New Explicit Expressions for Relative Frequencies of Single-Nucleotide Polymorphisms With Application to Statistical Inference on Population Growth

A. Polanski*,† and M. Kimmel*,[1]

*Department of Statistics, Rice University, Houston, Texas 77005 and †Institute of Automation, Silesian Technical University, 44-100 Gliwice, Poland

Manuscript received January 29, 2003
Accepted for publication May 30, 2003

ABSTRACT

We present new methodology for calculating sampling distributions of single-nucleotide polymorphism (SNP) frequencies in populations with time-varying size. Our approach is based on deriving analytical expressions for frequencies of SNPs. Analytical expressions allow for computations that are faster and more accurate than Monte Carlo simulations. In contrast to other articles showing analytical formulas for frequencies of SNPs, we derive expressions that contain coefficients that do not explode when the genealogy size increases. We also provide analytical formulas to describe the way in which the ascertainment procedure modifies SNP distributions. Using our methods, we study the power to test the hypothesis of exponential population expansion *vs.* the hypothesis of evolution with constant population size. We also analyze some of the available SNP data and we compare our results of demographic parameters estimation to those obtained in previous studies in population genetics. The analyzed data seem consistent with the hypothesis of past population growth of modern humans. The analysis of the data also shows a very strong sensitivity of estimated demographic parameters to changes of the model of the ascertainment procedure.

A lot of research has been done to develop methods for discovery of single-nucleotide polymorphisms (SNP) and to characterize distributions of SNPs across the genome (Collins *et al.* 1997; Wang *et al.* 1998; Cargill *et al.* 1999; Marth *et al.* 1999; Picoult-Newberg *et al.* 1999; Altshuler *et al.* 2000). SNP data have already been used in association studies of complex diseases (Boerwinkle *et al.* 1996; Halushka *et al.* 1999; Bonnen *et al.* 2000; Trikka *et al.* 2002), and it is believed that eventually they will enable creation of fine genetic maps for complex traits analysis (Kruglyak 1999; Rish 2000). Databases, like that of the SNP Consortium, at http://snp.cshl.org, contain massive amounts of data on positions of SNPs in the human genome, but it is likely that most of these SNPs are very rare and therefore of limited value in gene association studies. Estimates of distributions of expected relative frequencies of SNPs result from studies that use population genetics models, *e.g.*, Durrett and Limic (2001) and Wang *et al.* (1998), and the predicted excess of rare alleles is explained as resulting from expansion of human populations.

Using population genetics methods to model and analyze SNPs opens an area for investigating problems like predicting frequencies of SNPs under various demographic scenarios, inferring demographic parameters and history from sampling frequencies of SNPs, comparing estimates obtained on the basis of SNP data to those obtained with other methods, and evaluating efficiency of using SNP data for estimation of population parameters. Several interesting studies were carried out in this area. Studies by Durrett and Limic (2001) and Wang *et al.* (1998) estimated frequencies of SNPs under the hypothesis of population growth. A problem of how sampling frequencies of SNPs are influenced by ascertainment procedures was investigated by Eberle and Kruglyak (2000), Yang *et al.* (2000), and Renwick *et al.* (2002). Using SNPs for estimation of the scaled product parameter $\theta = 4N_e\mu$ of the effective population size N_e and mutation rate μ, under assumption of constant population size, was studied by Kuhner *et al.* (2000). They took into account various hypotheses of spatial (chromosomal) distributions of SNPs such as complete or partial linkage or occurrence of linked segments of nonrecombining SNPs and, on the basis of extensive simulations, evaluated accuracy of estimates and possible sources of bias. Studies by Nielsen (2000) and Wakeley *et al.* (2001) were devoted to detection of signatures of human population growth in SNP data. Nielsen (2000) fitted the scenario of exponential expansion to SNP data of Picoult-Newberg *et al.* (1999). Wakeley *et al.* (2001) used the model of stepwise change of the population size with population subdivision (Wakeley 2001). They fitted their model to SNP data from Wang *et al.* (1998), Cargill *et al.* (1999) and Altshuler *et al.* (2000). Parameter-space regions corresponding to the highest likelihoods were not inconsistent with the hypothesis of population growth. Moreover, if the ascertainment bias was not considered, less realistic shapes of parameter regions were obtained. Comparison of cases in which population substructure was not consid-

[1]*Corresponding author:* Department of Statistics, Rice University, M.S. 138, 6100 Main St., Houston, TX 77005. E-mail: kimmel@rice.edu

ered with those in which it was considered seems to support the latter scenario. To evaluate SNP frequencies, these studies used the standard coalescence approach and Monte Carlo simulations.

