Inference about Population Growth from Single Nucleotide Polymophisms Frequencies

A. Polanski[1,2], M. Kimmel[1]
[1]Department of Statistics, Rice University, Houston, TX, USA
[2]Institute of Automation, Silesian Technical University, Gliwice, Poland

March 29, 2002

Abstract

We have evaluated probability distributions of estimates of parameters of population growth, based on data on frequencies of alleles of unlinked SNP sites in DNA, modeled with the use of time dependent coalescence process acting together with mutation of very low intensity. Probability distributions of maximum likelihood estimates of product parameter of present population effective size and exponent coefficient, for exponential scenario, have atoms at zero and long tails to the right. For stepwise scenario, log likelihood functions typically have very long ridges (covering many decades of the scale) of almost the same value of log likelihood. Observational data from (Picoult Newberg et al.1999) are not inconsistent with the hypothesis of population growth.

1 Introduction

Single Nucleotide Polymorphisms (SNP) seem to be most promising genetic markers due to their high density in human genome. Publicly available SNP databases constantly increase in number and size. A lot of research was done to develop methods for SNP discovery and to characterize distributions of SNPs across the genome (Wang et al. 1998), (Collins et al 1997), (Marth et al. 1999), (Picoult Newberg et al.1999), (Cargill et al. 1999), (Altshuler et al. 2000). SNP data has already been used in association studies of complex

diseases (Boerwinkle et al. 1996), (Bonnen et al. 2000), (Halushka et al. 1999); it is believed that eventually they will enable creating fine genetic maps for complex traits analysis (Kruglyak 1999), (Rish 2000).

Researches were also conducted to use SNP data in population genetics models, for inference on demographic parameters and history. Kuhner et al. (2000) analyzed estimation of the product parameter $\theta = 4N_e\mu$ of effective population size N_e and mutation rate μ, under assumption of constant population size and various hypotheses of spatial (chromosomal) distributions of SNPs: fully or partially linked, or linked segments of nonrecombining SNP sites. Basing on extensive simulations, accuracy of estimates and possible sources of bias were evaluated.

Studies by Nielsen (2000) and Wakeley et al. (2001) were devoted to detection of signatures of human population growth in SNP data. For estimation of growth parameters both researches used maximum likelihood method for unlinked SNP sites created by mutation process of very low intensity. The models included assumptions concerning SNP ascertainment procedures, since it is well known that strategy for SNP discovery significantly influences their sampling distributions (Renwick et. al, 2002), (Yang et al., 2000), (Eberle and Kruglyak, 2000). Nielsen (2000) fitted the scenario of exponential expansion $N_e(t) = N_{e0}\mathrm{e}^{-rt}$ ($N_e(t)$ - effective population size, t - time in generations measured backwards, r - growth exponent) to SNP data from the paper (Picoult Newberg et al.1999). Surprisingly, estimates of the product parameter rN_{e0} were equal to zero for both cases of unmodeled and modeled ascertainment procedure. Wakeley et al. (2001) used the model of stepwise change of population size $N_e(t) = N_e$ for $t < t_s$, and $N_e(t) = N_{ea}$ for $t > t_s$ (N_e, N_{ea} are present and ancestral effective population sizes; stepwise change of population size occurs at $t = t_s$ generations before now) with additional, hidden, population subdivision (Wakeley 2001). Fitting their model to SNP data from (Wang et al. 1998), (Cargill et al. 1999) and (Altshuler et al. 2000) they aimed at estimating ratios N_{ea}/N_e and $t_s/2N_e$. Due to divergent shapes of likelihood surfaces and limited accuracy of computed likelihoods, they were not able to find unique maximum points, but parameter - space regions corresponding to highest likelihoods were not inconsistent with the hypothesis of population growth. Moreover, unmodeled ascertainment led to less likely shapes of parameter regions, and comparison of cases of modeled and unmodeled population structure seemed to support the latter scenario.

In this paper we accept basic hypotheses used in the above mentioned studies: unlinked SNP sites and low intensity mutation. We address the

problem: "How reliable and accurate are estimates of population growth parameters, based on SNP data?". This problem was not fully explored in previous studies due to computational difficulties. Here we use the method for analytical calculation of distributions of coalescence times in time - varying population size evolution (Polanski and Kimmel, 2002) which greatly improves efficiency of numerical computations, and allows us to perform number of computational experiments enough to determine distributions and confidence regions of parameter estimates. Since it was demonstrated in several studies, e.g., (Kuhner et al. 1998), (Pybus et al. 2000), (Polanski et al. 1998), that population genetics processes can lead to unstable or biased parametric and non parametric estimates of population size history, then exploring variabilities and biases of estimates of population growth parameters, based on SNP frequencies, seems a reasonable step towards understanding observational SNP data.

2 Methods

We analyze the situation where the data under study comes from a number of unlinked SNP sites, obtained at random positions in the genome. Denote number of SNP loci by K, and let the observed diallelic data be given by

$$X = \{X_1, X_2, \ldots, X_K\} = \left\{(x_1^R, x_1^F), (x_2^R, x_2^F), \ldots, (x_K^R, x_K^F)\right\} \qquad (1)$$

where x_k^R is the number of copies of less frequent (rare) allele, x_k^F - number of copies of more frequent one, in the sample of $n_k = x_k^R + x_k^F$ taken at SNP site no k. It is possible that $x_k^R = x_k^F$ for some indices k. We assume that it is not known which one of the two alleles is mutant and which is ancestral (wild). Due to independence between sites, the likelihood function for the whole sample $L(P|X)$, given the vector of parameters P related to populations demographic history, is the product of likelihoods $L(P|X_k)$ for sites $1, 2, ...K$:

$$L(P|X) = \prod_{k=1}^{K} L(P|X_k) \qquad (2)$$

We accept standard coalescent assumptions. Random variables given by coalescence times for the sample of size n are denoted by $T_n, T_{n-1}, \ldots, T_2$, and their realizations by corresponding small letters $t_n, t_{n-1}, \ldots, t_2$. Times between coalescence events are denoted by $S_n, S_{n-1}, \ldots, S_2$, and $s_n, s_{n-1}, \ldots, s_2$;

this notation is shown in figure 1, for $n = 5$. When referencing times T_k, or S_k we do not add index n that would define the sample size. The underlying value of n is always clear from the context. Mutation is modeled by a Poisson process with intensity μ per generation, per site. Following discussions in referenced papers (Wang et al. 1998) and (Nielsen 2000) we use statistics which follow from passing to the limit $\mu \to 0$.

