldstein and Pollock, 1997), mutations change the length of an allele by a single unit, with an equal chance of increasing or decreasing the length of the allele. Using this model, we estimate the probability that a mutation will transform a non-ancestral allele into an ancestral one as $b/2$, where b is the proportion of alleles that are one mutation step away from the ancestral haplotype, and the division by 2 occurs because only half of mutations at these alleles produce the ancestral type. Note that in the case of a mutation that occurs on an ancestral allele, the outcome is even simpler: the probability that an allele will remain ancestral is 0.

$$\mathbf{M} = \begin{bmatrix} 0 & b/2 \\ 1 & 1-b/2 \end{bmatrix} \tag{5}$$

To find b in any generation, we require information that is not contained in the two-dimensional state vector: specifically, the frequencies of alleles that are one mutation step away from the ancestral chromosome. Thus, to describe the frequencies of all k possible alleles in the system, we require a k-dimensional state vector, which complicates matters because the **R** and **M** matrices would now have to be k x k rather than 2x2. Nevertheless, it is often possible to simplify the analysis when recombination occurs much more frequently than mutation. In this case, the distribution of non-ancestral alleles among mutant chromosomes is expected to be the same as in the control population, and b can be estimated directly from the proportions of alleles in the control population.

We now use equation (3), and the matrices **R** and **M**, to obtain the Markov transition matrix **K**. Errors in **K** could arise either from misestimation of c and μ (since information about these parameters is often inaccurate), or from errors in a and b that might occur due to inappropriate selection of control populations, a failure to type a sufficient number of chromosomes in the control population or changes in the proportions of alleles in the population over the course of recent history. Since none of these sources of error is taken into account in our method for estimating a date of mutation, experimenters should consider a range of possible values of c, μ, a and b, as a way of assessing how much variability in the estimate of the age of the mutation could arise from misestimation of parameters.

Under the assumption that **K** is correct, we can now consider a particular lineage of the genealogy—the chain of ancestors linking a present-day haplotype to the haplotype at coalescence—and use **K** to determine the probability that the lineage remains ancestral at any given generation. We begin with the vector representing the ancestral mutant chromosome, which has the form (1 0) where the first entry is the probability that the lineage has the ancestral type. To evaluate the fate of the lineage in every subsequent generation, we multiply **K** by the vector, using conventional matrix multiplication for a vertically oriented vector, until we obtain a probability of observing an ancestral haplotype that is closest to the observed proportion, p, of mutant chromosomes. The number of times that **K** has been multiplied tells us the number of generations that have passed since the ancestral mutant chromosome.

(D) Variance of the Age Estimate

The variance of the age estimate (unlike the age estimate itself) is systematically affected by a population's demographic history. The reason for this is that populations with different demographic histories have differently shaped genealogical trees. For example, in a population that has undergone a relatively recent and dramatic expansion, almost all lineages will trace their ancestry independently back to the time of the expansion, and the number of independent assessments of the age of the tree will be equal to the number of samples. For a constant-sized population, there will be high degree of shared ancestry among sampled chromosomes, as explained above, and the number of independent assessments of the age of the tree will be therefore much smaller than the number of sampled chromosomes. The relatively large number of age assessments in an expanding population means that the date estimate is more accurate.

To determine confidence intervals for the date, we use computer simulations based on a coalescent algorithm by R. Hudson (1991) to describe a wide variety of histories from constant population size to fast growth. The "population" in the simulation is actually the subpopulation of chromosomes that are mutant. Note that final population size and exponential growth rate are the variable parameters in our simulation. For each set of demographic parameters, the simulation generates a large number of genealogical trees and distributes mutation and recombination events along them according to a random (Poisson) process (we use the Markov transition matrix to determine which events turn an ancestral haplotype into a non-ancestral one and vice-

versa). Thus, the final distribution of haplotypes along a genealogical tree is affected by two sources of error: first, variability in the shapes of the genealogical tree, and second, variability in the mutation and recombination events that occur on those trees. The simulations allow us to take account of both these sources of error—generating a 95% central confidence interval for the number of ancestral haplotypes that could be expected to be seen in such a sample. We can then reject certain combinations of demographic parameters if the confidence intervals do not include the number of ancestral haplotypes that was actually observed.

To find allowed dates for the MRCA, we consider each combination of demographic parameters separately, simulating many genealogical trees and considering only those simulations that result in the observed number of ancestral haplotypes (i.e., we condition the simulations on the observed results). From the subset of trees obtained in this manner, we can then produce a 95% central confidence interval for the date of the MRCA. To obtain an allowed range of dates that is inclusive of all possible demographic histories, we then take the union of confidence intervals for each combination of parameters. The range of allowable dates can be constricted even further if we have additional information about the demographic history—for example, if the observed distribution pattern of non-ancestral haplotypes forbids particular combinations of demographic parameters, as explained in section (F) below.

(E) Age of the CCR5-Δ32 AIDS Resistance Allele

The CCR5 gene encodes a protein that serves as part of the primary entry port for HIV-1 in immune cells (Deng et al., 1996). Individuals homozygous for a particular 32 base pair deletion mutation in the gene, which we designate as CCR5-Δ32, are resistant to HIV-1 infection (Dean et al., 1996). Indeed, as many as 26% of northern Europeans carry at least one deleted copy of the gene (Stephens et al., 1998), while the frequency of carriers drops to zero along a north-south gradient (no copies are observed among Africans). The distribution pattern of the allele makes it seem likely that the mutation occurred recently, and it is therefore of interest to obtain a direct estimate for the date of the MRCA.

The data we use consist of 46 chromosomes carrying the CCR5-Δ32 deletion, and 146 controls that do not carry the mutation. Each chromosome was typed at two microsatellite markers on the same side of the CCR5 gene: GAAT12D11 (GAAT) and AFMB362wb9 (AFMB), with GAAT closest to the deletion locus. The ancestral haplotype is taken to be the one in which the GAAT marker carries the 197 base pair allele and the AFMB marker carries the 215 base pair allele. This haplotype occurs in 85% of mutant chromosomes, but only 36% of controls (Table 1).

Table 1: Distribution of Haplotype Frequencies for CCR5 Date Estimation

	Control Chromosomes	Mutant Chromosomes
$GAAT_{191}$, $AFMB_{215}$	3/146	
$GAAT_{191}$, $AFMB_{217}$	2/146	
$GAAT_{193}$, $AFMB_{215}$	20/146	2/46
$GAAT_{193}$, $AFMB_{217}$	21/146	
$GAAT_{197}$, $AFMB_{213}$		1/46
$GAAT_{197}$, $AFMB_{215}$	*** 53/146**	*** 39/46**
$GAAT_{197}$, $AFMB_{217}$	45/146	3/46
$GAAT_{197}$, $AFMB_{219}$	2/146	1/46

* indicates the putative ancestral type

To calculate the Markov transition matrix for this system, we note that two polymorphic markers were typed and hence there are a number of possible states in the system. However, we collapse the total number of allowed states into a smaller number of states that are more convenient to manipulate mathematically. Specifically, we consider the following three states: (1) both GAAT and AFMB are ancestral; (2) only GAAT is ancestral; and (3) neither GAAT nor AFMB is ancestral. The state of the system in any generation can be represented as $(q_1 \quad q_2 \quad 1-q_1-q_2)$, which should actually be represented as a vertical vector, and the transition matrices, corresponding to mutation at GAAT, mutation at AFMB, recombination at GAAT or recombination at AFMB, will be 3 dimensional (3x3) as well. The overall equation for the transition matrix **K** is then:

$$\mathbf{K} = \mu_{GAAT}\mathbf{M}_{\mathbf{GAAT}} + \mu_{AFMB}\mathbf{M}_{\mathbf{AFMB}} + c_{GAAT}\mathbf{R}_{\mathbf{GAAT}} + c_{AFMB}\mathbf{R}_{\mathbf{AFMB}} + (1\text{-}c_{GAAT}\text{-}c_{AFMB}\text{-}\mu_{GAAT}\text{-}\mu_{GAAT}) \quad (6)$$

where μ_{GAAT}, μ_{AFMB}, c_{GAAT} and c_{AFMB} the rates of mutation and recombination for the GAAT and AFMB markers, and $\mathbf{M}_{GAAT}$, $\mathbf{M}_{AFMB}$, $\mathbf{R}_{GAAT}$ and $\mathbf{R}_{AFMB}$ are mutation and recombination matrices.

We must now estimate the parameters μ_{GAAT}, μ_{AFMB}, c_{GAAT} and c_{AFMB}. To estimate the mutation rates, we use $\mu_{GAAT} = \mu_{AFMB} = 0.00053$, based on an estimated value for dinucleotide microsatellites (J. Weber, unpublished data). To obtain the recombination rates, we use $c_{GAAT} = 0.0021$ and $c_{AFMB} = 0.0072$, which are obtained by converting physical distances from radiation hybrid mapping to recombination distances using a linear regression that applies on the chromosome on which CCR5 is found (Stephens et al. 1998). In this analysis, error in estimation of the recombination rate was much more of a worry to us than error in the mutation rate, since the recombination rate is so much larger in absolute terms.

To obtain the mutation matrices, we use the frequencies of alleles in the control population that are one mutation step away from the ancestral GAAT (b_1) and ancestral AFMB (b_2) alleles (see equation (4)). It follows that for mutation at the GAAT marker, the matrix is $\mathbf{M}_{GAAT}$, while for mutation at the AFMB marker, the matrix is $\mathbf{M}_{AFMB}$.

$$\mathbf{M}_{GAAT} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & b_1/2 \\ 1 & 1 & 1-b_1/2 \end{bmatrix} \qquad \mathbf{M}_{AFMB} = \begin{bmatrix} 0 & b_2/2 & 0 \\ 1 & 0 & b_2/2 \\ 0 & 1-b_2/2 & 1-b_2/2 \end{bmatrix} \tag{7}$$

Table 1 implies that $b_1 = 0$ and $b_2 = 53/100 = 0.53$.

To obtain the recombination matrices, we follow expression (5), dealing first with the case in which the recombination occurs between the gene locus of interest and GAAT, and then the case in which the recombination event occurs between GAAT and AFMB. In the first case, the situation is exactly analogous to equation (5), and the frequencies of each possible outcome can be estimated as the proportion of alleles in the control population that are of each haplotypic state. We designate these frequencies, respectively, as a_1, a_2, and a_3, recalling that $a_3 = 1\text{-}a_1\text{-}a_2$. The resulting matrix is designated $\mathbf{R}_{\mathbf{GAAT}}$. In the second case, in which the recombination occurs between GAAT and AFMB, the alleles change at only a single locus (AFMB), and the only relevant parameters are the frequency of alleles for which the AFMB marker had the ancestral type (a_1), and the frequency of alleles for which the AFMB marker was non-ancestral (a_2+a_3). The overall 3x3 transition matrix, $\mathbf{R}_{\mathbf{AFMB}}$, then becomes:

$$\mathbf{R}_{\mathbf{GAAT}} = \begin{bmatrix} a_1 & a_1 & a_1 \\ a_2 & a_2 & a_2 \\ a_3 & a_3 & a_3 \end{bmatrix} \qquad \mathbf{R}_{\mathbf{AFMB}} = \begin{bmatrix} a_1 & a_1 & 0 \\ a_2 + a_3 & a_2 + a_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

Table 1 implies that $a_1 = 53/146$, $a_2 = 47/146$ and $a_3 = 46/146$.

We now use equation (6) to calculate **K**. Ignoring any error in the Markov transition matrix (more likely to be due to errors in estimation of the recombination rate and recombination parameters rather than errors in the mutation rate), the most likely age for the MRCA of the CCR5-Δ32 mutation is 31 generations, or 775 years assuming a generation time of 25 years. For comparison, if the calculation is done

according to the method of part (B), the estimate is 29 generations, slightly younger because no Markov transition matrix is used to take into account regeneration of the ancestral haplotype. Note that the estimated date is likely to be systematically younger than the date of first appearance of the mutation, since the estimation procedure only dates the most recent common ancestor of the sampled chromosomes. Thus, our estimate of the date must be interpreted with caution: in particular, if a dramatic expansion occurred in the population of mutant chromosomes, it is likely that the mutation occurred before the expansion, even though the MRCA may not date back earlier than the expansion. Note that Slatkin and Rannala (1997) provide an approach for dating mutations that takes this systematic bias into account, although their procedure depends on assumptions about a population's demography around the time of the mutation, which can rarely be known sufficiently well.

(F) Estimating a Variance for the Date by Reconstructing the Genealogy of CCR5-Δ32

To obtain a confidence interval for the date estimate, we use simulations that take into account all possible combinations of demographic parameters and genealogical trees, as described in section (D). To place further restrictions on the allowed dates of the mutation, we forbid certain genealogical trees—in the simplest case by using prior knowledge of population history. For the CCR5-Δ32 data, for example, we assume that during the past 10,000 years, northern European populations have had a certain minimum size. By specifying that the final effective population sizes of our sample was at least 5,000, we conclude that the date of the most recent

common ancestor was between 11 and 75 generations in the past (275–1875 years, assuming 25 years per generation) (Figure 1).

In a much more fundamental way, it is also possible to use the distribution of non-ancestral haplotypes among mutant chromosomes to put restrictions on the shape of the genealogical tree. For example, if the haplotypes all derive from separate mutation or recombination events, the lineages of the genealogical tree are uncorrelated and consistent with a dramatically expanded population. If the non-ancestral haplotypes derive from relatively few mutation or recombination events (which have been recopied and amplified within the lower branches of the genealogical tree), then the history of the mutant chromosomes is more likely to be consistent with a constant-sized population. By focusing on the distribution of non-ancestral haplotypes among CCR5-Δ32 chromosomes, we are then able to directly assess the degree of correlation in the tree, and from there to assess the variability of the date estimate.

To implement this approach, we consider the fact that of the seven non-ancestral CCR5-Δ32 chromosomes that were observed, there were four distinct haplotypes. The number of mutation and recombination events that actually gave rise to the four haplotypes was probably larger than four, since the distribution of non-mutant CCR5 chromosomes indicates that given six or seven chances, several haplotypes would be generated more than once (and, as expected from this hypothesis, the non-ancestral haplotypes we observe are the ones that are most frequent in the control population). We surmise that the non-ancestral haplotypes derive for the most

part from separate mutation and recombination events, and that in the present sample, we are observing the results of at least six and perhaps seven different events. Note that it would have been possible to determine the number of events with even more precision if more than two microsatellite markers had been typed.

To make explicit use of this information, we modify the simulation described in section (D) to report not only the number of non-ancestral haplotypes but also the number of distinct mutational and recombinational events that gave rise to these haplotypes. Thus, for each set of demographic parameters in the CCR5-Δ32 data set, we simulate a large number of genealogical trees that gave rise to 7 out of 46 non-ancestral haplotypes and then determine the proportion of these replicates that were derived from seven distinct events. If we require that no fewer than 5% of replicates have fewer than 7 distinct haplotypes, we can restrict the date of the mutation to between 9 and 214 generations in the past (225-5,350 years, assuming 25 years per generation). While this restriction on the date of the mutation is less stringent than the one derived from a historical assumption about effective population sizes, it is valuable because it is independent of such assumptions.

(G) The Analysis of New Data Sets

In applying this method to a new data set, it is always appropriate to begin by picking microsatellite markers that have the proper distance from the gene locus of interest. The markers should be chosen to be close enough to the locus of interest to define the ancestral haplotype, but far enough away to allow as many lineages as

possible to have had a chance to become non-ancestral. A good strategy for identifying markers is to select a panel that are at varying distances from the gene locus of interest, and then to pick out ones that comply with the criteria described above.

The analysis of data from a single microsatellite locus can often extract most of the relevant information about the date of a mutation. However, the use of multiple markers (e.g., in the CCR5-Δ32 experiment) may have a particular value in assessing the variance of the date estimate, allowing for a better assessment of the shape of a genealogical tree than would be possible with a single marker. The reason for this is that multiple markers allow us to more accurately reconstruct the history of mutation and recombination events. If even more markers are typed, it becomes possible to pinpoint the exact number of distinct mutation and recombination events that had led to the observed number of non-ancestral haplotypes, further restricting the allowed range of genealogical trees. On the other hand, multiple markers have a drawback because they can make an analysis more complicated, forcing the estimation of more matrix parameters, recombination distances and mutation rates.

Finally, another factor to consider in designing future experiments is that some mutations will be sufficiently old in a population that only markers close to the locus will display disequilibrium. In this case, it will be difficult to determine the recombination distances of markers from the locus, and it is appropriate to use markers that are sufficiently close to the gene that mutation serves as the main molecular clock for estimating a date for the MRCA. Errors in estimating the mutation

rate (and not the recombination rate) then become the main source of systematic error in determining the age of the MRCA, and to reduce this error, it is appropriate to use several markers that are close to the gene locus of interest, with an average mutation rate that in general will be more predictable than that of a single marker (Goldstein and Pollock, 1997). In practice, however, it may be difficult to find enough markers that are sufficiently close to the gene locus of interest to make this possible, except perhaps on the Y-chromosome where a large number of microsatellites are completely linked.

REFERENCES

Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, et al. (1996) Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science* **273**, 1856-1862.

Deng HK, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, DiMarzio P, Marmon S, Sutton RE, Hill CM, Davis CB, Peiper SC, Schall TJ, Litman DR, Landau NR (1996) Identification of a major coreceptor for primary isolates of HIV-1. *Nature* **381**, 661-666.

Goldstein DB, Pollock DD (1997) Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. *The Journal of Heredity* **88**, 335-342.

Hudson RR (1991) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) **7,** 1-44 (Oxford University Press, Oxford).

