Maximum likelihood estimation of population growth rates based on the coalescent

Mary K. Kuhner*, Jon Yamato, and Joseph Felsenstein†

Department of Genetics, University of Washington

Seattle, WA 98195, USA

*Internet address *mkkuhner@genetics.washington.edu*

†Internet address *joe@genetics.washington.edu*

Running head: Estimating growth rates

Keywords: coalescent, growth, Metropolis-Hastings, maximum likelihood

Corresponding author:

Mary K. Kuhner

University of Washington

Department of Genetics

Box 357360

Seattle, WA 98195-7360, USA

Phone (206) 543-8751

FAX (206) 543-0754

Internet *mkkuhner@genetics.washington.edu*

ABSTRACT

We describe a method for co-estimating $4N_e\mu$ (four times the product of effective population size and neutral mutation rate) and population growth rate from sequence samples using METROPOLIS-HASTINGS sampling. Population growth (or decline) is assumed to be exponential. The estimates of growth rate are biased upwards, especially when $4N_e\mu$ is low; there is also a slight upwards bias in the estimate of $4N_e\mu$ itself due to correlation between the parameters. This bias cannot be attributed solely to METROPOLIS-HASTINGS sampling but appears to be an inherent property of the estimator, and is expected to appear in any approach which estimates growth rate from genealogy structure. Sampling additional unlinked loci is much more effective in reducing the bias than increasing the number or length of sequences from the same locus.

INTRODUCTION

The genealogical structure of a sample from a population contains information about that population's history. The distribution of coalescence times (times at which two of the sampled individuals have a common ancestor) depends on the effective population size N_e: in a diploid population the distribution is proportional to $4N_e$. Since coalescence times cannot be directly observed in most cases, but only inferred from the accumulation of mutations, we rescale time proportional to the per-site neutral mutation rate μ. Thus, though we cannot estimate $4N_e$ directly, we can estimate the product $4N_e\mu$ which we will call Θ.

If the population size has changed over time, the distribution of coalescence times will differ from its expectation in a population where Θ is constant, and in principle this should be detectable. In particular, if the population has been growing the most rootward branches will be relatively short, whereas if it has been shrinking the most rootward branches will be relatively long.

We have previously described a method for estimating Θ in a population of constant size (KUHNER et al. 1995), using METROPOLIS-HASTINGS sampling (METROPOLIS et al. 1953, HASTINGS 1970) of genealogies. The basic strategy is to sample genealogies based on their posterior probability with regard to the data and a trial value of Θ, and then use the sampled genealogies to evaluate the relative likelihood of other values of Θ. This importance sampling approach concentrates the sampled genealogies in regions of high posterior probability, which is much more efficient than using random genealogies, and avoids the bias of using only a single genealogy reconstruction. This algorithm is implemented in our COALESCE program.

In this paper we extend the METROPOLIS-HASTINGS genealogy sampling approach to the case of a population experiencing exponential growth or decline. In this case population size is represented by two parameters: the exponential growth rate g and the present-day value of Θ (that is, the value at the time when the organisms were sampled). The parameters are not independent: the more rapidly a population has grown, the larger its current size is expected to be compared to its "average" size. We have written a program, FLUCTUATE, which implements this sampler.

Both analytic and simulation results show that the estimate of the growth rate g is biased upwards when a finite number of individuals are analyzed. At least two factors are at work in this bias: the non-linear relationship between coalescence times and the estimate of g, and truncation of the coalescent distribution, in genealogies of finite numbers of individuals, by the bottommost coalescence. There is also a smaller upwards bias in Θ due to the correlation between the two parameters. The bias in these estimators can most effectively be reduced by sampling multiple loci.

The method proposed by GRIFFITHS and TAVARÉ (1994) for estimation of growth rate uses a different strategy for defining and sampling genealogies, but shares a common mathematical rationale. It should therefore experience the same bias. Further testing will be required to compare the effectiveness of these two methods. Other approaches to estimating growth, such as the pairwise measures of SLATKIN and HUDSON (1991) and ROGERS and HARPENDING (1992), use less of the information present in the data and should be less efficient (FELSENSTEIN 1992); the genealogical methods are at a particular advantage when the growth rate g is low or negative, a case in which pairwise methods tend to fail due to the confounding influence of the genealogical structure (SLATKIN and HUDSON 1991).