Sampling distributions of SNP frequencies in populations with time-varying size can be calculated with the use of analytical expressions for the expected lengths of branches in the coalescence tree derived in the articles by Griffiths and Tavare (1998), Wooding and Rogers (2002), and Polanski *et al.* (2003). Analytical expressions allow computations, which are faster and more accurate than Monte Carlo simulations. However, the approaches shown in the articles by Griffiths and Tavare (1998), Wooding and Rogers (2002), and Polanski *et al.* (2003) suffer from one common difficulty, numerical instability for larger genealogies. When the analyzed genealogy size is >50, these analytical methods are either inapplicable or difficult to apply, due to the explosion of coefficients in equations. Wooding and Rogers (2002) give a method to stabilize numerical computations, which is valid for the case where effective population size changes in a stepwise manner. Here we show another approach, which is more general in the sense that it does not require assumption of piecewise constant history of effective population size. We transform equations for the relative frequencies of SNP to the form with nondiverging coefficients and we provide expressions, obtained with the use of methods of hypergeometric summation, to compute these coefficients. We also provide analytical expressions to describe the influence of the discovery procedure (ascertainment) on SNP frequencies. Our methods allow us to perform tasks that otherwise are prohibitive or cumbersome, like analyzing large genealogies, estimating confidence limits for parameters by resampling studies, and studying sensitivity of models to parameter changes. Using our methods we study our power to test the null hypothesis of evolution with constant population size *vs.* the alternative hypothesis of population expansion, for SNP data, under the exponential model of population size change. We also analyze some of the available SNP data (Picoult-Newberg *et al.* 1999; Trikka *et al.* 2002) and we compare our results to those obtained in previous studies concerning estimation of populations size changes (Slatkin and Hudson 1991; Rogers and Harpending 1992; Polanski *et al.* 1998; Weiss and Haeseler 1998).

METHODS

We consider the process of coalescence with time-changing effective population size. Notation for the coalescence tree, for the sample of size $n = 5$ DNA sequences, is shown in Figure 1. Time t is measured, in number of generations, from the present to the past. Random times between coalescence events are denoted by $S_n, S_{n-1}, \ldots, S_2$ and $s_n, s_{n-1}, \ldots, s_2$. Cumulative times to coalescence, from sample of size n to sample of size $k-1$, are denoted by T_k, $k = 2, 3, \ldots, n$, and their realizations by corresponding lowercase letters $t_n, t_{n-1}, \ldots, t_2$, $0 < t_n < t_{n-1} \ldots < t_2$.

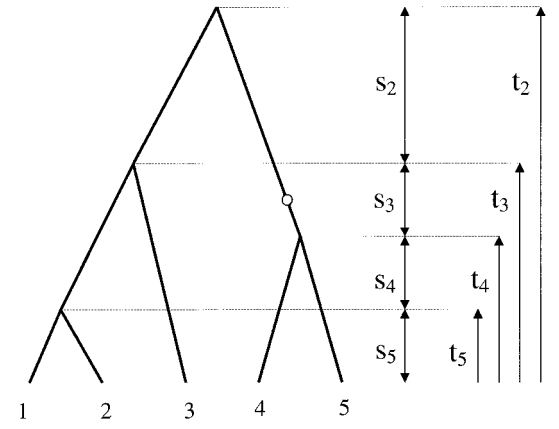

Figure 1.—Notation for ancestral history of a sample of DNA sequences. Coalescence times for the sample of size $n = 5$ are denoted by $T_5, T_4, \ldots, T_2$ and their realizations by corresponding lower case letters $t_5, t_4, \ldots, t_2$. Times between coalescence events are denoted by $S_5, S_4, \ldots, S_2$ and $s_5, s_4, \ldots, s_2$. A mutation event is marked by an open circle. Sequences 4 and 5 have mutant alleles (bases), while sequences 1–3 have ancestral ones.