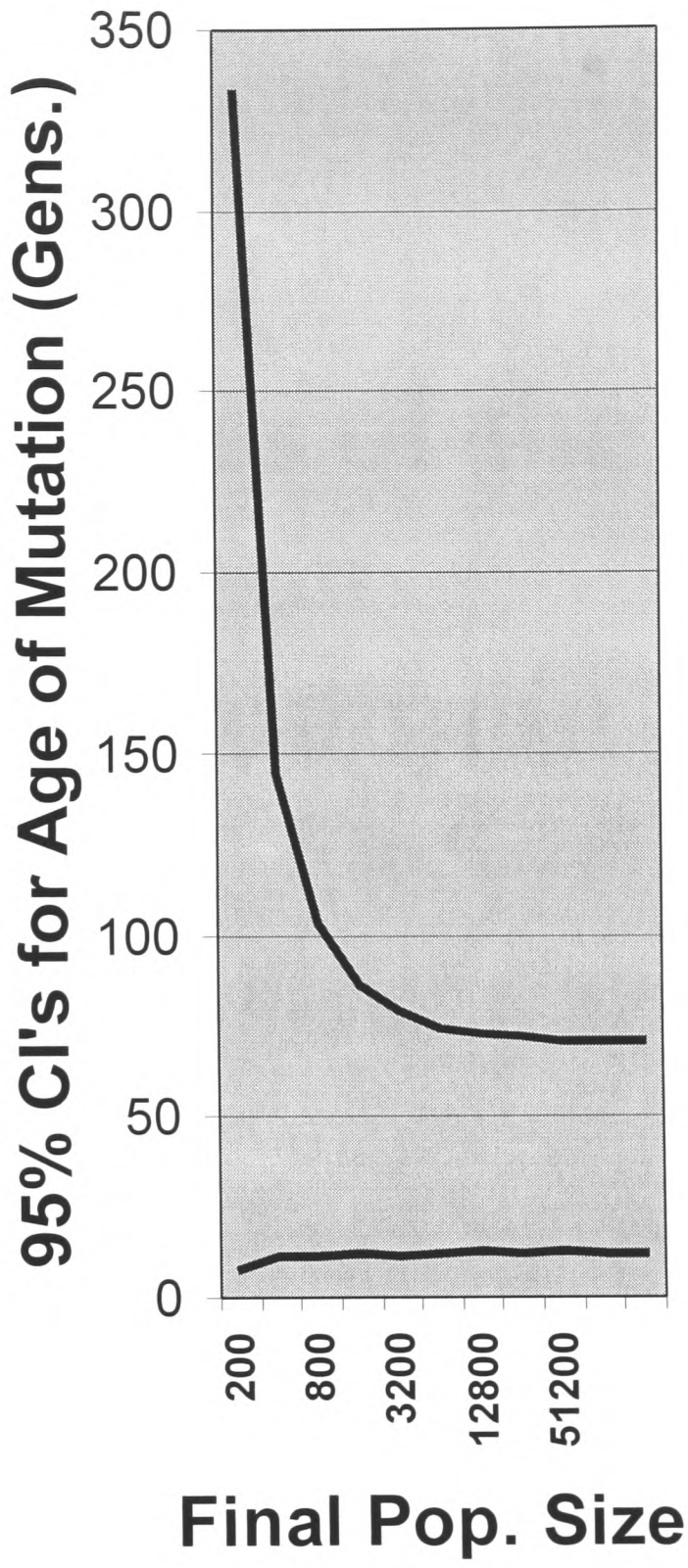
Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Fahn S, Breakefield X, Bressman S (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nature Genetics,* **9,** 153-159.

Slatkin M, Rannala B (1997) Estimating the age of alleles by use of intrallelic variability. *American Journal of Human Genetics,* **60**, 447-458.

Stephens JC, Reich DE, Goldstein DB, Doo Shin H, Smith MW, et al. (1998) Dating the origin of the CCR5-Δ32 AIDS resistance gene allele by the coalescence of haplotypes. *American Journal of Human Genetics,* **62,** 1507-1515.

FIGURE 1

Final Population Size	**95% CI's for Age of Mutation (Generations)** Low	High
200	**8**	**332**
400	**11**	**144**
800	**11**	**103**
1600	**12**	**86**
3200	**11**	**79**
6400	**12**	**74**
12800	**13**	**73**
25600	**12**	**72**
51200	**13**	**71**
102400	**12**	**71**
204800	**12**	**71**



95% confidence intervals for the age of the most recent common ancestor of the CCR5-Δ32 mutation, in tabular as well as graphical form. For each final population size, a range of exponential growth models are considered, ranging from extremely slow to extremely fast growth. For large final population sizes (greater than 100,000), the confidence interval converges to 12-75 generations, which corresponds to the limiting case of a starlike genealogy.

Estimating the ages of the most recent common ancestors of two common mutations causing Factor XI deficiency in Ashkenazi and Iraqi Jews

David B. Goldstein and David E. Reich

Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK.

Introduction

Factor XI blood-coagulation deficiency occurs in unusually high frequency in Jewish populations (Seligsohn 1978; Saito et al. 1985; Bolton-Maggs et al. 1992), and most cases are due either to the type III or to the type II mutations, which refer to two different unique mutations on the *FXI* gene. The type III mutation occurs with a frequency of 2.54% in Ashkenazi Jews, but has not been observed in other Jewish populations. The type II mutation is present in both Ashkenazi (2.17%) and Iraqi (1.67%) Jews (Shpilberg et al. 1995; Peretz et al. 1997). The high frequency of type II in both groups indicates a genetic relationship between the groups, which is interesting in light of the fact that they are believed by some authors to have diverged more than 2,500 years ago, at the time of the Babylonian exile (Peretz et al. 1997). To explore these issues further, we estimated dates for the most recent common ancestors of the type II and type III mutations.

Data and Analysis

When a new mutation appears, it occurs on a single chromosome and thereby generates linkage disequilibrium with any linked polymorphic markers. The rate at which this disequilibrium breaks down depends on the recombination and mutation rates; hence, if estimates of these are available, the observed level of disequilibrium can be used to estimate the date of the most recent common ancestor of mutant chromosomes in the sample (Risch et al. 1995; Stephens et al. 1998). To achieve statistical power, however, it is important to use a marker at an appropriate

recombinational distance—that is, great enough to ensure the generation of nonancestral haplotypes at moderate frequency but, ideally, not so great as to equilibrate allele frequencies within the time frame of interest. Of 74 informative meioses, we observed a single recombination event between *FXI* and D4S171 (a dinucleotide microsatellite), implying a recombination rate of .0135, appropriate for estimating coalescent times on the order of hundreds of generations. Frequencies of alleles at this marker were determined in 99 chromosomes carrying the type II mutation and in 73 chromosomes carrying the type III mutation. We also characterized 103 chromosomes from healthy Ashkenazi Jews, to estimate the proportion of recombination events that are expected to result in a nonancestral haplotype (Table 1).

In their study of idiopathic torsion dystonia, Risch et al. (1995) used a method for estimating the coalescent time (G) of mutant chromosomes that focuses on the proportion of lineages not having undergone a mutation or recombination event. If no regeneration of ancestral haplotypes is assumed, the expected proportion of ancestral haplotypes is effectively $p = e^{-Gr}$, where G is the number of generations since the coalescence of the sampled chromosomes and r is the effective mutation and recombination rate (see eq. [1]). When there is a moderate to high proportion of nonancestral alleles among the mutant chromosomes, however, it is necessary to model the regeneration of ancestral haplotypes by the recombination process. This can be done in a number of ways, but we favor the Markov model representation of Reich and Goldstein (1999) for its flexibility. Under this formulation, it is straightforward to include an arbitrary number of markers and both recombination and mutation.

The full behavior of the system is easily represented as a Markov process in which the state space is the proportion of chromosomes in the ancestral and

nonancestral categories, and the transition matrix **K** gives the probabilities that each haplotype will be transformed into the other in a single generation (Stephens et al. 1998; Reich and Goldstein 1999). The transition matrix **K** is given by:

$$\mathbf{K} = c\mathbf{R} + \mu\mathbf{M} + (1 - c - \mu)\mathbf{I} \qquad (1)$$

where c and μ are scalars reflecting the recombination and mutation rates, respectively. In this case, $\mu = .00056$ (Weber and Wong 1993) and $c = .0135$. The matrices **R** and **M** reflect the probabilities of producing nonancestral haplotypes should a recombination or mutation event occur, whereas **I** is the identity matrix. If a represents the frequency of the ancestral allele in the control population, the matrix **R** has the elements $R_{11} = a$, $R_{12} = a$, $R_{21} = 1 - a$, and $R_{22} = 1 - a$. Formally, **M** would depend on the frequencies of marker alleles on mutant chromosomes, requiring a larger state space than ancestral/nonancestral alleles. In our case, however, because the recombination rate is much greater than the mutation rate, we assume that the distribution of allele sizes on mutant chromosomes matches that seen in the control population. For simplicity, we also assume a strict stepwise mutation model (Goldstein and Pollock 1997). Under these assumptions, **M** has the elements $M_{11} = 0$, $M_{12} = b/2$, $M_{21} = 1$, and $M_{22} = 1 - b/2$, where b is the frequency of all one-mutant neighbors of the ancestral allele in the control population.

With the parameters of **K** specified, the coalescent time is estimated by multiplying the state vector by **K** iteratively until the observed proportion of ancestral haplotypes is reached. Iteration begins at a frequency vector of (1,0), corresponding to a starting point of only ancestral haplotypes. The analysis requires identification of the

ancestral allele, which cannot be determined with certainty when the frequency of the nonancestral type is moderately high, as is the case for both the type II and type III mutations. For the type II mutation, the data are consistent with the ancestral haplotype having carried either the 151- or 153-bp allele, which is consistent with the high frequency of these alleles in the control population. If the 151-bp allele is assumed to have been ancestral, the proportion of ancestral haplotypes among mutant chromosomes (p) is .48, and the frequency of the ancestral allele in the control population (a) is .38, leading to an estimated coalescent time of 120 generations. If the 153-bp allele is assumed to have been ancestral, p = .27 and a = .22, and the estimated coalescent time is 189 generations.

The analysis of the type III mutation is complicated by the bimodal frequency distribution among the mutant chromosomes of an allele (145 bp) that is in low frequency in the control populations. In theory, this discrepancy could be caused by any of four factors: (1) multiple origins for the type III mutation; (2) frequencies of the allele in the control population being nonrepresentative of those in the population in which the mutant chromosomes have been recombining; (3) the mutation having occurred originally on the 145-bp allele, despite its rarity in the population; or (4) the occurrence, very early after the appearance of the mutation, of a recombination event between the chromosomes carrying 145- and 151-bp alleles or, equivalently, of a multistep mutation between these alleles. The first explanation is ruled out by the observation of Peretz et al. (1997) that the type III mutation is associated with a single, rare haplotype defined by four closely linked markers. The second explanation also appears unlikely because of the similarity of marker allele frequencies in the Ashkenazi control chromosomes and the CEPH chromosomes (Table 1). Although

distinguishing the latter two explanations would require more detailed information about the genealogy than is available, we favor the fourth explanation, because it is consistent with the substantially larger differential between the 151- and 153-bp alleles among mutant chromosomes than is observed in the control population. If we accept this explanation, we need not be concerned with whether the mutation actually occurred on a chromosome carrying a 145- or 151-bp allele, but rather we can treat both alleles collectively as ancestral, yielding $p = .78$ and $a = .41$ and leading to an estimated coalescent time of 31 generations. Of course, the coalescent time for the true ancestral haplotype must be longer than this estimate. If we treat the mutation as having occurred singly on either the 145- or 151-bp allele, the estimated age of the mutation would be >100 generations.

Although the expected coalescent times are independent of the shape of the genealogy, confidence intervals are strongly dependent on the precise shape. One approach to the estimation of confidence intervals is to carry out coalescent simulations assuming a range of population growth rates, resulting in gene genealogies ranging from the highly correlated trees typical of constant population size to the star-shaped genealogies typical of very rapid growth. When the data provide information on the number of recombination and mutation events responsible for the observed nonancestral haplotypes, it is possible to select, at least roughly, among these types of genealogies to construct confidence intervals (Reich and Goldstein 1999). Our data set, however, provides relatively little information about the shape of the genealogy, because of the combination of a single marker locus and high frequencies of nonancestral haplotypes. We therefore use a simple heuristic argument to provide a rough guide to the degree of confidence in these estimates, under different

assumptions concerning the gene genealogy. For convenience, we shall use the term "confidence interval," but it is important to appreciate that these calculations are meant to be illustrative of how genealogical shape influences confidence and cannot be construed as formal confidence intervals.

First, we assume a star genealogy in which all lineages are uncorrelated—that is, lineages trace their ancestries independently back to the root of the genealogy. In this case, where n is the number of sampled chromosomes and p is the proportion of ancestral chromosomes, there are n independent observations of the time that has elapsed since the appearance of the common ancestor of the sample. In the case of the type II mutation, $n = 99$, $p = .48$, and the confidence interval for the number of ancestral chromosomes observed can be estimated as $\pm 2\times$ the SD of a binomial distribution with parameters .48 and 99, leading to a confidence interval for p of .38-.58. When we apply the recursion in equation (1) to this range, the confidence interval becomes 75-254 generations. A similar calculation for the case of the 153-bp allele having been ancestral leads to a lower bound on the age of the mutation of ~120 generations. An upper bound does not exist in this case, because the lower bound on p is below the equilibrium frequency set by the observed frequency of the ancestral allele in the control population.

The conceptual basis of our approach for assessing confidence intervals in a correlated tree depends on estimating an effective number of lineages that would result in an uncorrelated tree with properties similar to those of a correlated tree with a greater number of sampled chromosomes; that is, we assess the extent to which correlations between lineages in the tree reduce the number of independent observations of the time to the common ancestor. To estimate the effective number of

lineages, we determine how many independent lineages would be required to produce a match between the mutant and control chromosome marker allele frequencies at least as close as was actually observed. For example, in the case of the type II mutation, <14 independent lineages would have a <5% chance of leading to a match that is as close as was actually observed.

To better understand the argument, imagine a genealogy in which every chromosome is included within a set of 10 exact copies, but each of these sets traces its ancestry independently to the time of origin. In this case, the number of independent lineages would be 10; however, we would not expect to have an allele-frequency distribution that matches the control distribution as closely as does the observed distribution. That point is reached only when there are at least 14 independent lineages. Therefore, if we take 14 as the lower bound on the "effective number" of lineages, the youngest possible ages for the type II mutation become 34 and 72 generations for the mutation having occurred on chromosomes carrying the 151- and 153-bp alleles, respectively. In the case of 14 lineages, there is no upper bound on the coalescent time, because the lower bound on p is below the equilibrium value. Applying a similar logic to the type III mutation, we obtain 23 effective lineages, and if we treat both the 145- and 151-bp alleles as ancestral, the estimated confidence interval is 5-70 generations.

This analysis considers uncertainty resulting from the evolutionary process but ignores uncertainty in our estimation of the mutation and recombination rates. Finally, we note that the approach taken here estimates the coalescent time of affected chromosomes, as opposed to the time at which the mutation first appeared, which must predate the coalescent time. Accurate estimation of the date of the mutation

would require detailed information about the population's demography at the time of the mutation (see Slatkin and Rannala 1997).

Discussion

The date estimate for the Factor XI type II mutation is consistent with the hypothesis that Ashkenazi and Iraqi Jews share a common genetic origin, but did not exchange genes subsequent to the Babylonian exile ~2,500 years ago. The date estimate does not support the occurrence of a more recent bottleneck in Ashkenazis, even though such a hypothesis has been used by many authors to explain the high frequency of monogenic diseases in this group (Goodman 1978; Risch et al. 1995). The type II mutation thus provides the first example of a monogenic disease in high frequency in Ashkenazis that cannot be ascribed to a medieval bottleneck.

References

Bolton-Maggs PHB, Wensley LJ, Tuddenbham EDG (1992) Genetic analysis of 27 kindreds with factor XI deficiency from north west England. Paper presented at 24th Congress of the International Society of Hematology, London, August 23-27, abstract 511a.

Goldstein DB, Pollock DD (1997) Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J Hered 88:335-342.

Goodman RM (1978) Genetic disorders among the Jewish people. Johns Hopkins University Press, Baltimore.

Peretz H, Mulai A, Usher S, Zivelin A, Segal A, Weisman Z, Mittelman M, et al (1997) The two common mutations causing factor XI deficiency in Jews stem

from distinct founders: one of ancient Middle Eastern origin and another of more recent European origin. Blood 90:2654-2659.

Reich DE, Goldstein DB (1999) Estimating the age of mutations using the variation at linked markers. In: Goldstein DB, Schlötterer C (eds) *Microsatellites: evolution and applications*. Oxford University Press, Oxford.

Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, et al (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat Genet 9:152-159.

Saito H, Ratnoff OD, Bouma BN, Seligsohn U (1985) Failure to detect variant (CRM+) plasma thromboplastin antecedent (factor XI) molecules in hereditary plasma thromboplastin antecedent deficiency: a study of 125 patients of several ethnic backgrounds. J Lab Clin Med 106:718-722.

Seligsohn U (1978) High gene frequency of factor XI (PTA) deficiency in Ashkenazi Jews. Blood 51:1223.

Shpilberg O, Peretz H, Zivelin A, Yatuv R, Chetrit A, Kulka T, Stern C, et al (1995) One of the two common mutations causing factor X1 deficiency in Ashkenazi Jews (type II) is also prevalent in Iraqi Jews, who represent the ancient gene pool of Jews. Blood 85:429-432.

Slatkin M, Rannala B (1997) Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60:447-458.

Stephens JC, Reich DE, Goldstein DB, Doo Shin H, Smith MW, Carrington M, Winkler C, et al (1998) Dating the origin of the *CCR5-Δ32* AIDS resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507-1515.

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123-1128.

TABLE 1

Allele Size (b.p.)	FREQUENCY			
	CEPH*	**Control (n = 130)**	**Type II (n = 99)**	**Type III (n = 73)**
143	.01	.02	.00	.00
145	.01	.03	.03	.26
147	.09	.06	.02	.05
149	.04	.08	.02	.01
151	.34	.38	.48	.52
153	.39	.22	.27	.10
155	.08	.12	.06	.01
157	.02	.00	.02	.01
159	.00	.05	.07	.01
161	.01	.00	.01	.00
163	.00	.02	.01	.00
165	.00	.03	.00	.01

*** Data provided by James Weber**

Chapter 3

Connections between evolutionary history and medical genetics

Part A: "Detecting associations in a case-control study in the face of population stratification" was submitted to *The American Journal of Human Genetics* on October 15, 1999.