METHODS

The METROPOLIS-HASTINGS genealogy sampler for constant-sized populations (KUHNER et al. 1995) works by a two-phase process. It begins with an initial genealogy and an initial value of Θ, called Θ_0. In the first phase, a new genealogy is created by locally rearranging the previous genealogy in proportion to the coalescent prior probability $P(G|\Theta_0)$ (given by KINGMAN 1982a, 1982b). In the second phase, this genealogy is accepted or rejected based on $P(D|G)$, the probability of the sequence data on the genealogy. This is equivalent to sampling from the posterior probability, which is proportional to $P(G|\Theta_0)P(D|G)$. This process is repeated, with samples taken from it at intervals to produce a set of genealogies from which a maximum likelihood of Θ can be made. The estimate is most efficient when Θ_0 is close to Θ, so it is useful to run several iterations of the sampler, using the estimated Θ of each iteration as the starting Θ_0 of the next.

Like most calculations involving the coalescent, these equations hold exactly only in the limit as the population size N goes to infinity: in practice the approximation involved should be insignificant as long as the number of individuals sampled is less than the square root of the population size.

Mutational model: We used the DNA/RNA sequence model of FELSENSTEIN (1981) which allows unequal base frequencies and transition/transversion bias, extended as in FELSENSTEIN and CHURCHILL (1996) to allow for variable rates among sites and auto-correlation of those rates. It is simple to substitute any other mutational model for which $P(D|G)$ can be calculated: for example, models appropriate to protein or microsatellite data. The algorithm as designed does not estimate parameters of the mutational model.

Scaling for population growth: When the size of the population changes exponentially through time, the coalescent prior becomes $P(G|\Theta_0, g_0)$ where g_0 is a trial value of the exponential growth rate g. (Positive values of g indicate population growth, and negative values indicate decline.) The units of g are $1/\mu$ generations.

In order to sample coalescence times from this prior, we use a time rescaling under which it becomes identical to the simpler constant-population prior. Time is scaled proportional to growth, so that the same expected amount of coalescence occurs in one unit of time regardless of population size. Under this transformation, the coalescent structure of the genealogy becomes identical to the constant-population expectations.

The rescaled time T is derived from the original time t by the following relation (SLATKIN and HUDSON 1991). The negative sign in the exponent is due to the fact that we are considering times previous to the present.

$$T = \frac{1}{g}(1 - e^{-gt})$$

This rescaled time is then substituted for ordinary time in constructing rearrangements of the genealogy. In cases where g is less than zero, some proportion of the rescaled times will correspond to infinite ordinary time. Our implementation rejects genealogies which contain infinite times, on the grounds that their likelihood for biologically reasonable data will tend to be very small. An upwards bias may be created by this procedure, but in practice it should be trivial.

A series of genealogies generated under a given Θ_0 and g_0 can be used to determine the likelihood $L(\Theta, g)$ for other values of Θ and g. For each genealogy G a product is taken over all coalescence intervals i: in each interval, k is the number of lineages in the genealogy during that interval, t_s is the time at the tipward end, and t_e is the time at the rootward end. Note that these are not rescaled times.

$$P(G|\Theta, g) = \prod_i \frac{e^{gt_e} e^{\frac{k(k-1)}{\Theta g}(e^{gt_s} - e^{gt_e})}}{\Theta}$$

This formula can be show to be equivalent to that given in GRIFFITHS and TAVARÉ 1994, bearing in mind that they scaled time in units of N generations rather than $\frac{1}{\mu}$ generations and they considered a haploid rather than a diploid case: they also retained some combinatorial constants which we omit, since we are concerned only with ratios of probabilities.

This probability is then corrected for the importance sampling function $P(D|G)P(G|\Theta_0, g_0)$ (where n is the number of genealogies sampled):

$$\frac{L(\Theta, g)}{L(\Theta_0, g_0)} = \frac{1}{n} \sum_G \frac{P(G|\Theta, g)}{P(G|\Theta_0, g_0)}$$

(The terms $P(D|G)$ drop out as they are the same for all values of Θ and g.)

The maximum of this function, which is a joint maximum likelihood estimate of Θ and g, can be found by standard methods. Technical difficulties are often encountered due to arithmetic overflow in exponentiation and the characteristic curving-ridge shape of the likelihood surface.