We assume that an observed SNP was produced by a single, neutral mutation, like the one denoted in Figure 1 by an open circle. In Figure 1 sequences 4 and 5 have mutant alleles (bases), while sequences 1, 2, and 3 have ancestral ones. In the situation where it is not known which allele is mutant and which is ancestral, we use the terms rare and frequent allele. In other words, the SNP in Figure 1 has configuration $b = 2$ mutant *vs.* $n - b = 3$ ancestral, or $b = 2$ rare *vs.* $n - b = 3$ frequent alleles. We assume that mutation intensity for SNPs is very low; *i.e.*, they follow the infinite-sites mutation model.

Probability that a SNP has *b* mutant bases: Probability q_{nb} that a SNP site in a sample of n chromosomes has b mutant bases, under the infinite-sites mutation model, is given by Griffiths and Tavare (1998, Equation 1.3) in terms of expectations of times in the coalescence tree (see also articles by Fu 1995; Sherry *et al.* 1997; Nielsen 2000; Wooding and Rogers 2002). In our notation, this expression has the form

$$q_{nb} = \frac{((n-b-1)!(b-1)!/(n-1)!)\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}E(S_k)}{\sum_{k=2}^{n} kE(S_k)}, \quad (1)$$

where $0 < b < n$, $S_k = T_k - T_{k+1}$, and $T_{n+1} = 0$.

The above expression can be written as

$$q_{nb} = \frac{((n-b-1)!(b-1)!/(n-1)!)\sum_{k=2}^{n}\sum_{j=k}^{n} j(j-1)\binom{n-k}{b-1}A_{kj}^{n}e_j}{\sum_{k=2}^{n}\sum_{j=k}^{n} j(j-1)/(k-1)A_{kj}^{n}e_j} \quad (2)$$

(Polanski *et al.* 2003), where

$$e_j = \int_0^\infty t q_j(t)\,dt \tag{3}$$

are expectations of times distributed as

$$q_j(t) = \frac{\binom{j}{2}}{N_e(t)} \exp\left(-\int_0^t \frac{\binom{j}{2} d\sigma}{N_e(\sigma)}\right), \tag{4}$$

with the effective population size history described by a function of reverse time,

$$N_e(t), \quad t \in [0, \infty). \tag{5}$$

Coefficients A^n_{kj} are defined by the expression

$$A^n_{kj} = \frac{\prod^n_{l=k, l\neq j} \binom{l}{2}}{\prod^n_{l=k, l\neq j} \left[\binom{l}{2} - \binom{j}{2}\right]}, \quad k \le j \le n, \tag{6}$$

$A^{nn}_n = 1$.

Equation 2 is an analytic expression for probabilities q_{nb}. Wooding and Rogers (2002) derive equations with the structure analogous to Equations 2–5, which also contain expectations defined in (3). In contrast to (2), they do not provide explicit expressions for coefficients in the equations; instead they use linear algebra operations (matrix diagonalization) to compute them. Both articles (Wooding and Rogers 2002; Polanski *et al.* 2003) report that it is rather difficult to efficiently apply analytical formulas for genealogies of size $n > 50$ because of the occurrence of diverging terms with alternating signs.

Methods for computation of q_{nb} for large genealogies: To avoid large numerical errors in summations in (2) for genealogies $n > 50$, one needs to apply computations with precision of hundreds, or even thousands, of decimal digits (Wooding and Rogers 2002), which significantly slows down computational process and requires appropriate software. Such computations must be also carefully executed. It is necessary to repeat computations several times, with an increasing accuracy, and to examine the convergence of the returned values.

Wooding and Rogers (2002) developed a way to avoid the necessity of extending precision of the arithmetics, based on a uniformization technique of computing matrix exponents. It is applicable for the case when the population size changes in a stepwise (piecewise constant) manner, with a finite number of steps, and it allows evaluating the expressions in a standard double precision arithmetic. However, when the number of steps in the population size history becomes large, *e.g.*, if one attempts to approximate a given $N_e(t)$ by a piecewise constant function, this approach may be difficult to apply.