Part B: "Correlation of genetic associations across populations: preliminary investigations" carries out a simple simulation study and lays a blueprint for future work.

Detecting association in a case-control study in the face of population stratification

David E. Reich[1,2] & David B. Goldstein[3]

[1] Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK. [2] Whitehead Institute/MIT Center for Genome Research, One Kendall Square, Cambridge, Massachusetts 02139, USA. [3] The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London, NW1 2HE, UK. Correspondence should be addressed to D.E.R. (e-mail: dreich@mit.edu; tel: 617-252-1900; fax: 617-252-1902)

Summary

Case-control studies are susceptible to the problem of "population stratification," which occurs in ethnically mixed patient populations and has the potential to produce false-positive signals of genetic association. However, by taking into account information from unlinked markers in the same patients in whom the original association was detected, empirical significance cutoffs can be derived that distinguish true associations from spurious ones.

Introduction

Case-control association studies detect non-random association between an allele and a trait, and can be powerful tools for gene mapping (Risch and Merikangas 1996). However, when a patient population is ethnically mixed, or is derived from admixture during the past few generations, non-random associations can occur even at markers completely unlinked to a disease locus (Chakraborty and Weiss 1988; Lander and Schork 1994). For example, in a patient population that is a mixture of African Americans and Caucasians, cases of hypertension will occur disproportionately among African Americans, who have a higher incidence of this disease (Kaplan 1994). Any alleles that occur more commonly in African Americans will therefore tend to be associated to disease, even if they are completely unlinked to disease-causing loci.

The transmission disequilibrium test (TDT) and related methods (SDT, sib-TDT) use control chromosomes from the families of affected individuals to circumvent the problem of population stratification (Ewens and Spielman 1995; Horvath and Laird 1998; Spielman and Ewens 1998). However, since it is not always possible to collect DNA from family members, it

would be useful to have a method to modify case-control studies so that they can be used in the presence of stratification. This is particularly important for dealing with so-called cryptic stratification, which, by definition, cannot be eliminated by asking subjects about their ethnic background.

By utilizing unlinked markers, it should be possible to adjust for population stratification without resorting to family-based controls. An empirical distribution of association statistics can be built up at these markers that can be used to estimate the strength of the evidence for linkage. In order to be useful, however, the unlinked markers must satisfy two criteria. First, they must recombine freely with the candidate polymorphism, and second, they cannot be in linkage disequilibrium with each other in the ethnically homogenous populations that contributed to stratification. Finally, they must not be linked to any causal variants.

Pritchard and Rosenberg (1999) showed that by genotyping cases and controls at fewer than 30 unlinked biallelic markers, the probability of first detecting a nominally significant association at the candidate allele, but failing to detect population stratification, can be reduced to less than about 5%. Here, we show how to use the same type of information to go beyond the detection of stratification—to determine whether the associations observed at candidate alleles are more extreme than what could be produced by stratification alone.

Terminology and Simulation Framework

To analyze this, we use the terminology and simulation framework described by Pritchard and Rosenberg (1999). Two subpopulations are assumed to have contributed to stratification, and the

disease incidence in subpopulation 1 is p_1 while the disease incidence in subpopulation 2 is p_2. Stratification increases along with an increasing difference between p_1 and p_2, and can be conveniently described in terms of two parameters: the probability f that a diseased individual comes from subpopulation 1, and the probability g that a healthy individual comes from subpopulation 1. The equation relating these variables is $f(1-g)/(g(1-f)) = p_1(1-p_2)/(p_2(1-p_1))$; the right half of this equation is approximately the same as the familiar relative risk (p_1/p_2) in the case of a rare disease (Pritchard and Rosenberg 1999).

The f and g parameterization is useful because it allows the difference in disease incidence between the two subpopulations to be treated separately from the allele frequency differences across loci. The difference in disease incidence (p_1 *vs.* p_2) is fixed for a given sample of cases and controls, but the allele frequency differences vary across loci according to a distribution that is determined by the vagaries of demographic history. To explore the behavior of this distribution, we use a coalescent simulation that reconstructs the relationships between the modern samples in a way that is consistent with the assumed demographies (Hudson 1991; Reich *et al.* 1999). An ancestral population of constant size N is assumed to have split into the two subpopulations, both of size N, at a time $\tau = t/2N$ in the past (t is in generations). The division of samples between subpopulation 1 and subpopulation 2 is determined by taking a fraction f of the cases and a fraction g of the controls from subpopulation 1 (both rounded to the nearest even integer). Mutations are then placed on the gene trees at a rate of $0.2/(4N)$ per generation, which is on the order of 200 times the mutation rate per base pair in humans. One way to think about this is that 200 base pair DNA fragments were screened in order to obtain the studied markers—a somewhat arbitrary choice, but our simulations turn out to be insensitive to the choice of

mutation rate. The simulations are repeated until the frequency of the less common allele in the combined sample of cases and controls is within a specified range. Note that the "allele frequency" is defined here as the frequency in the sample, not the population frequency.

True and False-Positive Associations

It should be possible to distinguish true from false-positive associations even in the presence of population stratification. For this purpose, we use the standard χ^2 association statistic, which can be calculated from contingency Table 1 as:

$$\chi^2 = \left(\frac{n_{11}}{n_{1*}} - \frac{n_{21}}{n_{2*}} \right)^2 n_{**} \frac{n_{1*} n_{2*}}{n_{*1} n_{*2}}. \quad (1)$$

To illustrate how a signal of linkage can be detected in the presence of stratification, we simulated an example involving 100 cases and 100 controls, a candidate allele frequency around 25%, $g = 0.1$, $f = .31$, and two subpopulations that diverged at a time $\tau = 0.25$ in the past (Fig. 1). Population stratification dramatically elevates the proportion of false-positives: fully 27% of unlinked markers show a nominally significant association at the $P < 0.05$ level ($\chi^2 > 3.84$) (Fig. 1). On the other hand, if we assume a particular disease model for causal alleles, the expected χ^2 value at the locus is easily distinguished from what is expected at unlinked markers, even though both are affected by stratification (Fig. 1). For alleles that are not themselves causal for disease, but are only linked to causal loci, such a distinction should still be possible due to the presence of linkage disequilibrium inherited from the subpopulations that were mixed.

To apply the empirical test for statistical association, the most straightforward approach is to ask whether the candidate allele shows a χ^2 statistic that is outside the range of what is observed at unlinked markers. To simplify matters, we assume that the allele frequencies for the unlinked markers are the same as for the candidate polymorphism. Under the null hypothesis that the candidate polymorphism is not associated to disease, the χ^2 value should be drawn from exactly the same distribution as the χ^2 vales at the unlinked markers. Hence, by rank-ordering the χ^2 statistics, and assessing where the candidate allele χ^2 statistic falls within this rank-ordering, a *P*-value can be obtained. If j of the m markers have χ^2 values higher than the candidate allele, the significance level determined by rank-ordering is $P < (j+1)/(m+1)$ (e.g., $P < 0.05$ if $j = 0$ and $m = 19$). Unfortunately, if higher degrees of precision are required, the method is impractical due to a requirement for a very large number of unlinked markers. To establish a significance cutoff of $P < 0.005$, for example, m must be at least 199.

Parametric Approach

It is also possible to establish significance levels with much less genotyping than is necessary for the rank-ordering method. A parametric approach uses the actual values of χ^2 values at unlinked markers, instead of counting the number above and below the candidate χ^2 statistic. To explore this approach, we performed computer simulations for a wide range of population stratification scenarios, and discovered a useful relationship between the empirical k^{th} percentile cutoff in a stratified population, c_k, and the mean χ^2 value, μ (Fig. 2). The ratio, c_k/μ, is nearly constant, regardless of k or the number of cases and controls. In other words, the empirical χ^2 distribution

is just scaled by a constant factor μ compared to what is expected for an unstratified population (however, it maintains the same overall shape). The linear relationship between the mean and cutoffs can be summarized by the following equation:

$$c_k = \chi^2_{1,k} \cdot \mu \,, \qquad (2)$$

where $\chi^2_{1,k}$ is just the k^{th} percentile cutoff in the absence of stratification, which can be obtained from any standard statistical table. For an unstratified population $\mu = 1$.

Equation 2 models the simulated data remarkably well, except that the linear fit begins to break down for high levels of stratification, and in particular, μ tends to conservatively overestimate the significance cutoffs (Fig. 2). To explain this, we note that the χ^2 distribution is defined to describe the probability distribution for the variance of a standardized normal distribution (Rice 1995). When sample sizes are sufficiently large, the difference statistic n_{11}/n_{1*}-n_{21}/n_{2*} is distributed normally, and the squared difference is distributed as the variance of a normal distribution, which explains why the χ^2 distribution governs the behavior of expression 1. The difference statistic n_{11}/n_{1*}-n_{21}/n_{2*} appears to continue to be distributed normally in the face of population stratification, but the variance increases by a factor of μ, and the effect is to scale the χ^2 distribution by μ while preserving its overall shape. Indeed, if we generate random contingency tables according to this framework, we replicate the linear scaling of $c_{.95}$ with μ, which increases our confidence that this intuitive explanation is correct. The same line of

reasoning also explains the non-linear behavior: for extreme stratification, the variance can increase to such an extent that the n_{ij} values in some squares of Table 1 reach their minimal values and become non-normal. Hence, they lose their simple scaling properties.

Implementation

The statistic μ is a full measure of stratification in the population, since it determines the scaling of the empirical χ^2 distribution. To find the highest μ consistent with the data, which can be plugged in to Equation 2 to obtain a significance cutoff, an efficient approach is to identify the largest μ consistent with the observed mean χ^2 value across unlinked loci ($\bar{\chi}^2$). Equation 2 indicates that the shape of the empirical χ^2 distribution, as determined by the relative positions of all the percentiles in the distribution, is the same regardless of the level of stratification. Hence, the rate of convergence of $\bar{\chi}^2$ to μ, as increasing numbers of markers are typed, is stratification-independent. It is therefore valid to use the analytically known form of the χ^2 distribution in the absence of stratification to carry out a numerical study of this convergence. Table 2 gives the maximum factor by which the true μ could be in excess of $\bar{\chi}^2$, at the 95% confidence level, for a specified number of genotyped markers. To ensure that μ is within a factor of 1.5 of $\bar{\chi}^2$, for example, it is necessary to genotype at least 26 unlinked markers. Suppose that 5 unlinked markers have been genotyped from Distribution 1 in Fig. 1. If $\bar{\chi}^2$ for these markers is 4.0, the maximum μ consistent with the data is $4 \times 2.21 = 8.8$ (Table 2). The 95th percentile cutoff can

then be extrapolated conservatively to $3.84 \times 8.8 = 34$, which gives about 80% power to detect the scenario of disease association depicted in Fig. 1.

A final caveat is in order. Equation 2 and other inferences made in this paper are based on computer simulations. While we tried to investigate a very broad range of stratification scenarios in order to make generalizations, we could not investigate every possibility. In particular, we did not consider relative risks greater than 8 and τ values greater than 0.4, which have been proposed as reasonable upper bounds for stratification (Pritchard and Rosenberg 1999). We also did not discuss the mixing of 3 or more subpopulations, or complications of demographic history. Hence, we can only guarantee the guidelines given in this paper for a subset of parameter space, but the simple explanation for the linear scaling in Fig. 1 suggest that that they are more general.

The Issue of Frequency Matching

Allele frequencies at unlinked markers need to be exactly matched to that of the candidate polymorphism in order for the χ^2 values to be strictly comparable. In practice, however, it is not always necessary to match allele frequencies exactly. To investigate how much constraint this imposes on the markers useful for analysis, we carried out a wide-ranging simulation study.

Figs. 3 and 4 describe the behavior of the 95th percentile cutoffs as a function of allele frequency, for a range of demographic scenarios. The 5-50% frequency range is covered by Fig. 3, and the 1-10% frequency range is covered by Fig. 4. Several important behaviors can be deduced from the figures. First, for low levels of stratification (as measured by the average value of the 95th percentile cutoffs), the traces in Fig. 3 are generally flat, and therefore independent of

allele frequency for allele frequencies greater than 15%. Second, regardless of stratification, accurate significance cutoffs can usually be obtained by restricting the marker allele frequencies to a 5-10% frequency window around that of the candidate allele. Finally, for allele frequencies less than 10%, frequency matching unfortunately needs to be more stringent, since significance cutoffs change relatively quickly as a function of frequency in this range (Fig. 4).

The importance of frequency matching for a particular case-control study must be assessed on the basis of the overall level of stratification. The appropriate way to do this is to estimate μ, and then to multiply by 3.84 ($\chi^2_{1,.95}$) to see where the stratified sample falls in terms of Figs. 3 and 4. However, since Fig. 3 and 4 are generated for arbitrary numbers of cases and controls, it is necessary to make a further adjustment for sample size. Designating n_g as the total number of chromosomes in the simulations (400 for Fig. 3 and 1000 for Fig. 4), n_{**} as the total number of chromosomes in the study, and assuming that the number of cases is equal to the number of controls, the following transformation is remarkably accurate:

$$c_{.95}' = \frac{n_{**}}{n_g}\left(c_{.95} - 3.84\right) + 3.84, \qquad (3)$$

In other words, the degree to which the significance cutoff for a given stratification scenario exceeds the unstratified case scales linearly with sample size. This can be understood, intuitively, by noting that the χ^2 distribution is the distribution of a variance (the same argument that was discussed above), and variances increase linearly with sample size.

Estimation of μ, combined with Equations 2 and 3, thus provides a guideline for deciding how stringent frequency matching needs to be. For example, if the scenario of stratification is as depicted in Fig. 1, and $\bar{\chi}^2$ is 3.2, the estimated 95th percentile significance cutoff is $c_{.95} = 3.2 \times 3.84 = 12$. This is in the range of traces that are nearly although not completely flat (Fig. 3). In such a circumstance, frequency matching should not be an important concern as long as the allele frequencies at the unlinked markers are all greater than about 15%.

Discussion

The main challenge in implementing our approach is to identify a sufficient number of unlinked, frequency-matched markers (probably 5 to 30) to extrapolate percentile cutoffs with reasonable accuracy. While this prospect might at first seem daunting, especially when close frequency matching is required and a large proportion of screened markers have to be rejected, there are efficient ways of finding appropriate markers. First, high-throughput genotyping technologies will facilitate the screening of many markers, so that a large subset will have frequencies in the right range. A second way to speed up the identification of markers is to choose them from a database that includes allele frequencies (i.e., dbSNP: www.ncbi.nlm.nih.gov/SNP/index.html). A marker with a published allele frequency similar to the candidate allele frequency is more likely than a random marker to be frequency matched, even if the population used for ascertainment is different from the patient population. Third, it is quite likely that instead of having one candidate association that is verified by genotyping a large number of new unlinked markers, the candidate association and unlinked markers will all be identified in the same screen.

While this speeds the identification of unlinked markers, it also necessitates the use of a χ^2 value that corrects for multiple hypothesis testing. In addition, to be conservative, it is necessary to include the candidate polymorphism's χ^2 value among the set used for obtaining cutoffs.

If the unlinked markers are not biallelic, frequency matching might appear to be even easier. For example, if the markers are microsatellites, the allelic classes can be carefully sorted into two categories in order to produce a "lumped" polymorphism with a frequency as close as possible to that of the candidate allele. However, it is not clear than microsatellites respond to stratification in the same way as SNPs, and hence we cannot currently recommend microsatellites as surrogates for unlinked biallelic markers in any context.

Population stratification can produce "spurious" signals of association in case-control studies. The presence of stratification, however, does not preclude the demonstration of true linkage. Here, we provide a method for teasing out evidence of true linkage in the context of stratification. The method can be easily implemented in a case-control association study, and requires the genotyping of only a moderate number of unlinked markers in cases and controls. By using the evidence from these unlinked markers, it is possible to distinguish the level of association that arises on account of population stratification alone, and which can be measured simply with μ, from the quantitatively greater association that occurs when a candidate polymorphism is actually associated to a disease-causing allele.

Acknowledgements

We are very grateful to David Altshuler, Mark Daly, Joel Hirschorn, Eric Lander and Sue Povey for their comments on early drafts of this paper. DER was supported in this work by a National Defense Science and Engineering Graduate Fellowship.

References

Chakraborty R, KM Weiss (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. USA 85:9119-9123.

Ewens WJ, RS Spielman (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet. 57:455-464.

Horvath S, NM Laird (1998) A discordant-sibship test for disequilibrium and linkage: No need for parental data. Am. J. Hum. Genet. 63:1886-1897.

Hudson RR (1991) Gene genealogies and the coalescent process. In: Oxford Surveys in Evolutionary Biology. Oxford University Press, Oxford, pp 1-44.

Kaplan NM (1994) Clinical Hypertension, 6th ed. (Williams and Wilkins, Baltimore, MD).

Lander ES, NJ Schork (1994) Genetic dissection of complex traits. Science 265:2037-2048.

Pritchard JK, NA Rosenberg (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 65:220-228.

Reich DE, MW Feldman, DB Goldstein (1999) Statistical properties of two tests that use multilocus data sets to detect population expansions. Mol. Biol. Evol. 16:453-466.

Rice JA (1995) Mathematical Statistics and Data Analysis, 2d ed. (Duxbury Press, Belmont, California).

Risch N, K Merikangas (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517.

Spielman RS, WJ Ewens (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. 62:450-458.