Multiple loci: The likelihoods can be multiplied together across unlinked loci to generate an overall multi-locus likelihood. Doing so should greatly improve the efficiency of the estimate, especially for g, since doubling the number of loci doubles the amount of information available about the most rootward parts of the genealogy (which are the most informative for growth rate, since they represent the population size most divergent from the modern-day size). Adding additional sequences mainly adds information about the most tipward parts of the genealogy, which contain relatively little information about growth.

If the loci to be combined cannot be assumed to have the same values for the parameters, this must be taken into account when combining them. It is reasonable to assume that the population growth rate affects all loci equally (barring selection), but both the neutral mutation rate μ and the effective population size N_e can vary among loci (for example, N_e is lower for a mitochondrial locus than for a nuclear one).

This can easily be accomodated if the relative values of the parameters for different loci are known (or can be assumed): we simply replace N_e and μ with appropriate locus-dependent functions when calculating the multi-locus likelihood. In the future, a method for dealing with unknown variability in μ among loci could be developed by assuming Gamma distributions for the parameters and integrating over the range of possibilities.

Assessing the accuracy of the estimate: An advantage of likelihood methods is that information about the accuracy of the estimate can be gleaned from the likelihood curve. We will consider the confidence interval as the set of all parameter values which would not be rejected (via a likelihood ratio test) at the given level. Asymptotically, as the number of loci approaches infinity, the shape of the likelihood curve becomes Gaussian (normal) and we can construct a variance for it using a χ^2 metric with two degrees of freedom (Cox and Hinkley 1974). Using this approach, the area of the parameter space in which the log likelihood is no more than three units below the maximum can be taken as a rough 95% confidence interval.

Such confidence intervals will be approximate at best for finite numbers of loci. It is not obvious a priori whether bias present in the maximum likelihood parameter estimates will also strongly affect the confidence intervals. We have not solved this problem analytically, but we can assess the usefulness of the approximate confidence intervals by simulation.

Simulation procedures. Each simulation consisted of 100 replicates. Genealogies of 25 sequences were randomly generated according to given values of Θ and g, and DNA data were generated randomly from these genealogies using a Kimura 2-parameter model (Kimura 1980) with a transition/transversion ratio of 2.0. In the following description a "step" is the construction of a single genealogy; a "chain" is a set of such genealogies used to make a parameter estimate, which can then be used as to set initial parameters for the following chain. For both the exponential-growth program Fluctuate and our constant population size program Coalesce (used for comparison), we used the following search strategy: for each locus, 10 short chains of 1000 steps each were run, followed by 2 long chains of 15,000 steps each, sampling every 20th step. We provided the programs with the correct transition/transversion ratio. For initial estimates of Θ we used Watterson's estimate (Watterson 1975); for initial estimates of g we arbitrarily chose 1.0. Initial genealogies were generated using Phylip programs (Felsenstein 1993, version 3.5c): Dnadist to produce corrected distances from the sequence data, and NEIGHBOR to generate UPGMA genealogies from these distances.

We also performed simulations in which we made maximum likelihood estimates assuming that the true genealogy was known without error. This is equivalent to using infinitely long sequences, since with such sequences the METROPOLIS-HASTINGS sampler should unerringly generate the true genealogy. We have called these results "infinite sites" in the Tables.

For each estimation, we noted whether or not the log likelihood for the true Θ and g was within 3 units of the log likelihood at the maximum–i.e., whether or not the truth could be rejected at the approximate 95% level.

RESULTS

Table 1 shows results from simulation tests of FLUCTUATE. We do not present results for the case of $\Theta = 0.01$, $g = 100$ with finite numbers of sites because data sets simulated at these values frequently contained no variable sites. On theoretical grounds we expect an invariant data set to produce a zero estimate of Θ and an indeterminate estimate of g (all values are equally likely).

Cases where g is negative entail the possibility that infinite time will be required for coalescence when simulating the genealogy. The probability that this will happen depends on the product of Θ and g. In practice, the case of $\Theta = 0.01$, $g = -10.0$ could be simulated (less than 1% failure to coalesce), but with $\Theta = 0.1$ a substantial fraction of simulated genealogies failed to coalesce in finite time, and so no results are presented for this case.