Below, we present a method for computing q_{nb} for large genealogies, which is more general than the one developed by Wooding and Rogers (2002), in the sense that it does not require assumption of stepwise change of $N_e(t)$. The idea is to reverse the order of summation in both denominator and numerator in Equation 2. We observe that the resulting expressions contain coefficients that do not explode when n increases. Proceeding in this way we obtain

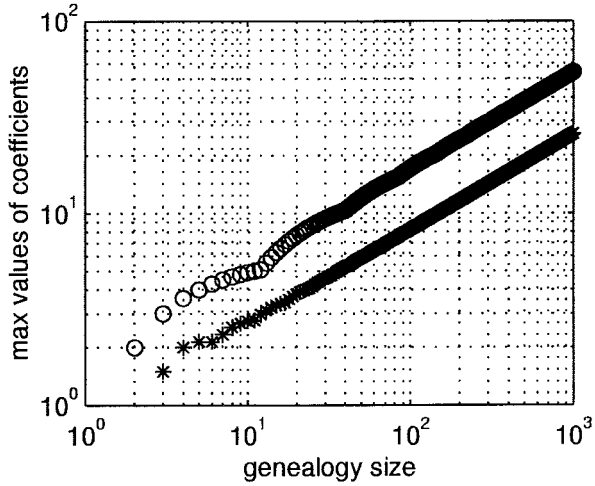

Figure 2.—Growth plots of $\max_{bj}|W^n_{bj}|$ (*) and $\max_j(V^n_j)$ (○) *vs.* n.

$$q_{nb} = \frac{\sum^n_{j=2} e_j \sum^j_{k=2} j(j-1)\binom{n-k}{b-1}((n-b-1)!(b-1)!/(n-1)!)\,A^n_{kj}}{\sum^n_{j=2} e_j \sum^j_{k=2} j(j-1)(A^n_{kj}/(k-1))} \tag{7}$$

$$= \frac{\sum^n_{j=2} e_j W^n_{bj}}{\sum^n_{j=2} e_j V^n_j}. \tag{8}$$

In the above, we introduced coefficients

$$V^n_j = \sum_{k=2}^{j} j(j-1)\frac{A^n_{kj}}{k-1} \tag{9}$$

and

$$W^n_{bj} = \sum_{k=2}^{j} j(j-1)\binom{n-k}{b-1}\frac{(n-b-1)!(b-1)!}{(n-1)!} A^n_{kj}. \tag{10}$$

For $k > n - b + 1$, the elements in the sum (10) become zero, so the upper limit j can be replaced by $\min(j, n - b + 1)$. Coefficients V^n_j and W^n_{bj} remain the same for all histories of effective population size $N_e(t)$. Once calculated, they can be stored in computer memory or tabularized and reused when we wish to analyze different histories $N_e(t)$, *e.g.*, when maximizing likelihood function with respect to population growth parameters. Their important property is that their growth, when genealogy size n increases, is very moderate; *e.g.*, for $n = 100$, $\max_j(V^{100}_j) = 17.13$, $\max_{bj}|W^{100}_{bj}| = 8.24$; for $n = 500$, $\max_j(V^{500}_j) = 38.36$, $\max_{bj}|W^{500}_{bj}| = 18.36$; and for $n = 1000$, $\max_j(V^{1000}_j) = 54.18$, $\max_{bj}|W^{1000}_{bj}| = 25.94$. In Figure 2 growth plots of $\max_{bj}|W^n_{bj}|$ and $\max_j(V^n_j)$ *vs.* n are shown. One can see that both plots are, asymptotically, of the power type with the exponent less than one.

Expressions in (10) and (9) are sums of hypergeometric series, which can be seen by factoring the denomina-

tors in (6), $\binom{l}{2} - \binom{j}{2} = \frac{1}{2}(l-j)(l+j-1)$, and then expressing coefficients A^n_{kj} in (6) as follows:

$$A^n_{kj} = \frac{n!(n-1)!}{(n+j-1)!(n-j)!}\frac{(2j-1)}{j(j-1)}\frac{(j+k-2)!}{(k-1)!(k-2)!(j-k)!}(-1)^{j-k}. \tag{11}$$