Table 1 • Contingency Table for Calculation of χ^2 Values

	Allele A	**Allele B**	
Cases	n_{11} = copies of allele A among cases	n_{12} = copies of allele B among cases	n_{1*} = *twice no. of cases*
Controls	n_{21} = copies of allele A among controls	n_{22} = copies of allele B among controls	n_{2*} = *twice no. of controls*
	n_{*1} = *copies of allele A*	n_{*2} = *copies of allele B*	n_{**} = *total no. of alleles*

Table 2 • No. of Markers vs. Precision of Estimation of μ

No. of Markers	Maximum Factor by Which μ can Exceed mean χ^2 value
1	3.84
2	3.01
3	2.60
4	2.37
5	2.21
6	2.10
7	2.01
8	1.94
9	1.88
10	1.83
11	1.79
12	1.75
14	1.69
16	1.65
18	1.61
20	1.57
22	1.54
24	1.52
26	1.50
28	1.48
30	1.46
32	1.45
34	1.43
36	1.42
38	1.41
40	1.40
45	1.37
50	1.35
55	1.33
60	1.32
65	1.30
70	1.29
75	1.28
80	1.27
85	1.26
90	1.26
95	1.25
100	1.24

Figures

Fig. 1 Empirical distributions of χ^2 values in the presence of population stratification, calculated for the demographic model described in the text. The disease model (for Distribution 2) is such that individuals who carry 2, 1 or 0 copies of the disease-associated allele contract disease with frequencies 16%, 4%, and 1% respectively. Population stratification alone causes 27% of replicates for unlinked markers to be nominally associated at the $P < 0.05$ level ($\chi^2 > 3.84$). However, χ^2 values drawn from Distribution 2 can be easily distinguished from Distribution 1.

Fig. 2 The mean of the empirical χ^2 distribution versus the 95th percentile and 99th percentiles, obtained by coalescent computer simulations for 100 cases and 100 controls. The following parameter combinations were considered: g = (0.05, 0.1, 0.3, and 0.4), τ = (0.025, 0.05, 0.1, 0.2 and 0.4), and relative risks of 2, 4 and 8. Eighty-five thousand simulations were performed for each of the 60 parameter combinations, and the results were binned into 7 allele frequency ranges (15-20%, 20-25%, 25-30%, 30-35%, 35-40%, 40-45% and 45-50%), each of which was used to calculate a mean and 95th percentile cutoff. The line fitted to the scatter-plot gives a remarkably good match to the data, except when stratification is severe, in which case the mean may overestimate the cutoff.

Fig. 3 χ^2 distributions exhibit a complex behavior as a function of allele frequency and population stratification scenario. Each trace corresponds to a unique combination of parameters, and plots 95th percentile cutoffs as a function of frequency. The following parameter

combinations were considered: g = (0.05, 0.1 and 0.3), τ = (0.025, 0.05, 0.1, 0.2 and 0.4), and relative risks of 2, 4 and 8. Eighty-five thousand simulations were performed for each of the 45 parameter combinations, for a sample size of 100 cases and 100 controls, and the results were binned into 9 allele frequency bins (from 5% to 50%), each of which was used to calculate a 95th percentile cutoff. For low values of g, the traces tend to decrease monotonically with allele frequency, while for high values, they tend to increase monotonically.

Fig. 4 95th percentile cutoffs as a function of allele frequency, for frequencies ranging from 1-10%. Fifty thousand simulations are performed for each stratification scenario, and there are 250 cases and 250 controls. Otherwise, parameters are identical to those in Fig. 3. For low frequencies, the 95th percentile cutoffs tend to be highly dependent on frequency, and to increase monotonically with frequency.

FIGURE 1

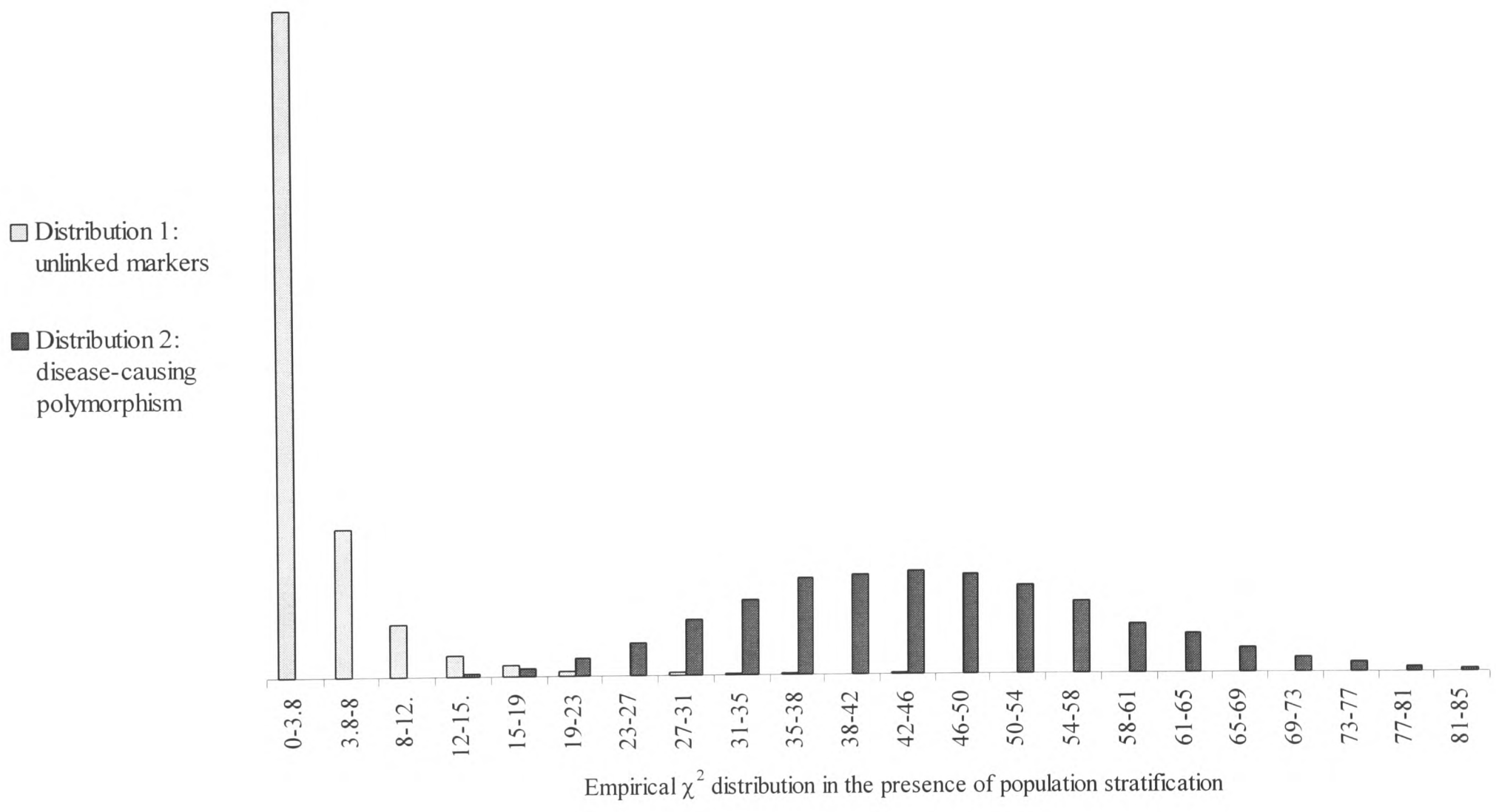

Distribution 1: unlinked markers
Distribution 2: disease-causing polymorphism
0-3.8
3.8-8
8-12.
12-15.
15-19
19-23
23-27
27-31
31-35
35-38
38-42
42-46
46-50
50-54
54-58
58-61
61-65
65-69
69-73
73-77
77-81
81-85
Empirical χ^2 distribution in the presence of population stratification

FIGURE 2

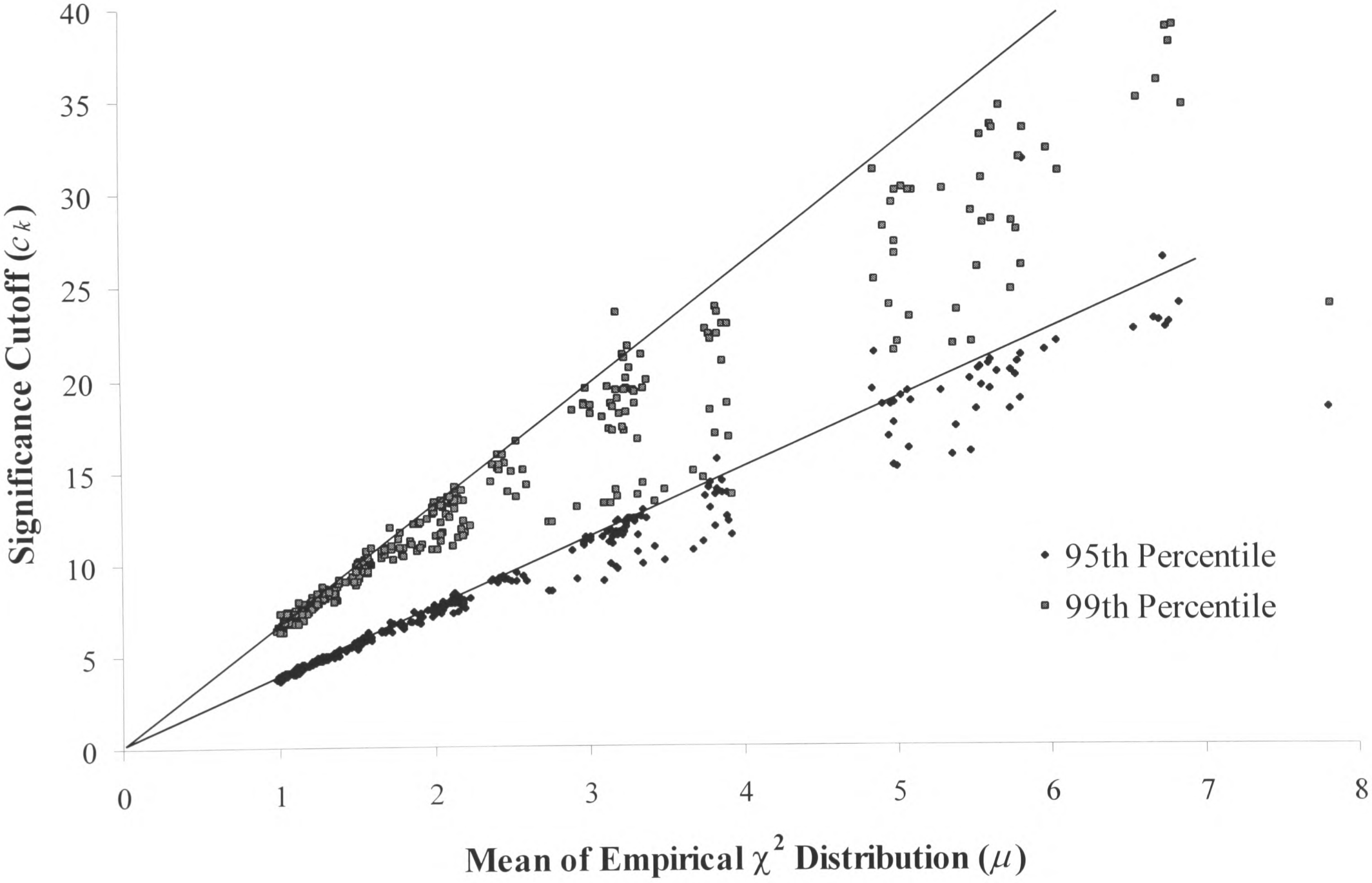

Significance Cutoff (c_k)
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
95th Percentile
99th Percentile
Mean of Empirical χ^2 Distribution (μ)