2.1 Case with no ascertainment condition

In the situation where DNA sample is scanned for SNPs unconditionally, probability $L(P|X_k)$ on the right hand side of (2) is given by (Nielsen 2000, eqns (5)-(6), Griffiths and Tavare, 1998, eq. (1.3) and X-Y Fu, 1995, eq. (14))

$$L(P|X_k) = \frac{E[S(X_k)|P]}{E(T_\Sigma|P)} \tag{3}$$

where T_Σ is the sum of branch lengths in coalescent tree

$$T_\Sigma = \sum_{j=2}^{n_k} jS_j = T_2 + \sum_{j=2}^{n_k} T_j \tag{4}$$

and $S(X_k)$ is the sum of lengths of all edges in which a single mutation could cause the site pattern X_k. The expectation in numerator in (3) can be computed as

$$E[S(X_k)|P] = \sum_{j=2}^{n_k} jE(S_j|P)\frac{\binom{x_k^R}{2} + (1-\delta_{x_k^R,x_k^F})\binom{x_k^F}{j-2}}{\binom{n_k-1}{j-1}}. \tag{5}$$

In the above, $\delta_{x_k^R x_k^F}$ is a Kronecker delta function.

2.2 Modeling ascertainment

Wakely et al. (2001) have an exhaustive discussion of possible ascertainment schemes and the corresponding models. Here we analyze the case, easily treated analytically, where DNA reading for larger sample is preceded by SNP discovery procedure based on the smaller number of chromosomes n_A. Assume that $n_A = 2$ and that the ascertainment sample $n_A = 2$ is included in the data sample of size n_k at k-th SNP locus. Then the likelihood $L(P|X_k)$

becomes (Renwick et. al, 2002), (Yang et al., 2000), (Eberle and Kruglyak, 2000), (Nielsen 2000)

$$L(P|X_k) = \frac{E[S(X_k)|P]x_k^R x_k^F}{n_k(n_k-1)E(T_2)} \tag{6}$$

2.3 Expectations of coalescence times

Likelihoods of samples depend on expected values of times $E(S_j|P)$ and $E(T_\Sigma|P)$ in the coalescence process. In previous studies, for the case of evolution with varying population size, these expectations were computed by Monte Carlo simulations. Here we compute expectations by using analytical expressions for marginal distributions of coalescence times (Polanski and Kimmel 2002). This saves a lot of computational effort and improves accuracy. Let us assume that the effective population size history is described by a function

$$N_e(t),\ t \in < 0, \infty) \tag{7}$$

where time t is measured in number of generations from now to the past. For a random sample of n DNA sequences, joint probability density function of the distribution of their coalescence times $T_n, T_{n-1}, ..., T_2$ is given by the expression (Griffiths and Tavare, 1994)

$$p(t_n, t_{n-1}, ..., t_2) = \prod_{j=2}^{n} \frac{\binom{j}{2}}{N_e(t_j)} \exp\left(- \int_{t_{j+1}}^{t_j} \frac{\binom{j}{2} d\sigma}{N_e(\sigma)} \right) \tag{8}$$

Marginal distributions $\pi_n(t_n), \pi_{n-1}(t_{n-1}), ..., \pi_2(t_2)$ of times $T_n, T_{n-1}, ..., T_2$ were computed by Polanski and Kimmel, 2002, as follows

$$\pi_j(t_j) = \sum_{l=j}^{n} A_l^j q_{\binom{l}{2}}(t_j) \tag{9}$$

where $q_{\binom{l}{2}}(t) = \frac{\binom{l}{2}}{N_e(t)} \exp\left(-\int_0^t \frac{\binom{l}{2} d\sigma}{N_e(\sigma)} \right)$ and $A_l^j = \prod_{\substack{s=j \\ s \neq l}}^{n} \binom{s}{2} / \prod_{\substack{s=j \\ s \neq l}}^{n} \left[\binom{s}{2} - \binom{l}{2} \right], A_n^n = 1.$

Denote by $e_{\binom{l}{2}}$ expected value $e_{\binom{l}{2}} = \int_0^\infty t q_{\binom{l}{2}}(t) dt$. Then from (9) expected values of times $E(S_j|P)$ and $E(T_\Sigma|P)$, with $P = N_e(t)$, can be expressed as

$$E(S_j|N_e(t)) = A_j^j e_{\binom{j}{2}} + \sum_{l=j+1}^{n} B_l^j e_{\binom{l}{2}} \tag{10}$$

and

$$E(T_\Sigma|N_e(t)) = \sum_{j=2}^{n} j \left(A_j^j e_{\binom{j}{2}} + \sum_{l=j+1}^{n} B_l^j e_{\binom{l}{2}} \right) \tag{11}$$

where: $B_l^j = \prod_{s=j}^{n} \binom{s}{2} / \prod_{\substack{s=j \\ s\neq l}}^{n} \left[\binom{s}{2} - \binom{l}{2} \right]$.

P can be specified as a finite dimensional vector by assuming parametric form for $N_e(t)$. For exponential model

$$N_e(t) = N_{e0} e^{-rt}, \tag{12}$$

$N_e(t)$ - effective population size, r - growth exponent, it becomes (Slatkin and Hudson, 1991)

$$e_{\binom{l}{2}} = -\frac{\exp(\frac{\binom{l}{2}}{rN_{e0}})}{r} Ei(-\frac{\binom{l}{2}}{rN_{e0}}) \tag{13}$$

where Ei denotes exponential integral (Gradshteyn, Ryzhik, 1980, §4.331.2).

For stepwise model

$$N_e(t) = \begin{cases} N_e for t < t_s \\ N_{ea} for t > t_s \end{cases}, \tag{14}$$

N_e, N_{ea} - present and ancestral effective population sizes, stepwise change of population size occurs at $t = t_s$ generations before now, it becomes

$$e_{\binom{l}{2}} = \frac{N_e}{\binom{l}{2}} \left[1 - \left(1 - \frac{N_{ea}}{N_e} \right) \exp(-\frac{\binom{l}{2}\, t_s}{N_e}) \right]. \tag{15}$$

From (13) and (15) it is clear that for exponential scenario $P = rN_{e0}$, and for stepwise scenario $P = (P_1, P_2) = (\frac{t_s}{N_e}, \frac{N_{ea}}{N_e})$. From now on we introduce notation $\kappa = rN_{e0}$, and $\tau = \frac{t_s}{N_e}$, $\delta = \frac{N_{ea}}{N_e}$.

3 Results

We have performed several series of numerical simulations, where SNP data were generated according to distributions (3)-(6). We have changed between the following parameters:

(1) Scenario of population growth (between exponential and stepwise),
(2) Values of entries of true parameters κ, and τ, δ.
(3) Ascertainment procedure (present or absent).

Sample size was assumed $n = 20$, and number of SNP loci was taken $K = 50$, in all simulations. Population size history parameters, κ, and τ, δ, were estimated by maximizing the likelihood function (2). We researched the effect on the estimates of parameters, of unmodeled ascertainment, i.e., on the estimates obtained when data was generated using expression (6), while likelihood function was computed using (5).

We have also reexamined observational data on SNPs from the paper (Picoult Newberg et al.1999), using experiences which follow from our computational experiments.

The obtained results are summarized below.