In general, estimates of g showed a strong upwards bias, decreasing somewhat with number of sites and more markedly with number of loci. The only exception was the case of $\Theta = 0.1, g = 100$ where the estimates appear biased downwards with finite amounts of data, possibly due to saturation of variable sites. The standard deviation of g was much less for high true values of Θ than for low ones, even with infinite numbers of sites.

Estimates of Θ also tended to be biased upwards, in contrast to the constant-population case, in which they appear nearly unbiased (KUHNER et al. 1995).

With few exceptions, doubling the number of loci was more effective in reducing bias and standard deviation than doubling the number of sites.

In most cases the true values of Θ and g were rejected at the 95% level slightly more often than the desired 5%.

Table 2 shows comparable results, for the case in which the true g was zero, from the program COALESCE (KUHNER et al. 1995) which uses a similar METROPOLIS-HASTINGS strategy but does not allow changes in population size. Examination of the results suggests that adding growth as a parameter approximately doubles the standard deviation of Θ.

DISCUSSION

Why is the estimate of g biased? We have identified two processes that contribute to this bias. Both are intrinsic to the estimation of exponential growth from genealogical data, and are not due to the METROPOLIS-HASTINGS sampler itself: they can be shown in simple cases that do not require any of the METROPOLIS-HASTINGS machinery.

One component of the bias results from the non-linear relationship between the coalescence times and the estimate of g. A simple two-sequence case provides a concrete demonstration. In genealogies of two tips where the true growth rate is zero and Θ is known without error, the distribution of the coalescence time t follows directly from coalescent theory (KINGMAN 1982a, b). Centiles of this distribution can then be used to make a distribution of $\hat{g}$ values (Table 3). The distribution of $\hat{g}$ is highly skewed, with a mean far above the true value. Essentially, the non-linear relationship between t and $\hat{g}$ transforms variance in t into bias in $\hat{g}$. Thus, bias is expected not only in our method but in any method that uses t (or measurements depending on it, such as number of mutations) as a basis for estimating exponential growth. For example, the star-phylogeny method of SLATKIN and HUDSON (1991), which counts variable sites, shows a similar upwards bias; we have confirmed this in simulation tests (data not shown).

However, even in the absence of variability in coalescence times some bias is present. Table 4 shows results based on analysis of a "perfect" coalescent genealogy, in which each interval has exactly its expected length; there is no variance in t. A bias is clearly visible in Table 4, although the 95% confidence intervals do include the true value. This component of the bias results from the fact that any genealogy with finite tips truncates the distribution of coalescence times; it has a "final coalescence" at the root, prior to which no further information is available. This presents likelihood estimation with an attractive hypothesis involving a population bottleneck at the time of the final coalescence; such a hypothesis has high likelihood because it maximizes the probability of the final event. Attraction towards this degenerate hypothesis produces a bias in $\hat{g}$.

Correctness of the sampler: It is difficult to prove a complex computer program correct, but we tested FLUCTUATE in several ways to help assure ourselves that the observed bias was not due to program error. If the sampler is run with 100% acceptance (that is, the data are ignored and every proposed genealogy accepted) the genealogies produced should be an autocorrelated but otherwise random sample from a coalescent distribution with the given Θ and g. We examined large samples of such genealogies and found them consistent with the random coalescent (data not shown).

We also tested the sampler with $g = 0$ and found its results substantively identical to our previous program COALESCE which dealt with the constant-population case (data not shown). Based on these tests, we believe the sampler to be correct. In any case, as is shown in Tables 3 and 4, bias would be expected in a perfectly functioning sampler.

Overcoming bias: Given that this method (and other methods involving use of t to estimate g) has bias, how can the most accurate results be obtained? Tables 3 and 4 show clearly that adding additional sites or sequences is ineffectual, whereas adding additional unlinked loci rapidly reduces the bias. Each new locus will provide additional information about the region of the early branchings, thereby fleshing out this part of the distribution, and the independent variation in coalescence times among loci helps counteract the bias introduced by non-linearity.

It appears that the small bias seen in Θ is a consequence of correlation between Θ and g, since it does not appear when g is held constant at zero (as in COALESCE). One positive aspect of these findings is that it is quite possible to estimate current Θ accurately even if the population has been growing or shrinking; the bias in Θ is small even when g is far from zero.