FIGURE 3

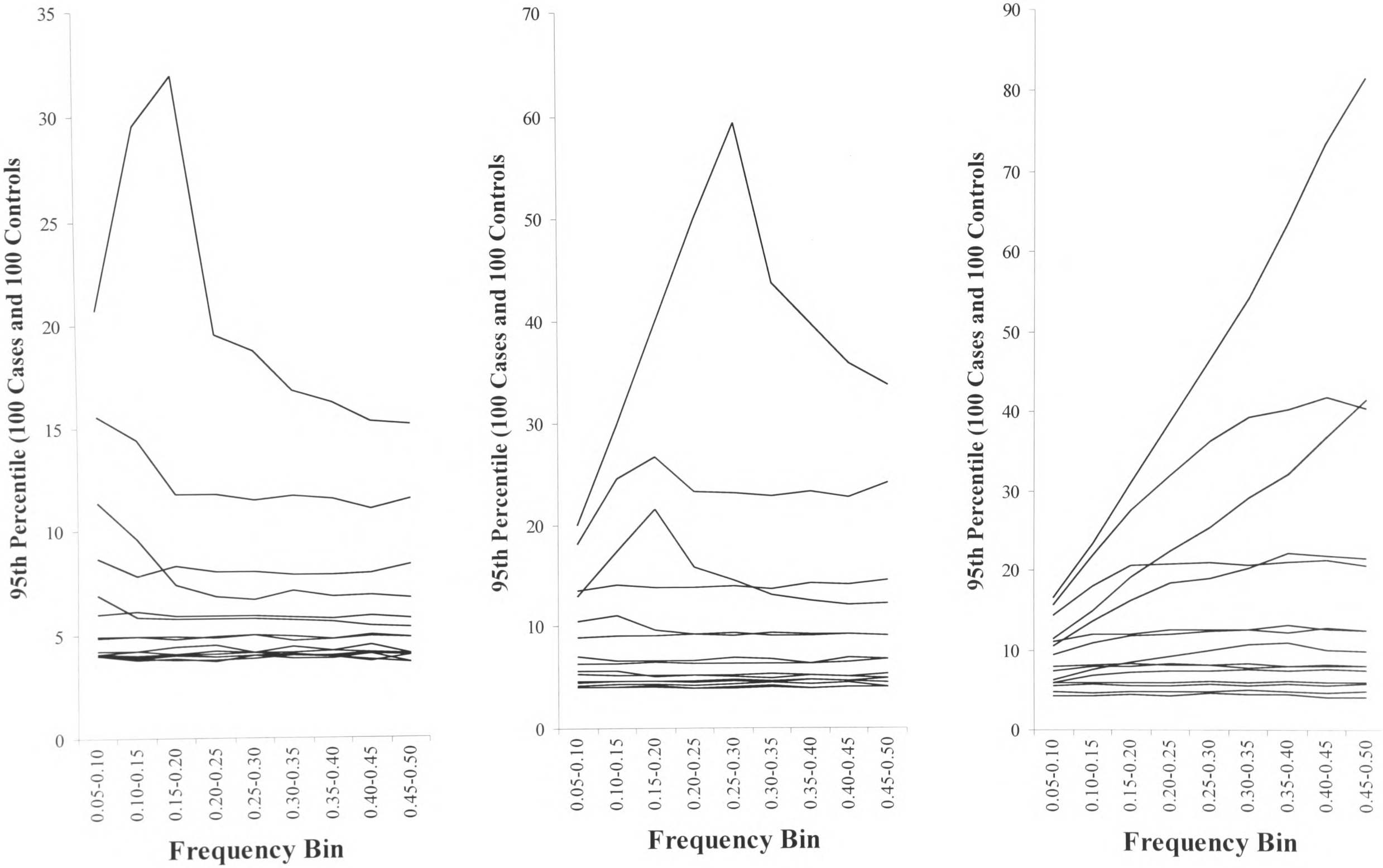
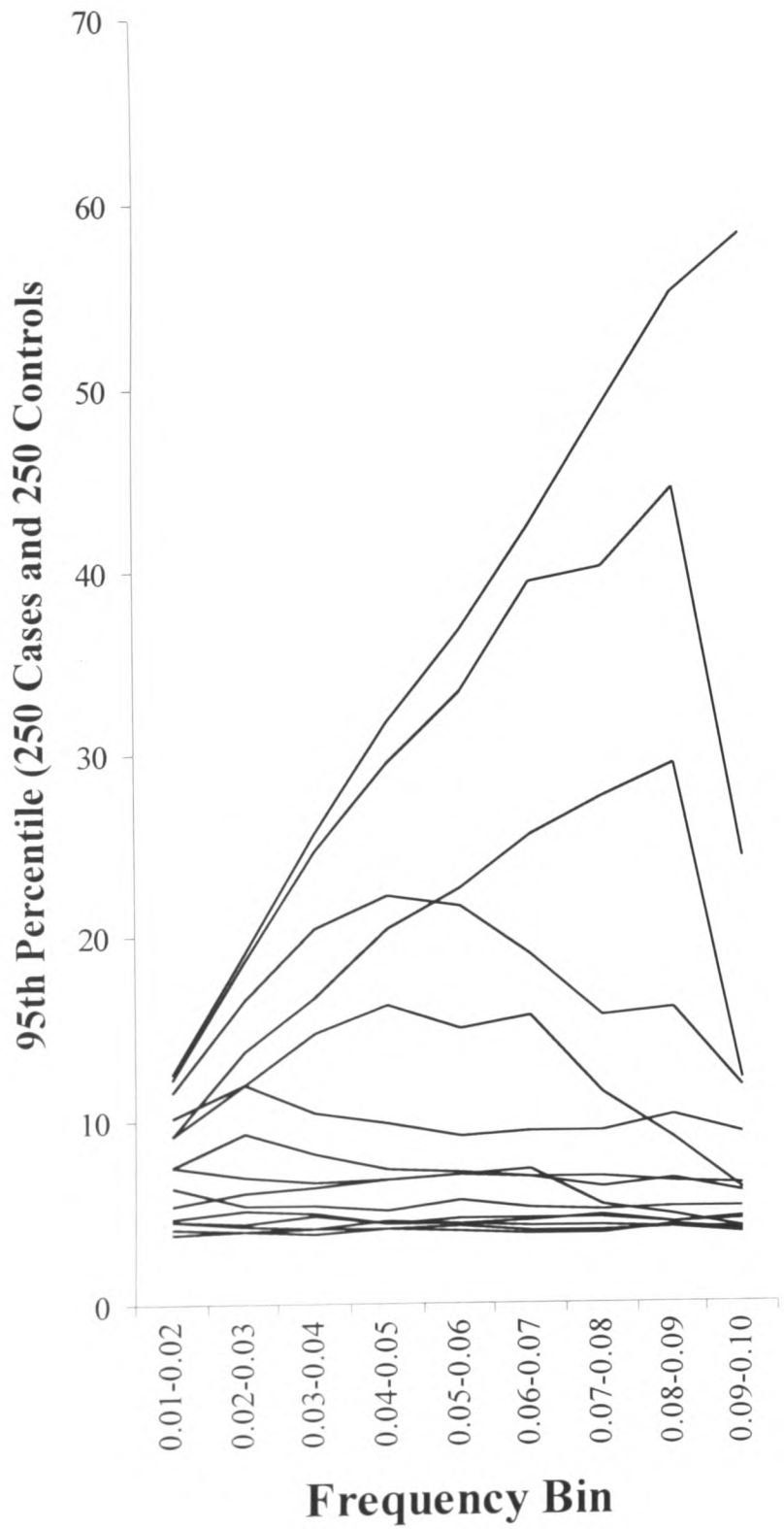
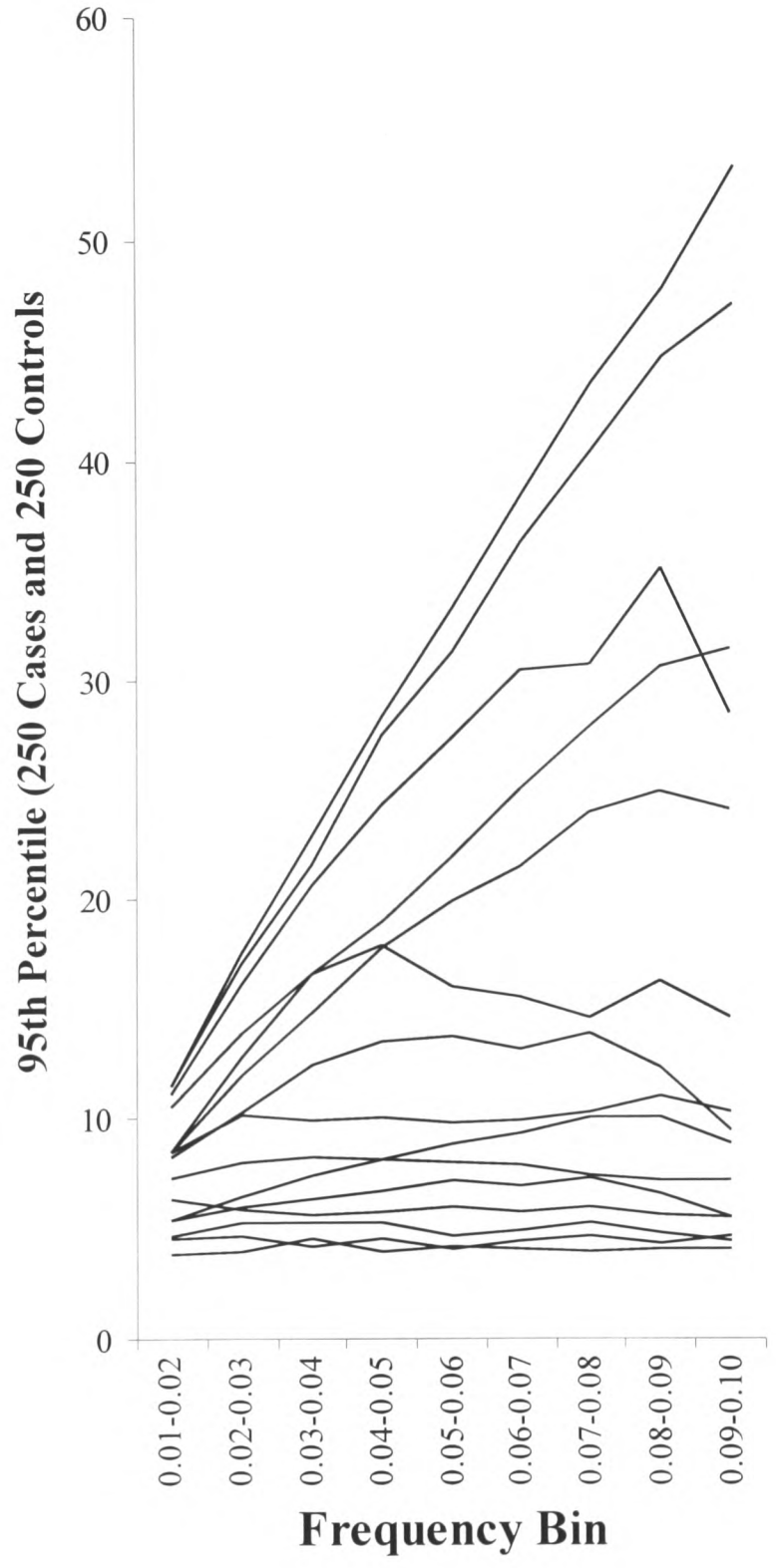
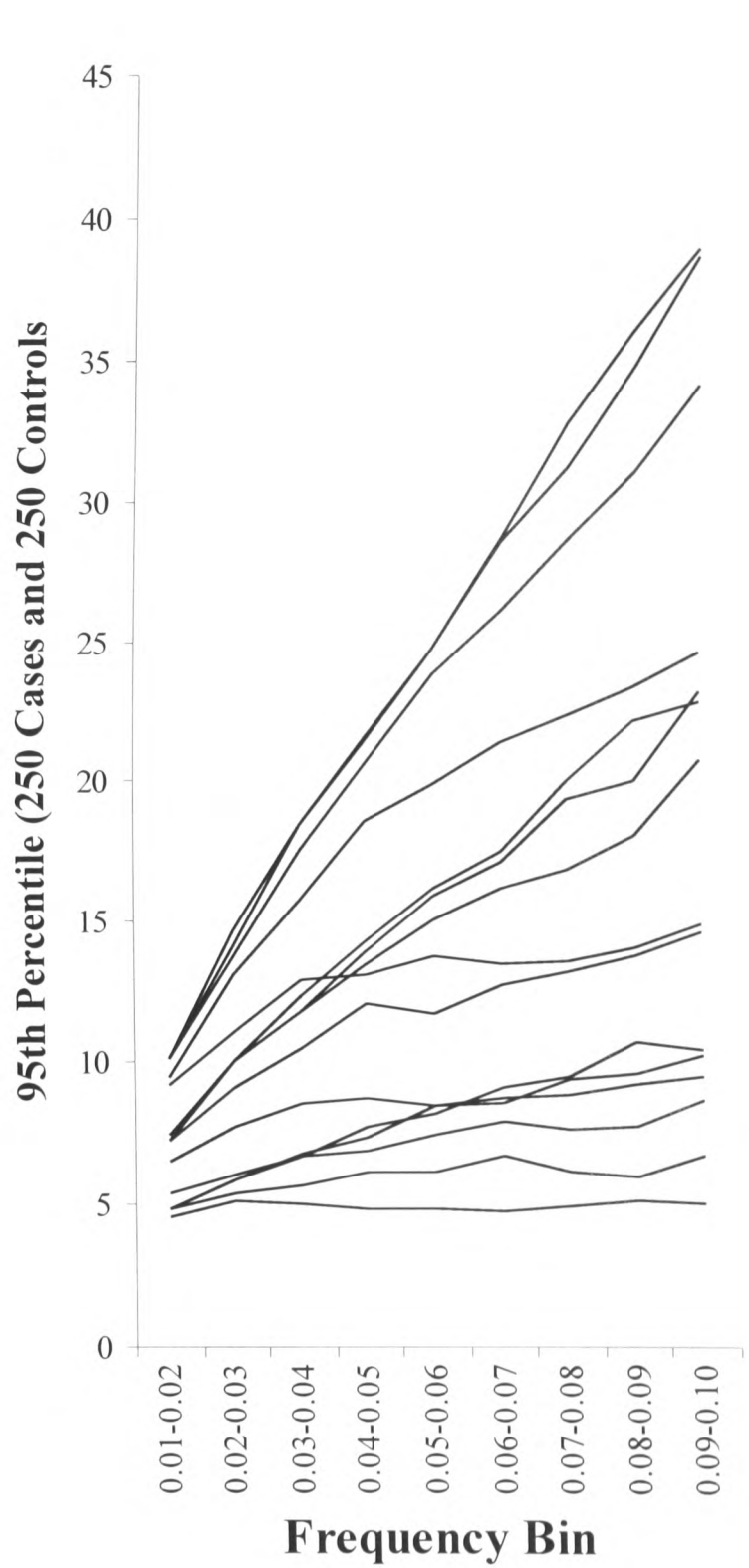

(A) g = 0.05
(B) g = 0.1
(C) g = 0.3
95th Percentile (100 Cases and 100 Controls
Frequency Bin
0.05-0.10
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50

FIGURE 4


(A) g = 0.05
95th Percentile (250 Cases and 250 Controls
0
10
20
30
40
50
60
70
0.01-0.02
0.02-0.03
0.03-0.04
0.04-0.05
0.05-0.06
0.06-0.07
0.07-0.08
0.08-0.09
0.09-0.10
Frequency Bin


(B) g = 0.1
95th Percentile (250 Cases and 250 Controls
0
10
20
30
40
50
60
0.01-0.02
0.02-0.03
0.03-0.04
0.04-0.05
0.05-0.06
0.06-0.07
0.07-0.08
0.08-0.09
0.09-0.10
Frequency Bin


(C) g = 0.3
95th Percentile (250 Cases and 250 Controls
0
5
10
15
20
25
30
35
40
45
0.01-0.02
0.02-0.03
0.03-0.04
0.04-0.05
0.05-0.06
0.06-0.07
0.07-0.08
0.08-0.09
0.09-0.10
Frequency Bin

Correlation of Genetic Associations Across Populations: Preliminary Investigations

David E. Reich[1,2] & David B. Goldstein[3]

[1] Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK. [2] Whitehead Institute/MIT Center for Genome Research, One Kendall Square, Cambridge, Massachusetts 02139, USA. [3] The Galton Laboratory, Department of Biology, University College London , Wolfson House, 4 Stephenson Way, London, NW1 2HE, UK.

SUMMARY

The correlation of genetic associations across populations has implications for medical genetics and population genetics. We present a blueprint for research into this correlation, and use simulations to explore its behavior for some simple scenarios.

INTRODUCTION

Linkage disequilibrium has become an increasingly useful tool for studying human genetics and human history. It has been used to map genes (e.g., Hästbacka et al. 1992; Kerem et al. 1989), and to make inferences about the histories of populations (Freimer et al. 1997; Laan et al. 1997). However, there are currently no approaches for making cross-population comparisons using linkage disequilibrium: i.e., for using genetic associations discovered in one population to predict the level of association in others. Intuitively, it is clear that if the populations are closely related, the predictability from one population to the other will be high. The quantitative level of correlation for more distantly related populations, however, is a complex function of the historical relationships and demographies of the populations.

Medical Applications:

(1) By using the known historical relationships among populations, which may be assessed by the degree of divergence in allele-frequencies among populations (Cavalli-Sforza et al. 1996), it may be possible to obtain precise expectations about

the extent to which a disease association discovered in one group (e.g., Caucasians) will be preserved in distantly related groups (e.g., Japanese). A quantitative understanding of this correlation may eliminate the need to repeat an expensive epidemiological study in every new population.

(2) If a genetic association to disease is detected, it can be difficult to know whether the associated allele is causal or merely linked to the causal allele. By repeating the study in several populations, however, and searching for a strikingly low level of fluctuation in the level of association from population to population (less than would be expected given the historical relationships among populations), it may be possible to detect the signature of an association. Such an assessment, however, requires a quantitative understanding of the patterns of correlation of linkage disequilibrium across populations, our main focus in this paper.

(3) Marker alleles may be positively correlated with disease in one population, but negatively correlated in another. Two populations may therefore separately show significant association between the allele and the disease, but if the populations are mixed, the associations may cancel. This cancellation will not cause problems for closely related populations. To make inferences for more distantly related populations, however, it is necessary to understand, at a quantitative level, the correlation of genetic associations across populations.

Population Genetic Applications:

The correlation may also be useful as population genetic distance, i.e., for assessing the degree of relatedness among populations. A distance based on the correlation of genetic associations will be complementary to traditional distances

based on allele frequency in the sense that it will respond in different ways to demographic scenarios. However, empirical characterization of the correlation in genetic associations across populations has direct medical implications as well. The medical applications in (1) - (3) may be best elucidated if empirically measured correlations in genetic associations across populations, rather than allele frequency differences, are used to gauge the degree of relatedness of populations.

SIMULATION METHODS

The coalescent simulations used in this paper are based on a published computer program, which was modified to allow for recombination between loci separated by a specified recombination distance *c* (Hudson 1991). The demographic model involves splitting of the populations at a time *t* in the past, and no exchange of genes between the populations subsequent to splitting. Figure 1 diagrams the model:

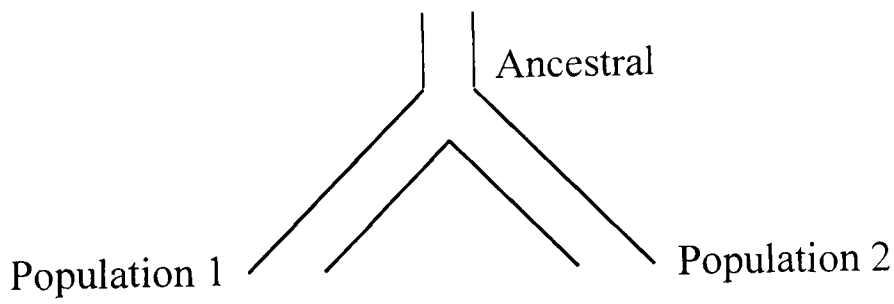


FIGURE 1

The ancestral population is constant in size, and the post-splitting population is assumed to be either the same size or expanded in size compared to the ancestral population. Samples of 200 chromosomes are drawn from each of the two modern populations (1 and 2), and gene genealogies are simulated, at both loci, under the

assumption that they are separated by a recombination fraction *c*. At each locus, a single mutation is placed on the gene genealogy, which corresponds to a scenario in which the overall branchlength of a gene tree does not affect the probability that a polymorphism will be observed at the associated locus (the expected scenario if very long DNA fragments are sequenced in order to discover polymorphisms). The probability that a mutation occurs on a branch of a particular tree is proportional to the length of the branch. At both loci, the mutations are only allowed to occur on the branches of the trees that result in the frequency of the less common allele in population 1 occurring within a narrow frequency range (25% ± 5%). The frequency of the less common allele in population 2 is required to be at least 10% in order to ensure a useful polymorphism in that population. Simulation results are dropped if they do not satisfy the frequency requirements in both populations.

THE SIMPLE CASE

Consider a population in which two polymorphic (biallelic) loci are separated by a genetic distance *c*. The population splits, at a time *t* in the past, into populations 1 and 2 (Figure 1), which initially have identical haplotype frequencies as the ancestral population (specified as p_{11}, p_{12}, p_{21} and p_{22}). The descendant populations are so large that there is no genetic drift, and hence no change in allele and haplotype frequencies subsequent to splitting. Recombination is then the only influence on haplotype frequencies in populations 1 and 2, breaking down disequilibrium at an exponential rate with a time constant *tc*. The projected contingency table is given in Table 1.

	Locus A, Allele 1	Locus A, Allele 2	*Marginals*
Locus B, Allele 1	$p_{11}e^{-tc} + p_{1*}p_{*1}(1-e^{-tc})$	$p_{12}e^{-tc} + p_{1*}p_{*2}(1-e^{-tc})$	p_{1*}
Locus B, Allele 2	$p_{21}e^{-tc} + p_{2*}p_{*1}(1-e^{-tc})$	$p_{22}e^{-tc} + p_{2*}p_{*2}(1-e^{-tc})$	p_{2*}
Marginals	p_{*1}	p_{*2}	

Table 1: Contingency table for biallelic markers in the case of no drift. Allele 1 at locus A is the less common allele at that locus, while allele 1 at locus B is the allele that is positively correlated with allele 1 at locus A.

In Table 1, p_{ij} is the population frequency of chromosomes that have allele i at locus A and allele j at locus B. The association, S, is defined as the ratio of the observed p_{11} haplotype, to the value predicted by the marginals: $S \equiv (p_{11}e^{-tc}+p_{1*}p_{*1}(1-e^{-tc}))/(p_{1*}p_{*1})$. S is related to the classical measure of disequilibrium, $D \equiv p_{11}e^{-tc}+p_{1*}p_{*1}(1-e^{-tc}) - p_{1*}p_{*1}$, in a simple way: $S = 1+D/(p_{1*}p_{*1}) = 1+D_0e^{-tc}/(p_{1*}p_{*1})$, where $D_0 = p_{11} - p_{1*}p_{*1}$ is the ancestral level of disequilibrium. S decays to its equilibrium frequency (1) as D decays to 0.

If there is no genetic drift, the underlying level of association, S, will be the same in populations 1 and 2 (there is no differentiation between the populations). To predict the level of association in population 2, it is therefore sufficient to measure the level in population 1. If there is substantial genetic drift separating the populations, however, the correlation in S values across populations is no longer expected to be high.

COMPLICATION: GENETIC DRIFT

If two populations are distantly related, substantial genetic drift is likely to have occurred since their separation, either in population 0, population 1, or both.

Drift may be due to a bottleneck at the time of population splitting, or small population sizes subsequent to splitting. Genetic drift can also generate new linkage disequilibrium and new genetic associations (Slatkin 1994), in addition to what already existed at the time of population separation. Hence, when genetic drift is present, Table 1 no longer gives correct values for the expected haplotype and allele frequencies.

To understand the effect of genetic drift, we note that the genetic associations inherited from the ancestral population decay at a rate of e^{-tc}, but drift also generates new linkage disequilibrium that may either increase or decrease the level of genetic association in both populations. This causes the value of the association to diverge between population 1 and population 2, and explains the differences that are observed in S between populations, above and beyond sampling effects. Simulations can then be used to explore how much the predictive power of the allelic state in population 1 is degraded, in relation to population 2, as a function of the amount of genetic drift that occurred subsequent to population splitting.