3.1 Exponential model

Assume that the true value of the parameter of exponential model (12) is $\kappa = 1$, and that the DNA sample is scanned for SNPs unconditionally. Sample size is $n = 20$ and number of SNP loci $K = 50$. When we (A) simulate frequencies of SNP alleles, by generating 50 independent realizations of the distribution given by (3) and (5), and (B) try to restore value of the parameter κ by maximizing likelihood (2), (3), (5), then typical log likelihood curve looks like that shown in fig. 2, upper plot. As steps (A) and (B) are repeated many times, the second possible shape of log likelihood curve, shown in fig. 2, lower plot, is also observed (about 10% of simulations). This curve has no maximum corresponding to $\kappa > 0$; $\kappa_{est} = 0$ is the most likely estimate. Repeating (A) and (B) 1000 times we got approximate distribution of estimate κ_{est}. The estimated cumulative probability function for this distribution is presented in fig. 3. As seen from fig. 3, the distribution of κ_{est} has an atom (of weight 0.104) at $\kappa_{est} = 0$, corresponding to log likelihood curves from lower plot of figure 2. It is also rather heavy tailed, with values reaching far above the true $P = 1$. Similar observations were made for the case of modeled ascertainment procedure, when probability distribution for X_k is given by (6).

Using the method as above, we have estimated probability distributions corresponding to true values of κ : 0.1, 1, 10, 100 and 1000 for both cases of unconditional scan for SNP in DNA data, and ascertainment procedure based on two chromosomes. With the notation,

- median(κ_{est})

- $P_{=0}$ - probability that $\kappa_{est} = 0$

- $P_{0.9-1.1}$ - probability that $0.9\kappa < \kappa_{est} < 1.1\kappa$, where κ is the true value of the parameter

- $P_{0.5-2}$ - probability that $0.5\kappa < \kappa_{est} < 2\kappa$,

- $P_{0.1-10}$ - probability that $0.1\kappa < \kappa_{est} < 10\kappa$,

results of performed simulations are presented in table 1 (a) (the case of unconditional scan for SNP in DNA data), and (b) (the case where ascertainment procedure is based on two chromosomes, as given by expression (6)).

3.2 Stepwise change model

We assumed that ascertainment method is based on two chromosomes as given by expression (6) and $n = 20$, $K = 50$. Again we used the procedure of (A) generating $K = 50$ independent realizations of the distribution (3), (5), (15), and (B) restoring values of τ, δ by maximizing likelihood (2). Typical plots of log likelihood level curves, on the plane τ - δ, for true values of parameters $\tau = 0.01$, $\delta = 0.01$ (left plot) and $\tau = 0.01$, $\delta = 0.01$ (right plot) are presented in fig. 4. Regions bounded by level curves $\max -0.5$, are shaded grey. Graphs of log likelihood function, on the plane τ - δ, show very long ridges of almost the same value of likelihood. The ranges of likely values of parameters cover many decades of log scales.

We have repeated steps (A)-(B) 500 times for combinations of parameters from fig. 4. Fig. 5 shows two dimensional histograms of the estimates τ_{est}, δ_{est} obtained from maximizations of log likelihoods. Data in fig. 5 are not suitable to estimate confidence regions or moments of estimates. They show, however, possible ranges of parameters.

3.3 Unmodeled ascertainment

By unmodeled ascertainment we mean the situation where generation of data in the above step (A) is done basing on the expression (6), while retrieving

parameters in step (B) uses expression (5). Table 2 shows statistics of estimates κ_{est} obtained in 1000 repeats of steps (A)-(B) under unmodeled ascertainment. In the exponential scenario of growth, unmodeled ascertainment results in large bias in estimate of κ - most often estimate falls to $\kappa_{est} = 0$.

In order to study results of unmodeled ascertainment in the stepwise scenario of population history, we have taken the same data (SNP frequencies) which was previously used to draw contour lines in fig. 4 left plot (true values of parameters $\tau = 0.01$, $\delta = 0.01$). Fig. 6 shows comparison of log likelihood contour lines for modeled and unmodeled ascertainment. Left plot in fig. 6, is the same as left plot in fig. 4, while right plot in fig. 6 shows log likelihood level curves computed with unmodeled ascertainment, i.e., computed not by formula (6), but with the use of expression (5). Comparing left and right plots in fig. 6. one can see that, for our data, unmodeled ascertainment shifts values of parameters corresponding to highest likelihoods towards the range of larger δ.

3.4 Observational data from (Picoult Newberg et al.1999)

Observational data from (Picoult Newberg et al.1999) were previously used by Nielsen (2000) in conjunction with the model of exponential population expansion. Nielsen (2000) confined his analysis to 37 polymorphic SNP sites from 44 shown in table 4 (Caucasians) in (Picoult Newberg et al.1999). He omitted 7 monomorphic sites from this table. Here we take both two - element ascertainment sample and data sample of 44 SNP sites table 4 (Caucasians) in (Picoult Newberg et al.1999). In the notation from Wakeley et al. (2001), we use $n_A = 0$, $n_O = 2$, and $n_D = 16$. This approach is consistent with expression (6). In our notation $n = 18$, and $K = 44$. Log likelihood curve for parameter κ for exponential model for these data is presented in figure 7. This curve attains maximum which leads to estimate $\kappa_{est} = 0.0732$. Level curves of log likelihood resulting of fitting stepwise change model to these data are given in fig. 8. Maximization procedure launched for these data gives following values: $\tau_{est} = 10^{-9.13}$, $\delta_{est} = 10^{-10.44}$. These values cannot be accepted as estimates of true parameters, but are quite consistent with ranges seen in fig. 5.

4 Discussion

We have evaluated probability distributions of estimates of parameters of population growth, based on data on frequencies of alleles of unlinked SNP sites in DNA. We were able to perform more, and more accurate, computational experiments than it was done in previous studies. Our study explores variability of estimates of exponential or stepwise scenarios of population growth, obtained when one uses time dependent coalescence process acting together with mutation of very low intensity to model SNP frequencies (Nielsen, 2000, Wakeley et al., 2001). Sample size, $n = 20$, and number of SNP sites $K = 50$, which we used in our simulations, are comparable to those reported in observational studies.

Probability distributions of maximum likelihood estimates of parameter κ for exponential scenario have atoms at $\kappa = 0$ and long tails to the right. Comparing probabilities $P_{=0}$, $P_{0.9-1.1}$, $P_{0.5-2}$ and $P_{0.1-10}$, in table 1 (a)(b), corresponding to different values of true κ, shows that quality of estimation of κ is highest for true $\kappa = 10$ and deteriorates for both $\kappa < 10$ and $\kappa > 10$. This irregular behavior is consistent with the result published by Pybus et al., (2000). Figure 2 in (Pybus et al., 2000) presents lower bounds of biases and variabilities of estimates of parameters r and N_{e0} of exponential model. Method for estimating lower bounds uses the assumption that coalescence times $t_n, t_{n-1}, ..., t_2$ are known exactly (Felsenstein, 1992). Lower bounds of biases and variabilities of estimates of r and N_{e0} depend only on the product of the true parameters $\kappa = rN_{e0}$. Moreover, changing κ has always opposite effects on estimates r_{est} and N_{e0est}, if bias and variability of r_{est} increase then corresponding parameters for N_{e0est} decrease, and conversely. We have repeated computations from Pybus et al. (2000) with the following modification: we have studied variability of estimate κ_{est} of product parameter rather then estimates r_{est} and N_{e0est} separately. By variability of estimate we mean $std(\kappa_{est})/\kappa$ (κ - true value of the parameter). The result, based on 10000 repeats of log likelihood maximization, on the grid r,$N_{e0} \in [10^{-2} \div 10^{2}]$ is shown in fig. 9. Variability curve shown in fig. 9 takes its minimum at $\kappa = 10$, consistently to our findings.