Future directions. Real biological populations often grow or decline in ways more complicated than simple exponential growth, but the bias in the estimator interferes with attempts to fit more complex models. For example, one could imagine fitting a two-stage model with exponential growth followed by a steady-state period; however, because of the sparseness of the rootward part of the genealogy this model would be attracted to wrong solutions featuring very rapid early growth. It is possible that using a sufficiently large number of loci would allow such models to work.

Since relatively little power is available for estimating growth, attempting to differentiate between different models of growth (for example, exponential versus geometric or linear) are unlikely to succeed with reasonably sized data sets. In principle, however, this method could accomodate any growth model for which the time transformation can be worked out.

The algorithm can readily be adapted to data types other than nucleotide sequence data, such as protein sequences, allozyme alleles, or restriction site polymorphisms, as long as an appropriate evolutionary model is available.

It is possible to extend this family of algorithms by including recombination, which will greatly facilitate the analysis of nuclear loci. This may also allow a single long locus to provide some of the advantages of multiple loci, since recombination turns the single genealogy into several partially correlated genealogies. However, the algorithm with recombination will be technically challenging due to the more complex data structures and rearrangement scheme required. GRIFFITHS and MARJORAM (1996) have developed an alternative approach to genealogical sampling in the presence of recombination, which is also computationally demanding: it will be interesting to compare these approaches in the future.

Availability of software: The METROPOLIS-HASTINGS Monte Carlo algorithm described here is available from the authors as program FLUCTUATE in the package LAMARC, which uses the same input/output formats as the PHYLIP package. The program is written in C and can be obtained by anonymous ftp from `evolution.genetics.washington.edu` in directory pub/lamarc or via the World Wide Web at `http://evolution.genetics.washington.edu/lamarc.html`.

ACKNOWLEDGMENTS

We thank MONTY SLATKIN and SIMON TAVARÉ for helpful discussion and PETER BEERLI for assistance in finding maxima of likelihood surfaces. We also thank the Organizing Committee of the 4th annual meeting of the Society of Molecular Biology and Evolution for inviting the first author to a highly productive meeting.

This research was supported by National Science Foundation grants BSR-8918333 and DEB-9207558 and National Institute of Health grant 2-R55GM41716-04 (all to J. F.).

LITERATURE CITED

COX, D. R., and D. V. HINKLEY, 1974 Theoretical Statistics, p. 314. Chapman and Hill, London.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Molecular Evolution **17:** 368-376.

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139-147.

FELSENSTEIN, J., 1993 PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

FELSENSTEIN, J., and G. CHURCHILL, 1996 A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13:** 93-104.

GRIFFITHS,, R. C., and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3:** 479-502.

GRIFFITHS, R. C., and S. TAVARÉ, 1994 Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B **344:** 403-410.

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

KIMURA, M., 1980 A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111-120.

KINGMAN, J. F. C., 1982a The coalescent. Stochastic Processes and Their Applications **13:** 235-248.

KINGMAN, J. F. C., 1982b On the genealogy of large populations. J. Applied Prob. **19A:** 27-43.

KUHNER, M., J. YAMATO, and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421-1430.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, and E. TELLER, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087-1092.

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 R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555-562.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276.

Table 1: FLUCTUATE simulation results

Estimates of Θ and g based on 100 simulated data sets each, with 25 sequences of the given number of base pairs. Columns headed $bp = \infty$ were created by assuming that the genealogy could be reconstructed without error. Table 1E shows the number of times that the true values of Θ and g could be rejected at the nominal 95% level, out of 100 data sets. nd, not determined.

A. Estimate of Θ

	Loci	$\Theta = 0.01$			$\Theta = 0.1$		
		bp=500	bp=1000	bp=∞	bp=500	bp=1000	bp=∞
$g = -10$	1	0.014	0.013	0.010	nd	nd	nd
	2	0.012	0.011	0.010	nd	nd	nd
$g = 0$	1	0.013	0.012	0.011	0.112	0.107	0.113
	2	0.012	0.011	0.010	0.104	0.106	0.104
$g = 100$	1	nd	nd	0.011	0.097	0.092	0.110
	2	nd	nd	0.011	0.103	0.097	0.106