Figure 2 presents results of the simulations for an allele frequency of 25% ± 5%, a recombination fraction between the linked loci of c = 0.05cM, and a pre-expansion, pre-splitting population size of N = 5,000 individuals. Two demographic scenarios are considered: sudden population expansion (by a factor of 1,000) that occurred after the split, and no expansion at all, which allows for genetic drift. Each of these scenarios is evaluated at two points in time: just after the split, and at time t = 0.25, corresponding to 100,000 years assuming 25 years per generation. The values of S in the two populations are plotted in Figure 2 (population 1 on the x-axis, and population 2 on the y-axis), and the degree of correlation can be quantitatively

assessed, in a rough way, by using the R^2 value for a least-squares fit going through the origin.

For a dramatic expansion, our simulations confirm that the correlation of the value of S in the two populations remains extremely high long after population splitting, even though the absolute value of S tends to decline (right column of Figure 2). In the presence of substantial genetic drift, however, the correlation in genetic association between the two populations decreases dramatically even though the absolute value of S may be unchanged or even higher in the descendant populations than in the ancestral population due to the generation of new linkage disequilibrium through genetic drift (left column of Figure 2).

CONCLUSION

In this preliminary investigation, we have used computer simulations and analytical modeling to develop intuitions about the correlation of genetic associations across populations. We showed that when there is no genetic drift—which occurs, for example, after a dramatic population expansion—the correlation of the genetic association across populations does not decline with time, even though the magnitude of the expected association decreases (Figure 2). However, for the more complex case of genetic drift, the opposite can occur: the correlation can decrease with time, while the magnitude of the genetic association can stay the same or even increase. The degree to which an association discovered in population 1 correlates with and predicts the association in population 2 is therefore highest for populations that have grown dramatically. Thus, for pairs of populations that have undergone rapid growth since

splitting (e.g., European Neolithic populations), a strong genetic association discovered in one population should be a good indication of an association in another, regardless of the degree of relatedness of the populations. In contrast, pairs of populations that have been subject to substantial genetic drift, even if they have diverged relatively recently, may be good candidates for differentiating between disease-causing alleles and alleles that are only linked to the disease-causing locus (point 2 in the Introduction).

Very few parameter combinations were simulated for this paper, and hence it is difficult to make generalizations, especially quantitative ones, from the results. To provide answers to the more important questions in the Introduction, however, it will be necessary to consider a full range of parameter combinations and to develop a quantitative framework for understanding the correlation of genetic associations across populations.

REFERENCES

Cavalli-Sforza, L.L., Menozzi, P. and Piazza A. (1996) *The History and Geography of Human Genes* (Princeton University Press, Princeton).

Freimer, N.B., Service, S.K. and Slatkin, M. (1997) Expanding on population studies. *Nature Genetics* **17:** 371-373.

Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, W. and Lander, E.S. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics* **2:** 204-211.

Hudson, R.R. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. & Antonovics, J.) 1-44 (Oxford University Press, Oxford, 1991).

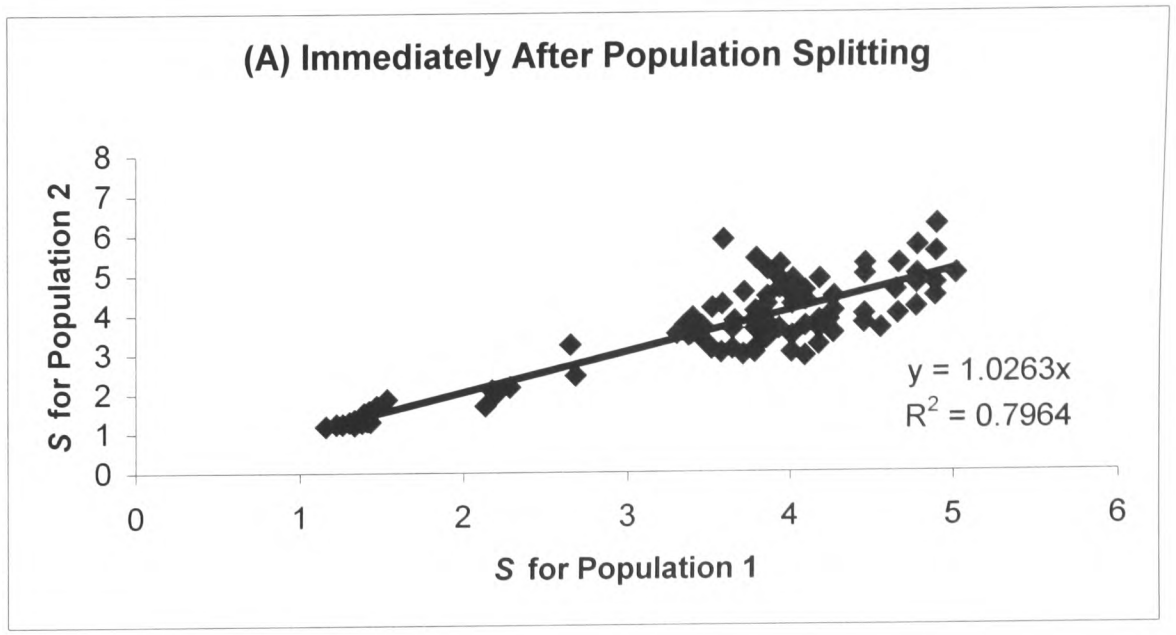
Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui, L.C. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* **245:** 1073-1080.

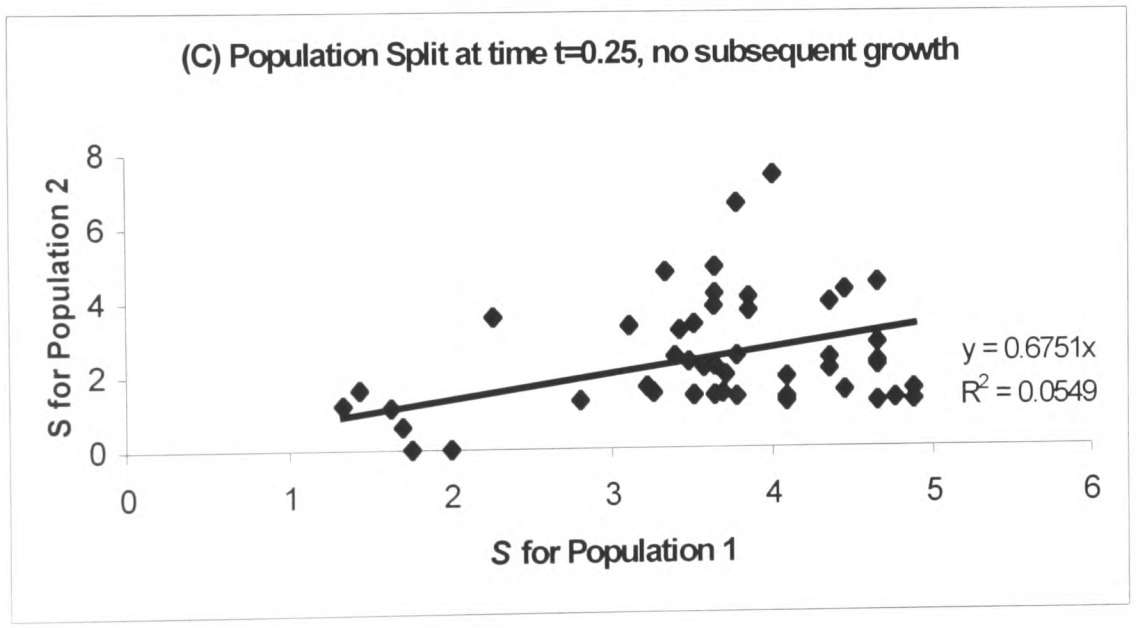
Laan, M. and Päabo, S (1997) Demographic history and linkage disequilibrium in human populations. *Nature Genetics* **17**: 435-438.

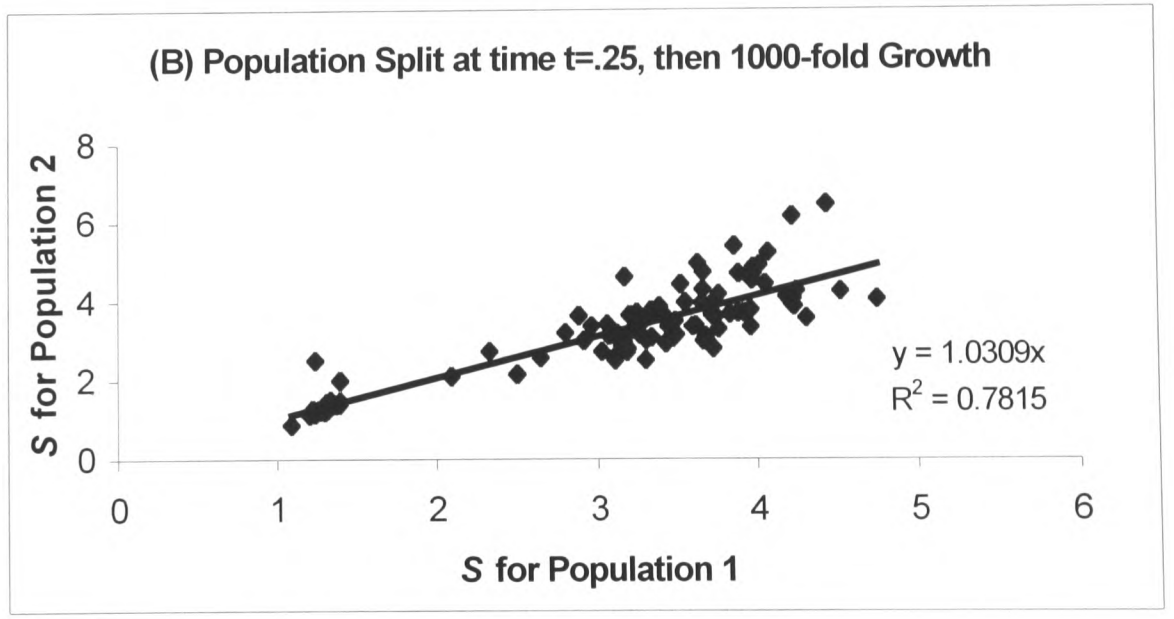
Slatkin, M. Linkage disequilibrium in growing and stable populations (1994) *Genetics* **137:** 331-336.

FIGURE 2

Fig. 2: Correlations in the genetic association (*S*) between populations for several demographic scenarios. All graphs are generated for allele frequencies at both loci of 25% ± 5%, a sample size of 200, c = 0.05cM, and a pre-expansion effective population size of 5,000. Correlations in the genetic association (*S*) across are considered for: (A) the populations just after splitting; (B) the populations at a time t=0.25 after splitting (the split is followed by 1000-fold growth); and (C) the populations at a time t=0.25 after splitting (the split is not followed by growth, so genetic drift occurs). The average value of S in figure (B) is less than in (A), due to the effects of recombination. However, the correlation of *S* between the two populations, as measured by the R^2 value, remains the same for (A) and (B).







Conclusion

The genomic revolution has resulted in dramatic improvements in the quality of data available for studying population history, as well as tools for identifying disease associations. Now, with the development of methods of data analysis that produce high-resolution pictures of the history and structure of populations, it is possible to begin using population genetic approaches to quantitatively inform the effort to choose populations that are optimal for addressing questions in genetic epidemiology and medical genetics.

The first chapter of this thesis introduces a novel, multilocus approach for studying population history, and applies statistical tests based on this approach to various data sets in order to identify populations with histories of growth and contraction. The second chapter, on date estimates, also has medical relevance: in particular, the recent date for CCR5-Δ32 suggests that this mutation confers resistance not only to AIDS but, in addition, to yet another disease. The third chapter discusses how to take account of population structuring in case-control associations studies, and raises the problem of how genetic associations detected in one population decay as a function of increasing genetic distance between populations.

Finally, several issues should be noted that emerge from the work described in this thesis but that were not discussed in its individual chapters.

The significantly reduced value of the interlocus *g* statistic in some African populations (Chapter 1) not only suggests a history of expansion, but is also interesting because populations with low values of *g* have important practical uses. Just as inbred and isolated populations (e.g., Finns, Icelanders, and Ashkenazi Jews) can serve as tools for mapping genes, populations with low *g* values can serve as tools for making estimates about biological parameters. For example, the signature of mutation rate variation across loci (greater variability in the density of polymorphisms across regions than would be expected by chance) can also be produced by variation in gene genealogies across loci. A low value of *g* minimizes this confounding effect, facilitating the detection of variation in the mutation rate across loci. As a general rule, if a population is going to be used as a tool for inferring a biological parameter, it is always better to use one with a low rather than a high value of *g*.

The data analysis in Chapter 2 data analysis not only reveals the histories of the CCR5 and Factor XI mutations, but also hints more generally at the power of methods that use disease genes to study certain aspects of population history, especially bottlenecks. Dates for disease mutations may coincide with the dates of historical bottlenecks, since all the sampled copies of a rare allele may share a most recent common ancestor that occurred around the time of the bottleneck, when there was intense genetic drift. However, for more common variants, the most recent ancestor is less likely to have occurred around the time of the bottleneck, because the genetic drift associated with the bottleneck is less likely to have reduced the ancestors of the modern samples to a singleton. It may therefore be no coincidence that the only genetic evidence for a

medieval bottleneck among Ashkenazi Jews comes from the study of a rare disease mutation causing idiopathic torsion dystonia (RISCH *et al.* 1997). Indeed, the power of disease mutations to elucidate certain types of historical events also explains, more generally, why clinical laboratories sometimes make discoveries about human history that are not accessible to traditional population genetic data sets. Many thousands of randomly chosen individuals have to be screened in a traditional population genetic approach in order to obtain the number of copies of a rare allele that are need to obtain a high-resolution estimate for the date. However, much less sampling is necessary when disease mutations are studied, because individuals carrying the mutations walk right into the clinic—in effect, they sample themselves.

Chapter 3 shows explicitly how assessments of population structure and history can improve the techniques of medical genetics. The first section demonstrates, in the context of case-control studies, how it is possible to calculate significance levels for a genetic association in the face of population stratification. Over the longer term, it seems promising to explore other connections between population genetics and medical genetics, applying insights from evolutionary history to choose optimal populations for disease-gene mapping. The selection of populations for fine-scale mapping is currently rather primitive, based on intuitions and prior experience. However, if tests of demographic history (e.g., Chapter 1) can be incorporated into the selection of populations, together with knowledge about the ages of typical mutations (e.g., Chapter 2), the potential for choosing populations that are optimal for mapping genes will be dramatically improved.

Reference:

RISCH N., D. DE LEON, L. OZELIUS, P. KRAMER, L. ALMASY, *et al.*, 1995 Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nature Genetics **9:** 153-159.

Acknowledgements

The research in this thesis has been enriched by conversations and collaboration with numerous scientists, including David Altshuler, Kristin Ardlie, Myles Axton, Ranajit Chakraborty, Mike Charleston, Peter Donnelly, Marc Feldman, Bob Griffiths, Paul Harvey, Joel Hirschorn, Eddie Holmes, Marek Kimmel, Eric Lander, Kerstin Lindblad, Stephen O'Brien, Jonathan Pritchard, Oliver Pybus, Daniel Reich, Nati Srebro and Jim Wilson. I am most grateful of all to my doctoral supervisor, David Goldstein, with whom I established a remarkably close working relationship during the past few years, and who has been a mentor without equal during my transformation from student to scientist.

Financial support for my work came from an Overseas Research Student Award, an Oxford Overseas Bursary, The Graduate Scholarship in Genetics at St. Catherine's College, and research assistantships from the Goldstein laboratory. During the academic year 1998-1999, I held a National Defense Science and Engineering Graduate Fellowship, which enabled me to take courses and to research in the United States while continuing my work collaboratively with the Goldstein laboratory. My time in the U.S. was especially useful for Chapter 3 of this thesis.

Appendix 1

List of publications and submitted research papers

* indicates a paper included in the thesis

** indicates a paper that is partially incorporated into the text of the thesis.

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

(2) * D.E. Reich and D.B. Goldstein (1998). Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA, 95:8119-8123.

(3) ** D.E. Reich, M.W. Feldman, D.B. Goldstein (1998). Statistical properties of two tests for demographic history, with application to human demography. Abstract presented at the American Society of Human Genetics annual meeting.

(4) * D.E. Reich and D.B. Goldstein (1999). Estimating the age of mutations using variation at linked markers. In *Microsatellites: Evolution and Applications*, ed. D.B. Goldstein and C. Schlötterer, Oxford University Press, 129-138.

(5) D.E. Reich, R.K. Wayne, D.B. Goldstein (1999). Genetic evidence for a recent origin by hybridization of red wolves. Molecular Ecology, 8:139-145.

(6) D.B. Goldstein, G.W. Roemer, D. Smith, D.E. Reich, A. Bergman, R.K. Wayne (1999). The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics, 151:797-801.

(7) * D.E. Reich, M.W. Feldman, D.B. Goldstein (1999). Statistical properties of two tests that use multilocus data sets to detect population expansions. Mol. Biol. Evol., 16:453-456.

(8) ** D.B. Goldstein, D.E. Reich, N. Bradman, S. Usher, U. Seligsohn, H. Peretz (1999). Age estimates for two common mutations causing factor XI deficiency: Recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. Am. J. Hum. Gen., 64:1071-1075.

(9) * D.E. Reich and D.B. Goldstein. Detecting associations in a case-control study in the face of population stratification. Submitted to Am. J. Hum. Gen.

(10) * D.E. Reich. Single nucleotide polymorphisms as tools for studying demographic history. Submitted to Genetics.