For stepwise scenario, log likelihood functions of τ and δ typically have very long ridges (over many decades of the scale) of almost the same value of log likelihood. Probability density functions of estimates τ_{est} and δ_{est} cover very wide range values ($10^{-16} - 10^{10}$). Therefore parameters of stepwise change cannot be obtained by maximization of log likelihood. Instead (like

in the study by Wakeley et al. 2001) one can only look at shapes of regions which correspond to high probabilities.

Unmodeled ascertainment has a very strong effect on estimation of parameter κ of exponential growth. As seen from table 2, estimates κ_{est}, in this case, are almost always equal to zero. In comparison, the same effect for the case of stepwise scenario, presented in fig. 5, seems much weaker. However: (1) The comparison in fig. 5 is only qualitative, since it does not concern numbers but only shapes of level curves; (2) There is a strong dependence of the effect of unmodeled ascertainment on true values of parameters τ and δ. For values other than those in fig. 5, qualitative effect can be much bigger (results not shown here).

Observational data from (Picoult Newberg et al.1999), (Caucasians), are not inconsistent with the hypothesis of population growth. However, there are some problems which need verification. We have compared our estimates of population history to those done previously by Rogers and Harpending (1994), Polanski et al. (1998), Weis and Haeseler (1998), Slatkin and Hudson (1991). Using predictions done by the above authors, reasonable ranges of values of growth parameters seem: for exponential scenario $\kappa = 100 \div 1000$, for stepwise scenario $\tau = 0.005 \div 0.05, \delta = 0.001 \div 0.01$. Under exponential scenario, our estimated value $\kappa_{est} = 0.0732$ is in large discrepancy with the above. It seems that there are three possible explanations: (1) The model for ascertainment procedure is still not enough adequate. Since ascertainment has very strong effect on the estimate (table 2) it must be modeled rather precisely to get reliable estimates of parameters. When comparing simple model (6) to the description of 4 filtering steps in (Picoult-Newberg, 1999) one can argue that a better model of ascertainment may be necessary. (2) For values of κ in the range $100 \div 1000$, variability of estimate does not leave κ_{est} close to zero or equal to zero very improbable (table 1). (3) Exponential model of population size history is not adequate. In our opinion joint effect of (1) and (2) is quite probable.

In contrast to the above, stepwise scenario with e.g., $\tau = 0.01$, $\delta = 0.05$ seems to fit quite well to observational data. This can be seen by comparing fig. 8 where level curves of log likelihood for data from (Picoult-Newberg, 1999) are shown with fig. 4, for similar parameters. It was not necessary to add mechanism like population substructure (Wakeley et al. 2001). The model is flexible enough without that assumption. However, a fit based on comparisons of shapes of likelihood level curves, rather then on values of estimated parameters, is much less reliable. Probably, its reliability could be

improved by increasing number of SNP loci.

References

[1] Altshuler D., V. J. Pollar, C. R. Cowles, W. J. Van Etten, J. Baldwin, L. Linton, E. S. Lander, 2000, A SNP map of the human genome generated by reduced representation shotgun sequencing, Nature 407:582-589.

[2] Boerwinkle E., D. L. Ellsworth, D. M. Hallman, A. Biddinger, 1996, Genetic analysis of artherosclerosis: a research paradigm for the common chronic diseases, Hum. Mol. Genet., 5: 1405-1410.

[3] Bonnen P. E., M. D. Story, C. L., Ashorn, T. A. Buchholz, M. M., Weil, D. L. Nelson, 2000, Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium, Am. J. Hum. Genet., 67: 1437-1451.

[4] Cargill M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, E. S. Lander, 1999, Characterization of single-nucleotide polymorphisms in coding regions of human genes, Nat. Genet., 22: 231-238.

[5] Collins F. S., M. S. Guyer, and A. Chakravarti, 1997, Variations on a theme: cataloging human DNA sequence variation, Science, 278: 1580-1581.

[6] M. A. Eberle, L. Kruglyak, 2000, An analysis of strategies for discovery of single nucleotide polymorphisms, Genet. Epidemiology, 19(Suppl 1): S29-S35.

[7] J. Felsenstein, 1992, Estimating Effective Population Size from Samples of Sequences, Inefficiency of Pairwise and Segregating Sites as Compared to Phylogenetic Estimates, *Genet. Res.*, vol. 59, pp. 139-147.

[8] Fu X.-Y., 1995, Statistical properties of segregating sites, Theoret. Popul. Biol., 48: 172-197.

[9] Gradshteyn I. S., I. M. Ryzhik, Table of integrals, series and products, fifth ed., Academic Press, 1980.

[10] Griffiths R.C., S. Tavare, 1998, The age of mutation in the general coalescent tree, Stochastic Models, 14: 273-295.

[11] Griffiths R.C., S. Tavare, 1994, Sampling theory for neutral alleles in a varying environment, Proc. R. Soc. Lond. B, 344: 403-410.

[12] Halushka M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, A. Chakravarti, 1999, Patterns of single-nucleotide polymorphisms in candidate genes for blood pressure homeostasis, Nat. Genet. 22: 239-247.

[13] Kruglyak L., Prospects for whole-genome linkage disequilbrium mapping of common disease genes, 1999, Nat. Genet. 22: 139-144.

[14] Picoult-Newberg L., T. E. Ideker, M. G. Pohl, S. L. Taylor, M. A. Donaldson, D. A. Nickerson, M. Boyce-Jacino, 1999, Mining SNPs from EST Databases, Genome Res, 9: 167-174.

[15] M. K. Kuhner, J. Yamato, J. Felsenstein, Maximum Likelihood Estimation of Population Growth Rates Based on Coalescent, Genetics, vol. 149, pp. 429-434, 1998.

[16] M. K. Kuhner, P. Beerli., J. Yamamoto, J. Felsenstein, 2000, Usefulness of Single Nucleotide Polymorphism Data for Estimating Population Parameters, Genetics, 156: 439-447.

[17] Mart G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu., H. Zakeri, N. O. Striziel, L. Hillier, P. Y. Kwok, W. R. Gish, 1999, A general approach to single-nucleotide polymorphism discovery, Nat. Genet., 23: 452-456.

[18] Nielsen R., 2000, Estimation of population paramters and recombination rates from single nucleotide polymorpisms, Genetics, 154: 931-942.

[19] A. Polanski, M. Kimmel, R. Chakraborty, 1998, Application of a Time - Dependent Coalescent Process for Inferring the History of Population Changes from DNA Sequence Data, Proc. Natl. Acad. Sci. USA, vol. 95, pp. 5456-5461.

[20] Polanski A., M. Kimmel, 2002, Marginal distributions in the time dependent coalescence process, submitted.

[21] O.G. Pybus, A. Rambaut, P. H. Harvey, 2000, An Integrared Framework for the Inference of Viral Population History from Reconstructed Genealogies, Genetics, vol. 155, pp. 1429-1437.