B. Standard deviation of Θ

	Loci	$\Theta = 0.01$			$\Theta = 0.1$		
		bp=500	bp=1000	bp=∞	bp=500	bp=1000	bp=∞
$g = -10$	1	0.009	0.006	0.002	nd	nd	nd
	2	0.004	0.003	0.002	nd	nd	nd
$g = 0$	1	0.009	0.004	0.002	0.032	0.027	0.028
	2	0.004	0.003	0.002	0.023	0.017	0.018
$g = 100$	1	nd	nd	0.003	0.043	0.038	0.031
	2	nd	nd	0.002	0.033	0.029	0.021

C. Estimate of g

	Loci	$\Theta = 0.01$			$\Theta = 0.1$		
		bp=500	bp=1000	bp=∞	bp=500	bp=1000	bp=∞
$g = -10$	1	257.3	165.8	50.0	nd	nd	nd
	2	130.2	71.3	27.8	nd	nd	nd
$g = 0$	1	360.2	145.9	45.1	14.6	6.2	12.7
	2	128.4	82.1	46.8	5.1	6.2	5.3
$g = 100$	1	nd	nd	227.4	73.6	69.7	119.1
	2	nd	nd	187.1	53.4	52.7	110.2

D. Standard deviation of g

	Loci	$\Theta = 0.01$			$\Theta = 0.1$		
		bp=500	bp=1000	bp=∞	bp=500	bp=1000	bp=∞
$g = -10$	1	567.1	286.6	116.9	nd	nd	nd
	2	463.2	149.2	67.8	nd	nd	nd
$g = 0$	1	1215.1	248.3	95.3	19.3	15.1	15.7
	2	298.5	144.6	88.1	8.1	10.2	8.5
$g = 100$	1	nd	nd	214.8	73.6	69.7	45.9
	2	nd	nd	144.7	53.4	52.7	27.8

E. Number of samples (out of 100) in which the true values were rejected at the 95% level

	Loci	$\Theta = 0.01$			$\Theta = 0.1$		
		bp=500	bp=1000	bp=∞	bp=500	bp=1000	bp=∞
$g = -10$	1	15	6	1	nd	nd	nd
	2	7	12	2	nd	nd	nd
$g = 0$	1	16	10	2	2	4	3
	2	13	7	5	9	1	3
$g = 100$	1	nd	nd	6	6	5	8
	2	nd	nd	7	7	4	3

Table 2: Coalesce (constant-population) simulation results

Estimates of Θ and g based on 100 simulated data sets each, with 25 sequences of the given number of base pairs. SD, standard deviation.

A: Low Θ (0.01), low g (0)

	$\hat{\Theta}$		SD of Θ	
	500 bp	1000 bp	500 bp	1000 bp
1 locus	0.0097	0.0099	0.0042	0.0034
2 loci	0.0102	0.0101	0.0028	0.0025

B: High Θ (0.1), low g (0)

	$\hat{\Theta}$		SD of Θ	
	500 bp	1000 bp	500 bp	1000 bp
1 locus	0.0982	0.1006	0.0191	0.0237
2 loci	0.1052	0.1116	0.0184	0.0167

Table 3: Theoretical results for tree of 2 tips

The expected distribution of t for trees of 2 tips was determined, and centiles of the distribution used to construct a distribution for $\hat{g}$. Given values are mean, standard deviation, and median of $\hat{g}$ for one, two and three loci. The true value of g was 0.0. Θ was assumed to be known without error. The result for 100 loci is an approximation based on 1000 replications using values of t drawn at random from the distribution for each locus.

loci	mean $\hat{g}$	SD of mean	median $\hat{g}$
1	20.3	75.4	2.2
2	3.1	12.3	0.8
3	1.3	3.6	0.5
100	0.02	0.07	0.01

Table 4: Results from perfectly coalescent genealogies

Estimates of Θ and g, and upper and lower approximate 95% confidence limits, for "perfect" genealogies of the given number of sequences. True $\Theta = 1.0$, true $g = 0.0$.

# tips	Θ	lower	upper	g	lower	upper
10	1.2093	0.7454	2.5514	1.012	-2.283	3.458
100	1.0200	0.9463	1.1127	0.497	-1.751	2.079
1000	1.0026	0.9913	1.0143	0.422	-1.634	1.892
10000	1.0003	0.9988	1.0018	0.409	-1.610	1.860