Appendix 2

Code for coalescent computer simulations

(A) Code for microsatellite-based tests of demography

The following coalescent simulation, a modified version of a C program by R. Hudson (1991; Oxf. Surv. Evol. Biol., **7:**1-44), calculates statistics for the within-locus and interlocus tests for a constant sized population, and for sudden expansions from an initial diploid population size of N_0 to a final population size of N_f. The time of expansion is specified in units of N_0 generations, and the mutation rate is given via the parameter $N\nu = \theta/4$.

The simulation assumes a single-step mutation model in which microsatellite mutations change allele lengths by exactly one step, and in which the change has an equal probability of occurring in either direction. The simulation can also enact more complex mutation processes, in which the average step size can be greater than one. Unless specified explicitly in the simulation, the mutation rate and mutation process are identical across loci. However, the program can also handle interlocus variation in the mutation rate, and interlocus variation in the propensity to multi-step mutation.

The simulation can be obtained in electronic form at the following URL:

www.ucl.ac.uk/biology/goldstein/davidr~1.htm

```
/* COALESCENT SIMULATION FOR STEPWISE MUTATION MODEL */
/* calculates statistics for within-locus and interlocus tests */
/* can incorporate:
        (1) interlocus variation in the mutation rate (see main routine)
        (2) interlocus variation in step size (see mutate_tree)
        (3) interlocus variation in the propensity to multi-step mutations
```

```
        (see mutate_tree) */
/* time is measured in units of N0 generations, with N0 the pre-expansion pop size */
/* August 23, 1998 */
/* by David Reich; adapted from Hudson's (1990) coalescent simulation */

#include <stdio.h>
#include <stdlib.h>
#include <math.h>

/* universal declarations of variables, output files, functions, and subroutines */
struct node {
        double time;
        int muts;
        int undmuts;
        struct node *desc1;
        struct node *desc2;
        struct node *ancestor;
        } ;
struct stats {
        double time;
        int undmuts;
        double mean;
        double var;
        double k;
        double test;
} ;

FILE *fp_out;

double ran1();
double rndgamma();
double rndgamma1();
double rndgamma2();
void make_tree();
void bottleneck();
double gammln();
double poidev();
double bnldev();
void mutate_tree();
int find_muts();
void calc_stats();
double find_median();

/* random number generator on a uniform distribution (from Numerical Recipes in C) */
double ran1(int *idum)
        #define IA 16807
        #define IM 2147483647
        #define AM (1.0/IM)
        #define IQ 127773
        #define IR 2836
        #define NTAB 32
        #define NDIV (1+(IM-1)/NTAB)
        #define EPS 1.2e-7
        #define RNMX (1.0-EPS)
{
        int j;
        int k;
        static int iy=0;
        static int iv[NTAB];
        double temp;

        if (*idum <= 0 || !iy) {
                if (-(*idum) < 1) *idum=1;
                else *idum = -(*idum);
                for (j=NTAB+7;j>=0;j--) {
                        k=(*idum)/IQ;
                        *idum=IA*(*idum-k*IQ)-IR*k;
                        if (*idum < 0) *idum += IM;
                        if (j < NTAB) iv[j] = *idum;
```

```
                iy=iv[0];
        }
        k=(*idum)/IQ;
        *idum=IA*(*idum-k*IQ)-IR*k;
        if (*idum < 0) *idum += IM;
        j=iy/NDIV;
        iy=iv[j];
        iv[j]=*idum;
        if ((temp=AM*iy) > RNMX) return RNMX;
        else return temp;
}

/* gamma function routine #1 from Yang's PAML package  */
double rndgamma (double s, int *idum)
{
        double ran1(int *idum);
        double rndgamma1(double s1, int *idum);
        double rndgamma2(double s2, int *idum);
        double  r=0.0;

        if (s <= 0.0)
                return 0;
        else if (s < 1.0)
                r = rndgamma1 (s, idum);
        else if (s > 1.0)
                r = rndgamma2 (s, idum);
        else
                r =- log(ran1(idum));
        return (r);
}

/* gamma function routine #2 from Yang's PAML package */
double rndgamma1 (double s, int *idum)
{
        double                  r, x=0.0, small=1e-37, w;
        static double  a, p, uf, ss=10.0, d;

        if (s!=ss)
                {
                a  = 1.0-s;
                p  = a/(a+s*exp(-a));
                uf = p*pow(small/a,s);
                d  = a*log(a);
                ss = s;
                }
        for (;;)
                {
                r = ran1(idum);
                if (r > p)
                        x = a-log((1.0-r)/(1.0-p)), w=a*log(x)-d;
                else if (r>uf)
                        x = a*pow(r/p,1/s), w=x;
                else
                        return (0.0);
                r = ran1(idum);
                if (1.0-r <= w && r > 0.0)
                        if (r*(w+1.0) >= 1.0 || -log(r) <= w)
                                continue;
                break;
                }
        return (x);
}

/* gamma function routine #3 from Yang's PAML package */
double rndgamma2 (double s, int *idum)
{
        double                  r ,d, f, g, x;
        static double  b, h, ss=0;

        if (s!=ss)
```

```
                {
                b  = s-1.0;
                h  = sqrt(3.0*s-0.75);
                ss = s;
                }
        for (;;)
                {
                r = ran1(idum);
                g = r-r*r;
                f = (r-0.5)*h/sqrt(g);
                x = b+f;
                if (x <= 0.0)
                        continue;
                r = ran1(idum);
                d = 64*r*r*g*g*g;
                if (d*x < x-2.0*f*f || log(d) < 2*(b*log(x/b)-f))
                        break;
                }
        return (x);
}

/* constructs the evolutionary tree (adapted from Hudson's coalescent simulation */
void make_tree(struct node *tree, int sample_size, int *idum)
{
        double ran1(int* idum);
        int in, pick;
        double t;
        struct node **list;

/* initializations */
        list = (struct node **)malloc( sample_size*
                        sizeof(struct node *) );
        for(in=0; in<sample_size; in++) {
                tree[in].time = 0.;
                tree[in].muts = 0;
                tree[in].desc1 = tree[in].desc2=0;
                list[in] = tree + in ;
        }
/* generate the times of the nodes */
        t=0.;
        for(in= sample_size; in>1; in--) {
                t += -2.0 * log(1.-ran1(idum)) / ( ((double)in)*(in-1) );
                tree[2*sample_size - in].time = t;
        }
/* generate the topology of the tree */
        for (in=sample_size; in>1; in--){
                pick = (int)floor(in*ran1(idum));
                list[pick]->ancestor=tree+2*sample_size-in;
                tree[2*sample_size-in].desc1=list[pick];
                list[pick]=list[in-1];
                pick = (int)floor((in-1)*ran1(idum));
                list[pick]->ancestor = tree + 2*sample_size - in;
                tree[2*sample_size - in].desc2 = list[pick];
                list[pick] = tree + 2*sample_size - in;
        }
        free(list);
}

/* rescales the tree in order to simulate a history of sudden growth or contraction.
   Adapted from Hudson's (1990) coalescent simulation. */
void bottleneck(struct node *tree,int sample_size,double tcrit,double f1)
{
        int in;

        for (in=sample_size;in<2*sample_size-1;in++)
                if (tree[in].time>tcrit) tree[in].time
                        = tcrit+(tree[in].time-tcrit)/f1;
}

/* gamma function routine, necessary for probability distribution functions */
```

```
/* from Numerical Recipes in C */
double gammln(double xx)
        #define PI 3.141592654
{
        double x,y,tmp,ser;
        static double cof[6]={76.18009172947146,
                -86.50532032941677,24.01409824083091,
                -1.231739572450155, 0.1208650973866179e-2,
                -0.5395239384953e-5};
        int j;

        y=x=xx;
        tmp=x+5.5;
        tmp -= (x+0.5)*log(tmp);
        ser=1.000000000190015;
        for (j=0;j<=5;j++) ser += cof[j]/+y;
        return -tmp+log(2.5066282746310005*ser/x);
}

/* generates poisson deviate (from Numerical Recipes in C) */
double poidev(double xm, int *idum)
{
        double gammln(double xx);
        double ran1(int *idum);
        static double sq,alxm,g,oldm=(-1.0);
        double em,t,y;

        if (xm < 12.0) {
                if (xm != oldm) {
                        oldm=xm;
                        g=exp(-xm);
                }
                em = -1.0;
                t = 1.0;
                do {
                        ++em;
                        t *= ran1(idum);
                } while (t > g);
        } else {
                if (xm != oldm) {
                        oldm=xm;
                        sq=sqrt(2.0*xm);
                        alxm=log(xm);
                        g=xm*alxm-gammln(xm+1.0);
                }
                do {
                        do {
                                y=tan(PI*ran1(idum));
                                em=sq*y+xm;
                        } while (em < 0.0);
                        em=floor(em);
                        t=0.9*(1.0+y*y)*exp(em*alxm-gammln(em+1.0)-g);
                } while (ran1(idum) > t);
        }
        return em;
}

/* return binomial deviates for stepwise mut. model (from Numerical Recipes in C ) */
double bnldev(double pp, int n, int *idum)
{
        double gammln(double xx);
        double ran1(int *idum);
        int j;
        static int nold=(-1);
        double am,em,g,angle,p,bnl,sq,t,y;
        static double pold=(-1.0),pc,plog,pclog,en,oldg;

        p=(pp <= 0.5 ? pp : 1.0-pp);
        am=n*p;
        if (n < 25) {
```

```
        bnl=0.0;
        for (j=1;j<=n;j++)
                if (ran1(idum) < p) ++bnl;
} else if (am < 1.0) {
        g=exp(-am);
        t=1.0;
        for (j=0;j<=n;j++) {
                t *= ran1(idum);
                if (t < g) break;
        }
        bnl=(j <= n ? j :n);
} else {
        if (n != nold) {
                en=n;
                oldg=gammln(en+1.0);
                nold=n;
        } if (p != pold) {
                pc=1.0-p;
                plog=log(p);
                pclog=log(pc);
                pold=p;
        }
        sq=sqrt(2.0*am*pc);
        do {
                do {
                        angle=PI*ran1(idum);
                        y=tan(angle);
                        em=sq*y+am;
                } while (em < 0.0 || em >= (en+1.0));
                em=floor(em);
                t=1.2*sq*(1.0+y*y)*exp(oldg-gammln(em+1.0)
                        -gammln(en-em+1.0)+em*plog+(en-em)*pclog);
        } while (ran1(idum) > t);
        bnl=em;
}
if (p != pp) bnl=n-bnl;
        return bnl;
}

/* distributes mutations on a tree according to poisson deviate.
   Adapted from Hudson's coalescent sim.; modified for stepwise mutation model */
void mutate_tree(struct node *tree,int sample_size,int *idum,double theta,
                                double param,int varq)
{
        double poidev(double xm, int *idum);
        double ran1(int* idum);
        double rndgamma (double s, int *idum);
        int in, nmuts,t1,j,steps,change;
        double td,pmean,pvar,a,c,par;

/* adapts to multi-step muts. or interlocus variation in multi-step mutations */
        if (varq==0 || varq==1) par=0.0;
        else if (varq==2) par=param;
        else { /* case of interlocus variation */
/* sets average propensity to multi-step mutation at pmean */
                pmean=0.20;
                pvar=param;
                a=pmean*pmean/pvar;
                c=pmean/pvar;
                par=(rndgamma(a,idum))/c;
        }
        nmuts=0;
        for (in=sample_size; in>1; in--){
                td=tree[2*sample_size-in].time - tree[2*sample_size-in].desc1->time;
                t1 = (int)poidev(td*theta/2.,idum);
                tree[2*sample_size-in].desc1->undmuts = t1;
                nmuts += t1;
                change=0;
                for (j=0;j<t1;j++) {
                        steps=1+(int)poidev(par,idum);
```

```
                        change=change+steps*(2*(int)bnldev(0.5,1,idum)-1);
                }
                t1=change;
                tree[2*sample_size-in].desc1->muts = t1;
                td=tree[2*sample_size-in].time - tree[2*sample_size-in].desc2->time;
                t1 = (int)poidev(td*theta/2.,idum);
                tree[2*sample_size-in].desc2->undmuts = t1;
                nmuts += t1;
                change=0;
                for (j=0;j<t1;j++) {
                        steps=1+(int)poidev(par,idum);
                        change=change+steps*(2*(int)bnldev(0.5,1,idum)-1);
                }
                t1=change;
                tree[2*sample_size-in].desc2->muts = t1;
        }
}

/* finds mutations from common ancestor */
int find_muts(struct node *node,double coatime,int sample_size)
{
        int in,muts;

        muts = 0;
        for (in=0;in<=sample_size;in++) {
                if ((node->time)<coatime) {
                        muts += node->muts;
                        node = node->ancestor;
                }
                else if ((node->time)==coatime) return(muts);
                else printf("ERROR IN FIND_MUTS!!!\n");
        }
}

/* calc_stats records run statistics in the "summary" data structure */
void calc_stats(struct node *tree, int sample_size, struct stats *summary, int * idum)
{
        int count,total_undmuts,st[200],j;
        double Var,VVVV,coatime,mean,n,f1,f2,f3,f4,Sig4,Gam4;
        double ran1(int* idum), poidev(double xm, int *idum);

        total_undmuts=0;
        for(count=0;count<sample_size*2-2;count++)
                total_undmuts += tree[count].undmuts;
        coatime = tree[2*sample_size-2].time;
/*  store the allele lengths and calculate the moments */
        for (j=0;j<sample_size;j++) st[j]=find_muts(tree+j,coatime,sample_size);
        mean=0.0;
        for(j=0;j<sample_size;j++) mean=mean+1.0*st[j]/sample_size;
        Var=VVVV=0.0;
        for(j=0;j<sample_size;j++) {
                Var += (st[j]-mean)*(st[j]-mean);
                VVVV += (st[j]-mean)*(st[j]-mean)*(st[j]-mean)*(st[j]-mean);
        }
/* calculate statistics for the run */
        n=1.0*sample_size;
        f1=(6*n-9)*Var*Var/(n*(n-1)*(n-2)*(n-3));
        f2=(n*n-2*n+3)*VVVV/((n-1)*(n-2)*(n-3));
        Gam4=f2-f1;
        f3=(n*n-3*n+3)*Var*Var/(n*(n-1)*(n-2)*(n-3));
        f4=VVVV/((n-2)*(n-3));
        Sig4=f3-f4;
        Var=Var/(sample_size-1);
/* output statistics for the run */
        summary->time = coatime;
        summary->undmuts += total_undmuts;
        summary->mean = mean;
        summary->var = Var;
        summary->test = 0.0;   /* user can calculate a statistic of choice */
/* requires sample_size>=20, and non-zero variance, for test to be applied */
```

```
        if (Var > 0.0 && Sig4!=0.0 && Gam4!=0.0 && n>=20 && fabs(2.45-
                Gam4/Sig4)<10000000.0) summary->k = 2.5*Sig4+0.28*Var-0.95/n-Gam4;
/*  median  */
        else summary->k = 0.0;
}

double find_median(double MM[10000],int number_samples,double percentile)
{
        int count,flag,in,im;
        double past,temp,min,temple;

        count=0;
        flag=1;
        past = -100000000.0;
        for (in=0;in<=number_samples;in++) if (flag) {
                min=1000000000.0;
                for (im=0;im<number_samples;im++) {
                        temp=MM[im];
                        if ((temp > past) && (temp <= min)) min=temp;
                        if (temp == past) {
                                count=count+1;
                                temple=MM[im];
                        }
                }
                if ((past == min) || (count > percentile*number_samples)) flag=0;
                past=min;
        }
        return (temple);
}

void main ()
{
        struct node *tree;
        struct stats summary;
        int sample_size,loci,toggle,i,j,k,kmax,pos,dropped,hist[20];
        int varq,bigreps,idum=(-13);
        double tgrowth,f1,theta,MK[10000],MG[10000],MVar[10000],G,K,aveVar,aveTime;
        double aveK,aveG,VarVar,tVar,G05,K05,max,thetatemp,thvar,a,c,tG,tKlow,tKeven;
        double store[8];

/* set parameters */
        varq=0;                                 /* if varq=0, no variation in mut. rate;
                                                   if varq=1, variation in the mut. rate;
                                                   if varq=2, variation in step size;
                                                   if varq=3, var. in propensity to
                                                   multi-step muts */
        printf("Type '0' for the Within-Locus Test and '1' for the Interlocus
Test:\n");
        scanf(" %d",&toggle);                   /* if toggle=0,
                                                   0-variance & low samp_size dropped */

        theta=3.5;                              /* pre-expansion theta */
        kmax=100;                               /* reps. of each parameter combination*/
        bigreps=8;                              /* no. of parameter combinations tried */
        loci=30;                                /* number of loci */
        sample_size=40;                         /* sample size */
        tgrowth=100.0;                          /* time of growth */
        f1=1.0;                                 /* factor of expansion */
        theta=theta*f1;                         /* use units of post-expan. pop size */
        tgrowth = tgrowth/(2*f1);               /* use units of post-expan. pop size */
        store[0]=0.00001; store[1]=0.01;        /* adjust parameters for the run */
        store[2]=0.02; store[3]=0.04;
        store[4]=0.08; store[5]=0.16;
        store[6]=0.32; store[7]=0.64;
        max=3.2;

/* output run parameters */
        printf("Pre-expansion Theta = %4.3lf  Sample Size = %d       Loci = %d\n",
                theta/f1,sample_size,loci);
        printf("Factor of Expansion = %2.1lf  Variance in Mut. Rate? (1=yes)
%d\n",f1,varq);
```

```
        if (toggle == 0) printf("Simulations are optimized for the Within-Locus
Test.\n\n");
        else printf("Simulations are optimized for the Interlocus Test.\n\n");

/* main simulation loop */
        fp_out = fopen("OutHist.txt","w");
        for (j=0;j<bigreps;j++) {
                aveTime=aveG=aveK=aveVar=0.0;
                for (i=0;i<20;i++) hist[i]=0;
                for (k=0;k<kmax;k++) MK[k]=MG[k]=0.0;
                for (k=0;k<kmax;k++) {
/*                      if (floor((k+1)/100.0)>=(k+1)/100.0) printf("k=%d\n",k+1);  */
                        tree= (struct node *)
                                malloc(2*sample_size*sizeof(struct node));
                        tVar=0.0;
                        dropped=pos=0;
                        for (i=0;i<loci;i++) MVar[i]=0.0;
                        for (i=0;i<loci;i++) {
                                if (varq==1) {
                                        thvar=store[j]*theta*theta;
                                        a=theta*theta/thvar;
                                        c=theta/thvar;
                                        thetatemp = (rndgamma(a,&idum))/c;
                                                /* mut. rate has gamma dist. */
                                }
                                else thetatemp=theta;
                                make_tree(tree,sample_size,&idum);
                                bottleneck(tree,sample_size,tgrowth,f1);
                        mutate_tree(tree,sample_size,&idum,thetatemp,store[j],varq);
                                calc_stats(tree,sample_size,&summary,&idum);
                                if (summary.k == 1.0*toggle) {
                                        /* is locus viable for within-test? */
/* WATCH OUT FOR INFINITE LOOP IF SAMPLE_SIZE<20 !! */
                                        i=i-1;
                                        ++dropped;
}
                                else {

                                        tVar=tVar+summary.var/loci;
                                        /* record variance */
                                        MVar[i]=summary.var;
                                        /* k statistic */
                                        aveTime=aveTime+summary.time/(loci*kmax);
                                        /* coalescence time */
                                        if (summary.k > 0.0) pos=pos+1;
                                        /* is k positive? */
                                }
                        }
/* calculate statistics for the run */
                        VarVar=0.0;
                        for (i=0;i<loci;i++)
                                VarVar=VarVar+(tVar-MVar[i])*(tVar-MVar[i])/(loci-1);
                        G=VarVar/(4*tVar*tVar/3+tVar/6);
                        K=1.0*pos/loci;
                        MG[k]=G;
                        MK[k]=K;
                        aveG=aveG+G/kmax;
                        aveK=aveK+K/kmax;
                        aveVar=aveVar+tVar/kmax;
                        if (dropped!=0)
                            printf("Cycle %d: %d dropped run(s).\n",k+1,dropped);
                        for (i=0;i<20;i++) if (G>=i*max/20.0 && G<(i+1)*max/20.0)
                                ++hist[i];
                }
/* calculate statistics over kmax replicates */
                G05=find_median(MG,kmax,0.05);
                K05=find_median(MK,kmax,0.05);
                tG=tKlow=tKeven=0.0;
                for (i=0;i<kmax;i++) {
                        if (MG[i]<0.333) tG=tG+1.0;
```

```
                        if (MK[i]<0.4) tKlow=tKlow+1.0;
                        if (MK[i]<=0.4) tKeven=tKeven+1.0;
                }
/* output results for each parameter combination */
                printf("Parameter (N.A. unless varq > 0) = %4.3lf    Within
        Interlocus\n",store[j],kmax);
                printf("5th Percentile Cutoffs:                                  g=%4.3lf
        k=%4.3lf\n",G05,K05);
                printf("Ratio of Deviation/NoDeviation Cutoffs:                  g=%4.3lf
        k=%4.3lf\n",K05/0.4,G05/0.333);
                printf("Probability of Type II Error:                    g=%4.3lf
        k=%4.3lf\n\n",tG/kmax,(tKlow+tKeven)/(2*kmax));
/* unbracket this, and it outputs a histogram (currently a g-value histogram) */
/*              for (i=0;i<20;i++) {
                        printf("%4.3lf-%4.3lf: ",i*max/20.0,(i+1)*max/20.0);
                        for (k=0;k<-0.5+80*hist[i]/kmax;k++) printf("=");
                        printf(" %d\n",hist[i]);
                }       */
                fprintf(fp_out,"\nLoci=%d     SampSize=%d
        RunReplicates=%d\n",loci,sample_size,kmax);
                fprintf(fp_out,"Theta=%4.3lf  GrowthFactor=%4.3lf    Param=%4.3lf\n",
                        theta/f1,f1,store[j]);
                fprintf(fp_out,"runsum:       %5.4lf  %5.4lf  %5.4lf  %5.4lf  %5.4lf\n",
                        store[j],K05/0.4,(tKlow+tKeven)/(2*kmax),G05/0.333,tG/kmax);

        }
        fclose(fp_out);
}
```