[22] Renwick A., P. Bonnen, D. Trikka, D. Nelson, R. Chakraborty, M. Kimmel, 2002, Sampling properties of estimators of nucleotide diversity at discovered SNP sites, submitted.

[23] Rish N. J., 2000, Searching for genetic determination in the new millenium, Nature, 405:847-856.

[24] A. R. Rogers, H. Harpending, Population Growth Makes Waves in the Distribution of Pairwise Genetic Differences, Molec. Biol. Evol., vol. 9., pp. 552-569, 1992.

[25] Slatkin M., Hudson R. R., 1991, Pairwise Comparisons of Mitochondrial DNA in Stable and Exponentialy Growing Populations, *Genetics*, 129: 555-562.

[26] Wakeley J., R.Nielsen, S. N. Liu-Cordero, K. Ardlie, 2001, The discovery of single-nucleotide polymorphisms - and inferences about human demographic history, Am. J. Hum. Genet., 69: 1332-1347.

[27] Wakeley J., 2001, The coalescent in an island model of population subdivision with variation among demes, Theor. Popul. Biol., 59: 133-144.

[28] Wang D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young et al., 1998, Large scale identifiacation, mapping and genotyping of single-nucleotide polymorphisms in the human genome, Science, 280: 1077-1082.

[29] Weiss G., A. Haeseler, 1998, Inference on population history using a likelihood approach, Genetics, 149: 1539-1546.

[30] Yang Z., G. Wong, M. A. Eberle, M. Kibukawa, D. A. Passey, W. R. Hughes, L. Kruglyak, J. Yu, 2000, Sampling SNPs, Nat. Genet., 26: 13-14.

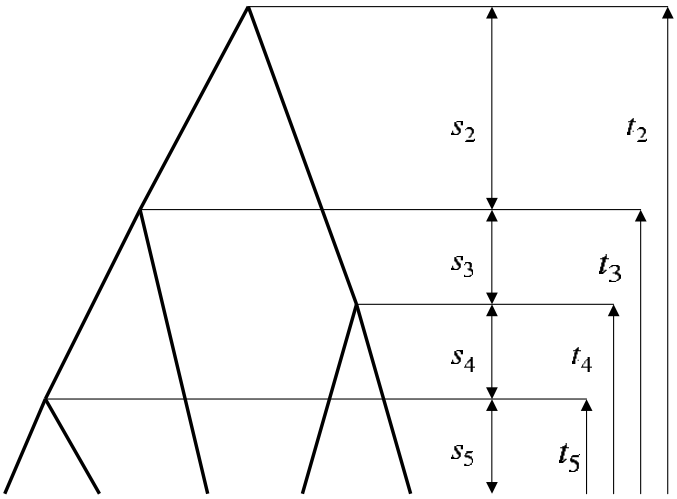

Figure 1. Notation for ancestral history of a sample of DNA sequences. Coalescence times for the sample of size $n = 5$ are denoted by $T_5, T_4, \ldots, T_2$, and their realizations by corresponding small letters $t_5, t_4, \ldots, t_2$. Times between coalescence events are denoted by $S_5, S_4, \ldots, S_2$, and $s_5, s_4, \ldots, s_2$;

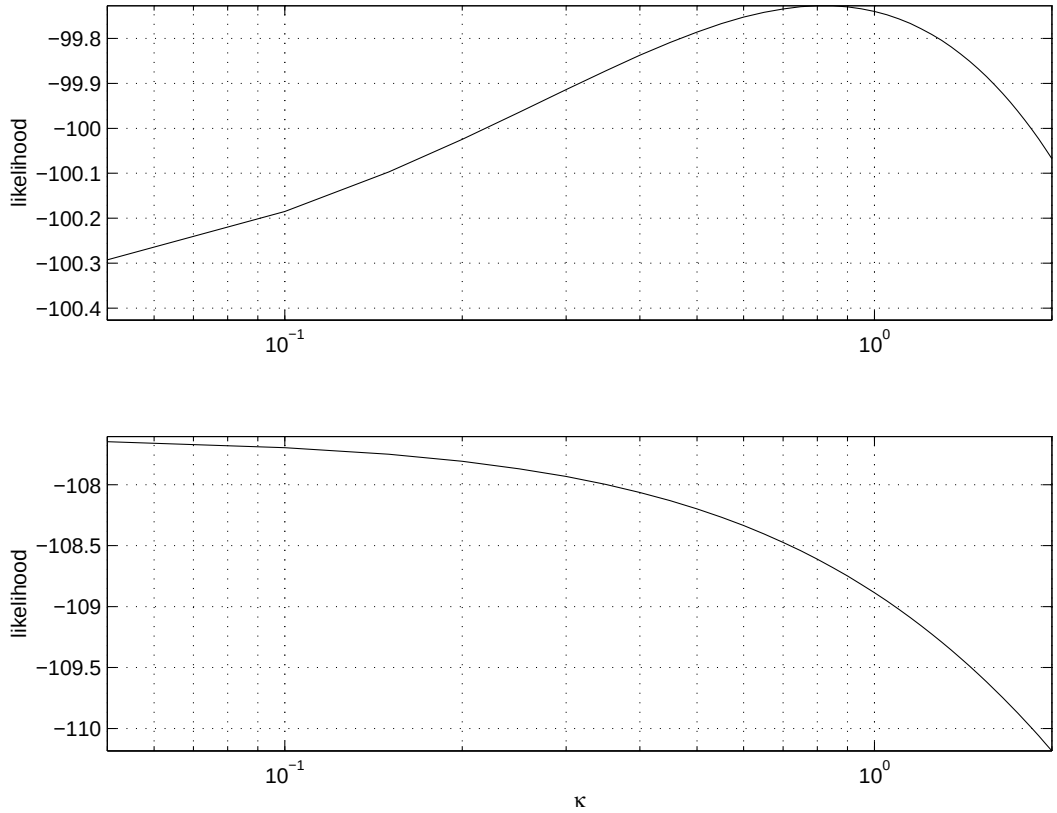

Figure 2. Log likelihood curves for exponential model of popultion growth. Two types of possible results of simulations. Upper plot: Curve which gives estimate $\kappa_{est} > 0$. Lower plot: Curve which attains maximum at $\kappa_{est} = 0$.

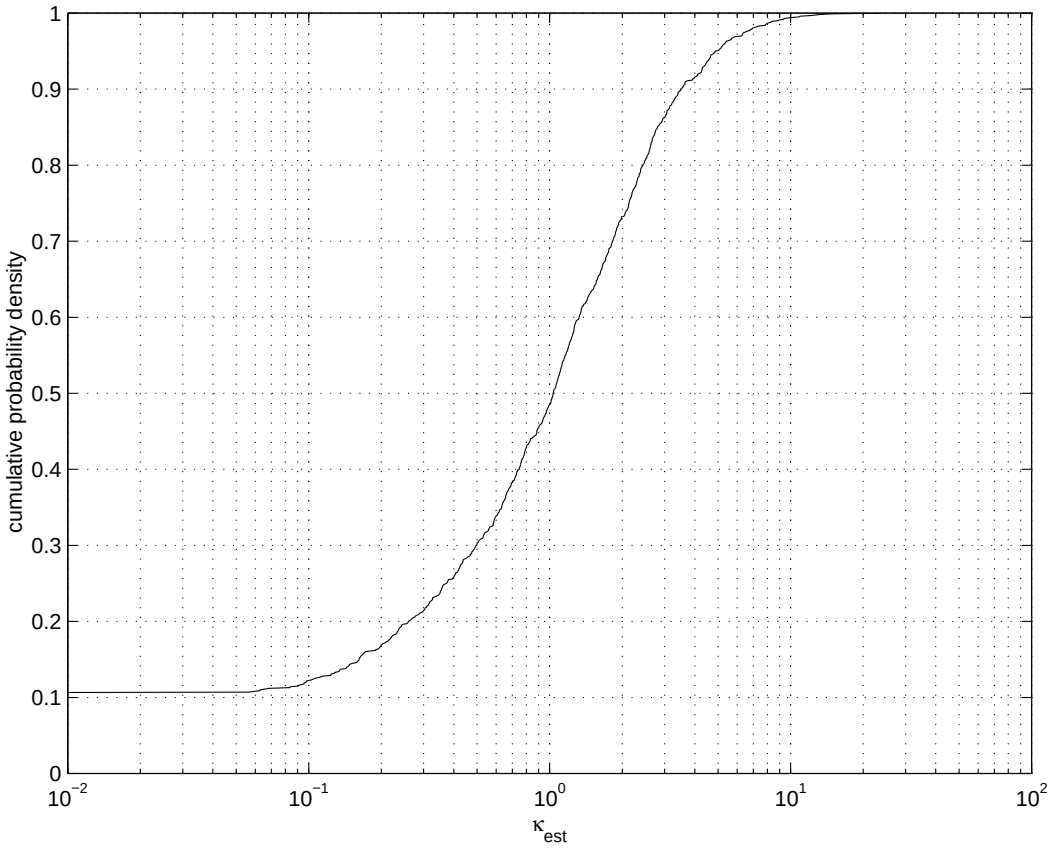

Figure 3. Cumulative probability function for distribution of estimate κ_{est}, obtained from 1000 repeats of maximization procedure. True value of the parameter was $\kappa = 1.0$.

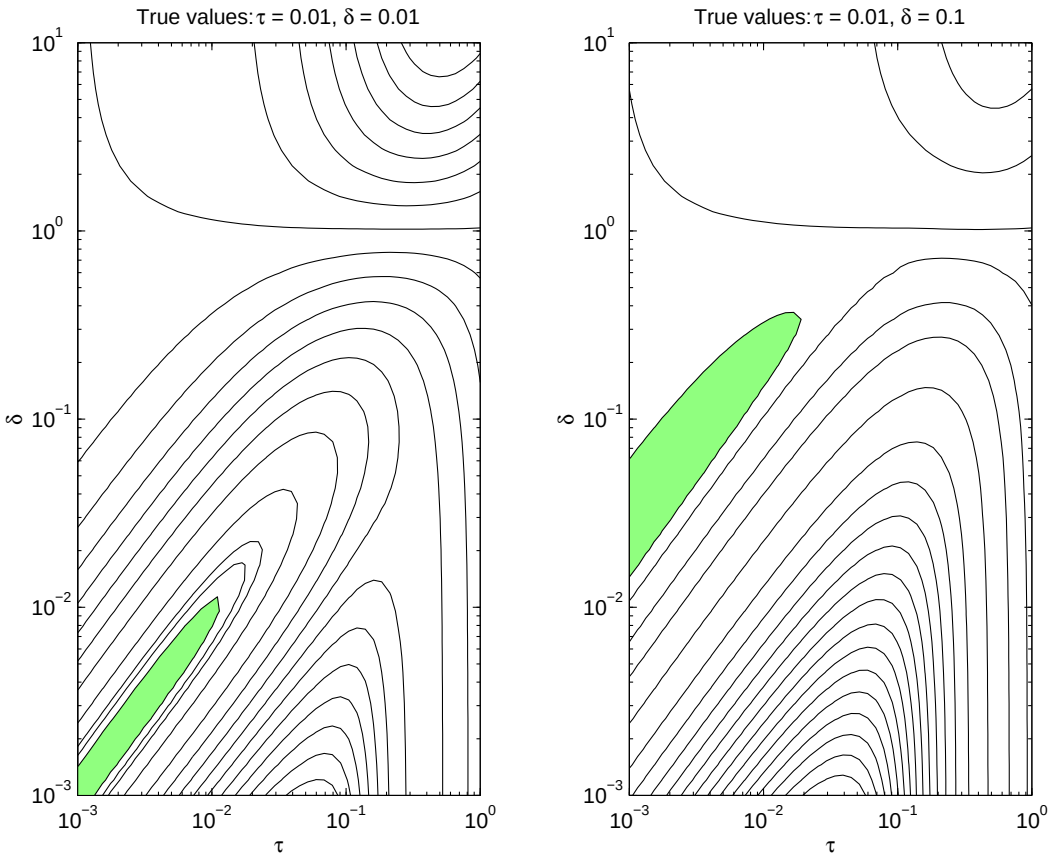

Figure 4. Plots of log likelihood level curves, on the plane $\tau - \delta$ for stepwise model of population history. Regions bounded by level curves $\max - 0.5$, are shaded grey. Maxima are: -85.09, -112.97.

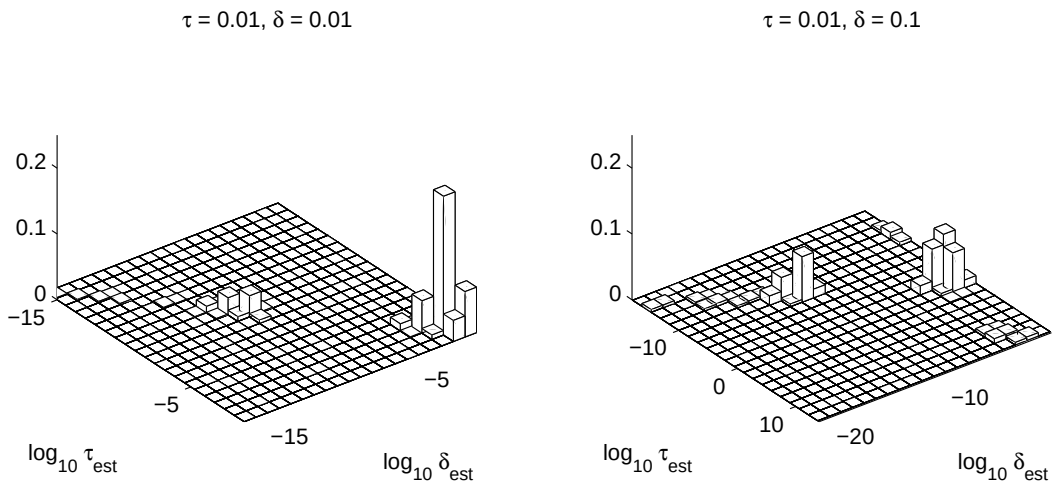

Figure 5. Histograms of estimates of parameters τ_{est} and δ_{est} obtained in 500 repeats of data generation - likelihood maximization procedure. True parameters are: in the left plot $\tau = 0.01$ and $\delta = 0.01$, in the right plot $\tau = 0.01$ and $\delta = 0.1$.