(B) Code for SNP-based tests of demography

This simulation generates SNP allele frequency distributions for various demographic scenarios, and obtains significance cutoffs for the test of expansion. The coalescent algorithm itself is nearly identical to that described in part A. The simulation will eventually be posted in electronic form at the following URL: **Error! Bookmark not defined.**.

```
/* WebPageBiallelicsSim */
/* flexible biallelic coalescent simulation */
/* August 30, 1999 */
/* by David Reich */

#include <stdio.h>
#include <stdlib.h>
#include <math.h>

#define REPS 1000                       /* repetitions of the simulation */

/* universal declarations of variables, output files, functions, and subroutines */
struct node {
        int id;
```

```
        int state;
        struct node *desc1;
        struct node *desc2;
        struct node *ancestor;
        double time;
        } ;
int all[3],mutct[3];
double brnlen;

FILE *fp1;

double ran1();
void make_tree();
void bottleneck();
void mutate_tree1();
void mutate_tree2();

/* random number generator on a uniform distribution (from Numerical Recipes in C) */
double ran1(int *idum)
        #define IA 16807
        #define IM 2147483647
        #define AM (1.0/IM)
        #define IQ 127773
        #define IR 2836
        #define NTAB 32
        #define NDIV (1+(IM-1)/NTAB)
        #define EPS 1.2e-7
        #define RNMX (1.0-EPS)
{
        int j;
        int k;
        static int iy=0;
        static int iv[NTAB];
        double temp;

        if (*idum <= 0 || !iy) {
                if (-(*idum) < 1) *idum=1;
                else *idum = -(*idum);
                for (j=NTAB+7;j>=0;j--) {
                        k=(*idum)/IQ;
                        *idum=IA*(*idum-k*IQ)-IR*k;
                        if (*idum < 0) *idum += IM;
                        if (j < NTAB) iv[j] = *idum;
                }
                iy=iv[0];
        }
        k=(*idum)/IQ;
        *idum=IA*(*idum-k*IQ)-IR*k;
        if (*idum < 0) *idum += IM;
        j=iy/NDIV;
        iy=iv[j];
        iv[j]=*idum;
        if ((temp=AM*iy) > RNMX) return RNMX;
        else return temp;
}

/* constructs evolutionary tree (adapted from Hudson's 1991 coalescent simulation */
void make_tree(struct node *tree, int q, int *idum)
{
        double ran1(int* idum);
        int in, pick;
        double t;
        struct node **list;

/* initializations */
        list = (struct node **)malloc( q*sizeof(struct node *) );
        for(in=0; in<q; in++) {
                tree[in].time = 0.;
                tree[in].desc1 = tree[in].desc2 = 0;
                list[in] = tree + in ;
```

```
        }
        for (in=0;in<=2*q-2;in++) tree[in].id=in;
/* generate the times of the nodes */
        t=0.;
        for(in= q; in>1; in--) {
                t += -2.0 * log(1.-ran1(idum)) / ( ((double)in)*(in-1) );
                tree[2*q - in].time = t;
        }
/* generate the topology of the tree */
        for (in=q; in>1; in--){
                pick = (int)floor(in*ran1(idum));
                list[pick]->ancestor=tree+2*q-in;
                tree[2*q-in].desc1=list[pick];
                list[pick]=list[in-1];
                pick = (int)floor((in-1)*ran1(idum));
                list[pick]->ancestor = tree + 2*q - in;
                tree[2*q - in].desc2 = list[pick];
                list[pick] = tree + 2*q - in;
        }
        free(list);
}

/* rescales the tree in order to simulate a history of sudden growth or contraction.
   Adapted from Hudson's 1991 coalescent simulation. */
 void bottleneck(struct node *tree,int q,double t2,double f2)
{
        int in;

        for (in=q;in<2*q-1;in++) if (tree[in].time>t2) tree[in].time=t2+(tree[in].time-
t2)*f2;
}

/* places a mutation on the gene tree, weighting the probability of a
   mutation by the length of the branch */
void mutate_tree1(struct node *tree,int *idum,int M,int N)
{
        int in,save,tp,j;
        double temp,runsum,temptime,tpt2;
        double ran1(int *idum);

        for (in=2*N-2;in>=0;in--) tree[in].state=0;
        brnlen=0.0;     for (in=0;in<=2*N-3;in++)
                                brnlen=brnlen+tree[in].ancestor->time-tree[in].time;
        temp=ran1(idum);        runsum=0.0;
        for (in=0;in<=2*N-3;in++) {
                if (temp>=runsum && temp<runsum
                        +(tree[in].ancestor->time-tree[in].time)/brnlen)
                        { tree[in].state=1; save=in; in=2*N-3; }
                runsum=runsum+(tree[in].ancestor->time-tree[in].time)/brnlen;
        }
        if (tree[save].time>0) for (in=0;in<N;in++) {
                tp=in;
                temptime=tree[save].time;
                tpt2=tree[tp].time;
                do { tp=tree[tp].ancestor->id;} while (tree[tp].time<tree[save].time);
                if (tree[tp].time==tree[save].time) tree[in].state=1;
        }
        for (j=0;j<3;j++) mutct[j]=0;
        for (in=0;in<M;in++) if (tree[in].state==1)  ++mutct[0];
        for (in=M;in<N;in++) if (tree[in].state==1) ++mutct[1];
        for (in=0;in<N;in++) if (tree[in].state==1) ++mutct[2];
        if (mutct[0]==0 || mutct[0]==M) all[0]=1; else all[0]=2;
        if (mutct[1]==0 || mutct[1]==N) all[1]=1; else all[1]=2;
        if (mutct[2]==0 || mutct[2]==N) all[2]=1; else all[2]=2;
}

/* THIS IS THE OLD WAY OF DOING MUTATIONS; IT WORKS FINE, AND IS JUST SLOW... */
void mutate_tree2(struct node *tree,int *idum,int M,int N,double theta)
{
        int in,tpid,flag,sum,mutid;
```

```
        double temp,inc,runsum;

        brnlen=0.0;     for (in=0;in<=2*N-3;in++)
                                brnlen=brnlen+tree[in].ancestor->time-tree[in].time;
        if (ran1(idum)<1-exp(-theta*brnlen/2)) {
                flag=0;temp=ran1(idum); runsum=0.0; in=N;
                while (flag==0) {
                        inc=(tree[in].time-tree[in].desc1->time)/brnlen;
                        if (temp>=runsum && temp<runsum+inc)
                                { mutid=tree[in].desc1->id; flag=1; }
                        runsum=runsum+inc;
                        inc=(tree[in].time-tree[in].desc2->time)/brnlen;
                        if (temp>=runsum && temp<runsum+inc)
                                { mutid=tree[in].desc2->id; flag=1; }
                        runsum=runsum+inc;
                        ++in;
                }
                for (in=0;in<N;in++) tree[in].state=0;
                for (in=0;in<N;in++) {
                        tpid=in;
                        while (tpid<mutid) tpid=tree[tpid].ancestor->id;
                        if (tpid==mutid) tree[in].state=1;
                }
                sum=0; for (in=0;in<M;in++) sum=sum+tree[in].state; mutct[0]=sum;
                if (tree[in].state==0 || tree[in].state==M) all[0]=1; else all[0]=2;
                sum=0; for (in=M;in<N;in++) sum=sum+tree[in].state; mutct[1]=sum;
                if (tree[in].state==0 || tree[in].state==N-M) all[1]=1; else all[1]=2;
                sum=0; for (in=0;in<N;in++) sum=sum+tree[in].state; mutct[2]=sum;
                if (tree[in].state==0 || tree[in].state==N) all[2]=1; else all[2]=2;
        }
        else all[0]=all[1]=all[2]=1;
}

void main ()
{
        struct node *tree;
        int idum=(-5),l,loci,i,k,ct,cut,freq,run1s,highs,mutproc,know;
        int M,N,histmode,inp,ctcuts;
        double theta,f2,t2,avelen,mean,bigmean;
        double hist[1000],oldhist[1000],store[REPS],factor,obs;

/* set parameters */
        loci=1;
        know=0;
        histmode=1;
        mutproc=0;      /* mutproc = 0 means that mutations are of fast type */
        theta=30.05;    /* usual population genetic parameter theta */
        factor=0.01;    /* factor of expansion */
        t2=0.0001;      /* time of expansion in units of current population size */
        N=8;
        M=0;            /* ascertainment sample; if 0, indicates no presampling */
/*      printf("SampleSize=%d AscertainmentSample=%d RepsPerRun=%d\n",N,M,REPS);
        printf("Expansion Factor = %2.1lf     ExpansionTime=%5.4lf\n",1/factor,t2); */
/* query the user */
        printf("Type (1) for a significance level, (2) for a freq. distribution:   ");
        scanf("%d",&histmode);
        inp=-5; do {
                printf("\nType a number to change the settings:\n");
                if (histmode==1) printf("  (0) simulation generates significance
levels\n");
                else printf("  (0) simulation generates allele freq. distributions
(output to Histogram.txt)\n");
                if (know==0) printf("  (1) frequency of rarer allele is used\n");
                else printf("  (1) frequency of the mutant allele is used\n");
                printf("  (2) sample size = %d\n",N);
                if (histmode==1) printf       ("  (3) %d SNP loci in sample\n",loci);
                if (M==0) printf("  (4) no presampling\n");
                else printf("  (4) presampling with %d samples\n",M);
                printf("Type 6 if the settings are OK...     ");
                scanf("%d",&inp);
```

```
                if (inp==0) histmode=3-histmode;
                if (inp==1) know=1-know;
                if (inp==2) { printf("Input the number of chromosomes:      ");
                        scanf("%d",&N); }
                if (inp==3) { printf("Input the number of SNPs:      ");
                        scanf("%d",&loci); }
                if (inp==4) { printf("Input the number of presamples:       ");
                        scanf("%d",&M); }
        } while (inp!=6);
        if (histmode==1) {
                printf("Input observed mean frequency (mean no. of copies of allele):
        ");
                scanf("%lf",&obs);
        }
        else loci=1;
        printf("\n");

/* main simulation loop */
        fp1 = fopen("Histogram.txt","w");
        if (histmode==2) fprintf(fp1,"sample size = %d      prescreening sample = %d
                repetitions = %d\n\n",N,M,REPS*loci);
        for (k=0;k<histmode;k++) {
                if (k==0) f2=1.0; else f2=factor;
                bigmean=avelen=brnlen=0.0;
                ctcuts=ct=runls=highs=0;
                for (i=0;i<N;i++) hist[i]=0.0;
                for (i=0;i<REPS;i++) {
                        mean=0.0;
                        for (l=0;l<loci;l++) {
                                tree= (struct node *) malloc(2*N*sizeof(struct node));
                                make_tree(tree,N,&idum);
                                bottleneck(tree,N,t2,f2);
                                if (mutproc==0) mutate_tree1(tree,&idum,M,N);
                                else mutate_tree2(tree,&idum,M,N,theta);
                                if (M==0 && all[2]==2) cut=1;
                                else if (M>0 && all[2]==2 && all[0]==2) cut=1;
                                else cut=0;
                                if (cut==1 && mutct[2]>0 && N-mutct[2]>0) {
                                        if (l==0 && ((int)i/100.0 ==
floor((int)i/100.0))) printf("%d cycles completed / %d total\n",i,REPS);
                                        freq=mutct[2];
                                        if (mutproc==0) {
                                                hist[freq]=hist[freq]+
                                                        1-exp(-theta*brnlen/2);
                                                avelen=avelen+
                                                        (1-exp(-theta*brnlen/2))/REPS;
                                        }
                                        else { hist[freq]=hist[freq]+1.0;
                                                avelen=avelen+1.0/REPS; }
                                if (know==0) { if (freq<=N/2.0) mean=mean+1.0*freq/loci;
                        else mean=mean+1.0*(N-freq)/loci;
                                        else mean=mean+1.0*freq/loci;
                                        ++ct;
                                }
                                else { l=l-1; ++runls; if (all[2]>2) ++highs; }
                        }
                        store[i]=mean;
                        bigmean=bigmean+mean/REPS;
                        if (mean<obs) ++ctcuts;
                }
/* output allele frequency distribution, and Figure 2 plot, to "Histogram.txt" */
                if (histmode==2) {
                        for (i=1;i<=N-1;i++) hist[i]=hist[i]/(REPS*avelen);
                        if (k==0) {
                                fprintf(fp1,"\nallele frequency distribution for a
constant-sized population\nfreq.      prob.\n",
                                        N,M,REPS*loci);
                                if (know==0) for (i=1;i<=N/2.0;i++) {
                                        if (i<N/2.0) fprintf(fp1,"%d
        %4.3lf\n",i,hist[i]+hist[N-i]);
```

```
                                        else fprintf(fp1,"%d   %4.3lf\n",i,hist[i]);
                                }
                                if (know==1) for (i=1;i<=N-1;i++)
                                        fprintf(fp1,"%d       %4.3lf\n",i,hist[i]);
                                for (i=1;i<=N-1;i++) oldhist[i]=hist[i];
                        }
                        if (k==1) {
                                fprintf(fp1,"\nratio of observed to expected value for
%1.0lf-fold growth %5.4lfN gens. ago\nfreq.  ratio\n",
                                        1/factor,t2*2);
                                if (know==0) for (i=1;i<=N/2.0;i++) {
                                        if (i<N/2.0) fprintf(fp1,"%d
        %3.2lf\n",i,(hist[i]+hist[N-i])/(oldhist[i]+oldhist[N-i]));
                                        else fprintf(fp1,"%d
        %3.2lf\n",i,hist[i]/oldhist[i]);
                                }
                                if (know==1) for (i=1;i<=N-1;i++) fprintf(fp1,"%d
        %3.2lf\n",i,hist[i]/oldhist[i]);
                        }
                }
        }
        printf("ctcuts=%d\n",ctcuts);
        fclose(fp1);
```