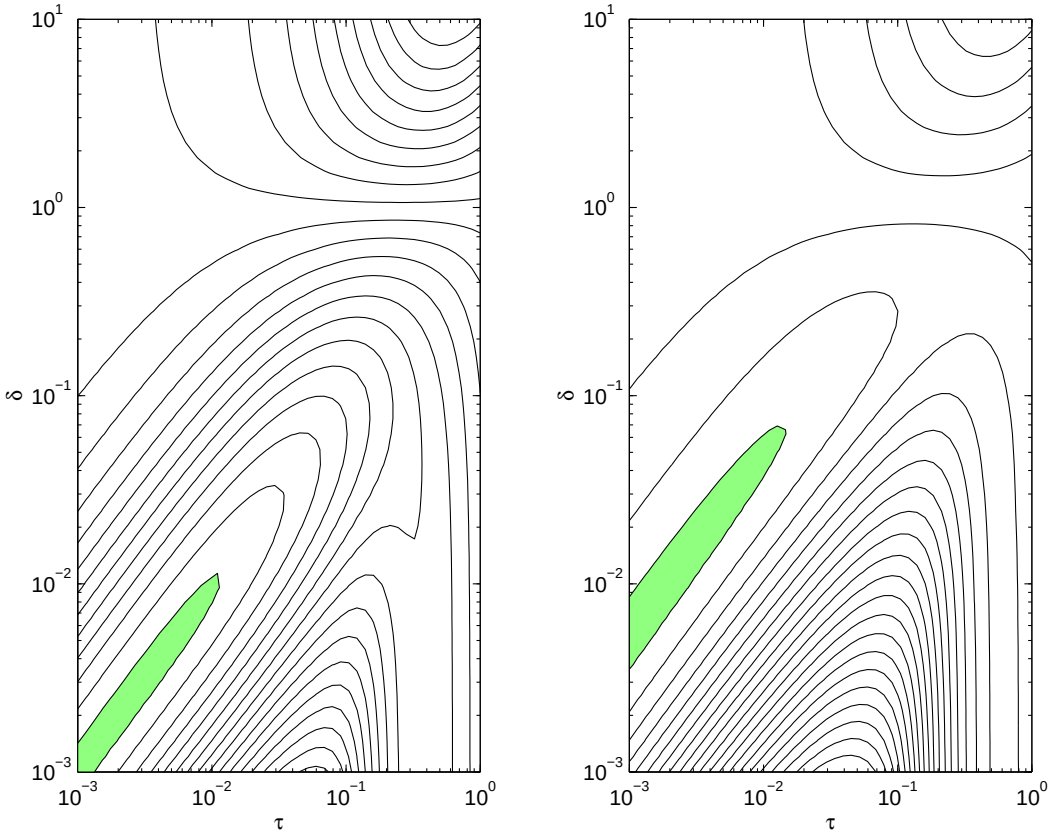

Figure 6. Left plot is the same as left plot from fig. 4. Right plot: contour lines of log likelihood computed not by formula (6) as in the left plot, but with the use of expression (5). Shaded regions are max -0.5. Maxima: -85.09 for left plot, -86.98 for right plot.

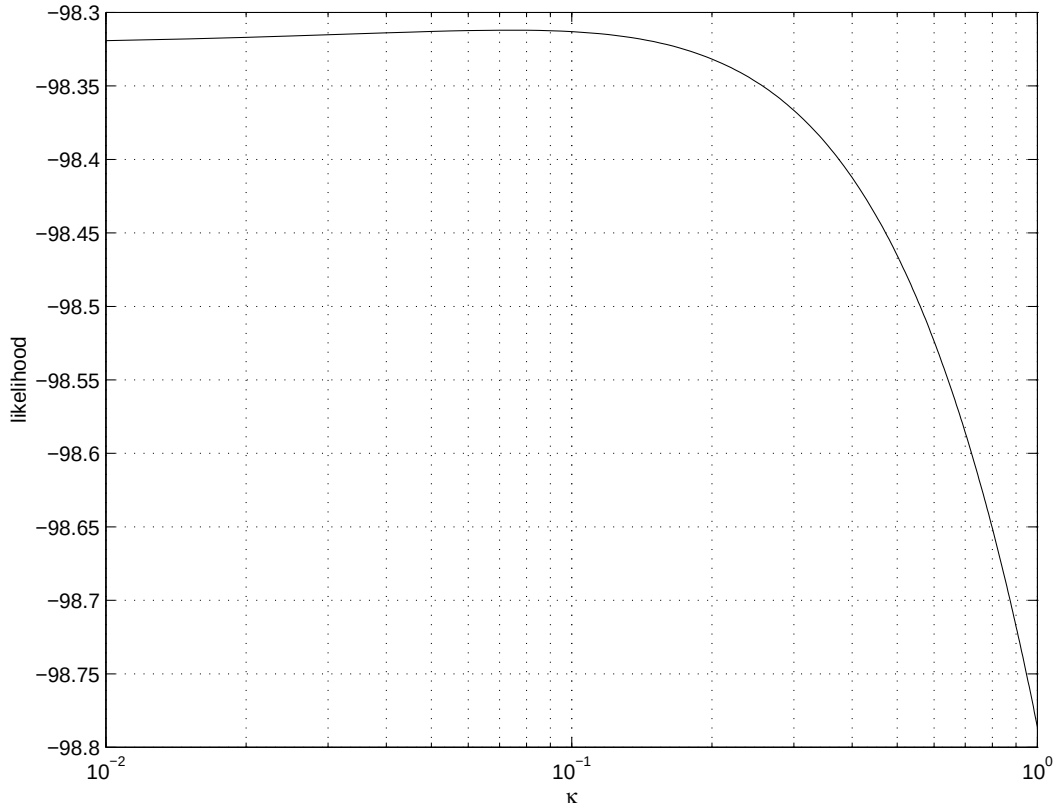

Figure 7. Likelihood curve for parameter κ resuling from fitting exponential model to data from (Picoult Newberg et al.1999). Likelihood curve attains its maximum at $\kappa_{est} = 0.0732$.

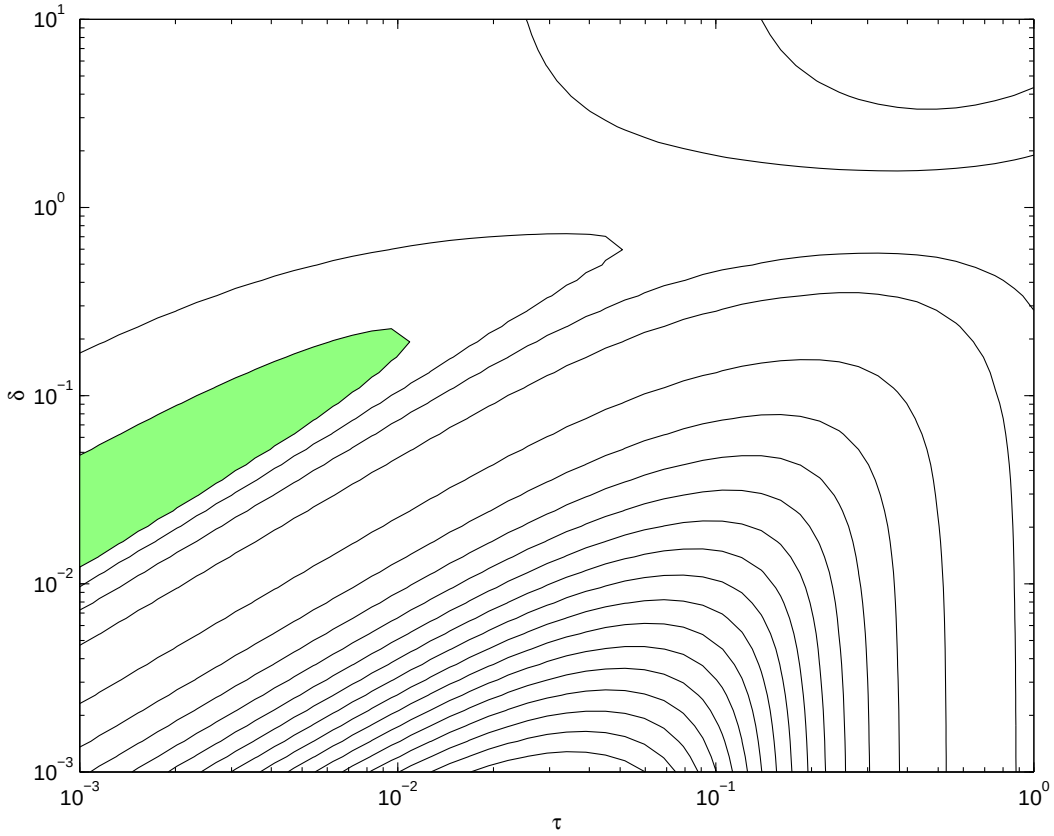

Figure 8. Level curves of likelihood resulting of fitting stepwise change model to data from (Picoult Newberg et al.1999). Region bounded by level curve $\max - 0.1$ is shaded grey. Maximum: $\max = -98.02$.

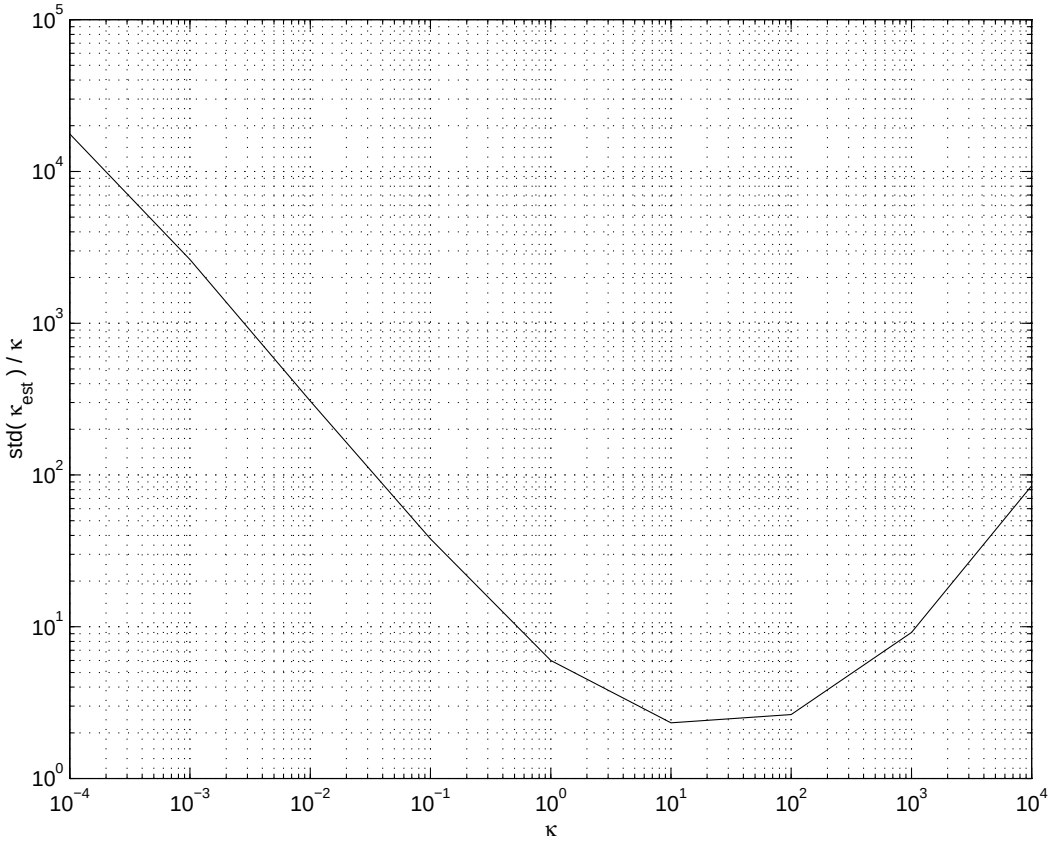

Figure 9. Lower bound of variability $std(\kappa_{est})/\kappa$ of estimate κ_{est} obtained with the use of the method from Pybus et al. 2000.

Table 1. Results of 1000 simulations of the procedure (A) - (B) for estimation of parameter κ. (a) The case of unconditional scan for SNP in DNA data. (b) The case where ascertainment is based on two chromosomes.

(a) Unconditional scan for SNP in DNA data

true κ	median(κ_{est})	$P_{=0}$	$P_{0.9-1.1}$	$P_{0.5-2}$	$P_{0.1-10}$
0.1	0.082	0.424	0.016	0.109	0.455
1	1.032	0.106	0.072	0.429	0.872
10	9.672	0.014	0.107	0.594	0.981
100	102.415	0.144	0.07	0.498	0.811
1000	997.63	0.254	0.071	0.351	0.607

(b) Ascertainment based on two chromosomes

true κ	median(κ_{est})	$P_{=0}$	$P_{0.9-1.1}$	$P_{0.5-2}$	$P_{0.1-10}$
0.1	0.058	0.447	0.01	0.1	0.411
1	0.980	0.120	0.057	0.418	0.84
10	9.468	0.009	0.096	0.615	0.989
100	99.300	0.118	0.101	0.571	0.860
1000	1031.97	0.191	0.064	0.430	0.743

Table 2. Estimation of κ in exponential scenario of population history, under unmodeled ascertainment. $N_{>0}$ denotes number of cases, in 1000 repeats of steps (A)-(B), such that $\kappa_{est} > 0$. Sample size: $n = 20$, number of SNP loci: $K = 50$.

true κ	$N_{>0}$
0.1	0
1	2
10	19
100	1
1000	2