Substituting (11) in (9) and using Chu-Vandermode identity (Graham *et al.* 1998, p. 212, Equation 5.93) we obtain

$$V^n_j = (2j-1)\frac{n!(n-1)!}{(n+j-1)!(n-j)!}[1+(-1)^j]. \tag{12}$$

Coefficients W^n_{bj} in expression (10) can be efficiently computed with the use of recursive procedures (Paule and Schorn 1994; Petkovsek *et al.* 1996). Several recursions for W^n_{bj} are possible, depending on which index one decides to consider as the running one. We used the implementation of Zeilberger's algorithm in Mathematica, developed by P. Paule and M. Schorn, available at http://www.risc.uni-linz.ac.at/research/combinat/risc/software/, to obtain recursions for W^n_{bj}. We found the following recursive scheme, with respect to the index j, very useful:

$$W^n_{b2} = \frac{6}{(n+1)}, \tag{13}$$

$$W^n_{b3} = 30\frac{(n-2b)}{(n+1)(n+2)}, \tag{14}$$

$$W^n_{b,j+2} = -\frac{(1+j)(3+2j)(n-j)}{j(2j-1)(n+j+1)}W^n_{bj} + \frac{(3+2j)(n-2b)}{j(n+j+1)}W^n_{b,j+1}. \tag{15}$$

The above recursions are numerically stable and very fast. We used them, implemented in a standard double-precision arithmetic, for genealogies consisting of thousands of DNA sequences (the largest value of n tested was $n = 5000$). We did not perform precise measurements, but usually, when calculating probabilities q_{nb}, according to (8), computing coefficients W^n_{bj} and V^n_j takes only a small fraction of the time, while most of the computing effort is needed to evaluate expectations e_j.

Influence of the ascertainment procedure on SNP sampling frequencies: Most of the published data on SNP sampling frequencies are obtained in a two-step process, where the first step involves discovering chromosomal locations of a number of SNPs, and the second one involves DNA sequencing of a sample of n chromosomes restricted to locations discovered in the first step. The first step is called SNP ascertainment and is based on number of chromosomes smaller than n. As described in previous studies, taking into account the ascertainment scheme is a very important aspect of SNP data analysis. Below we derive expressions for modeling the way in which ascertainment modifies SNP sampling frequencies.

We use the following notation introduced by Wakeley *et al.* (2001): Data set size is $n = n_D + n_O$, and ascertainment set size equals to $n_O + n_A$, where n_A stands for the number of ascertainment-only samples; n_O, the number of overlapping samples (both in the ascertainment study and in the later data set); and n_D, the number of data-only samples.

To determine how ascertainment modifies probability distribution (22), we merge ascertainment and data sets to obtain the joint set of size $n_J = n_D + n_O + n_A$. We treat the ascertainment procedure as sampling SNP alleles, without replacement, from the joint set. A SNP is discovered if (a) both alleles are present in the ascertainment sample and (b) none of the alleles in the ascertainment sample has number of copies less than G, where G is a predetermined threshold. Since the joint set contains elements of two types (two alleles), the number of copies of alleles in the ascertainment sample follows a hypergeometric distribution. We analyze two cases: (i) no overlap, which means $n_O = 0$, $n = n_D$, $n_J = n_D + n_A$; and (ii) overlap only, which means $n_A = 0$, $n = n_J = n_D + n_O$. The case where both overlap and ascertainment-only samples are present is obtained by combining i and ii. We compute frequencies of discovered SNPs in the data set, which follow from conditions a and b above. We analyze first the case i. If a SNP in the joint set has b mutant and $n_J - b$ ancestral bases, then the probability that a sample of size n_A from the joint set has β mutant and $n_A - \beta$ ancestral bases is

$$h(\beta, n_J, b, n_A) = \frac{\binom{b}{\beta}\binom{n_J - b}{n_A - \beta}}{\binom{n_J}{n_A}}. \tag{16}$$

For a SNP to be discovered, β must satisfy $G \le \beta \le n_A - G$, with G defined as above. Moreover, the following inequalities must hold: $\beta \le b$, $n_A - \beta \le n_J - b$. Consequently, the probability $\pi^A_{n\gamma}$ that a discovered SNP in the data-only set i has $\gamma = b - \beta$ mutant and $n_D - \gamma$ ancestral alleles is

$$\pi^A_{n_D\gamma} = \frac{\sum_{\beta=G}^{n_A-G} q_{n_J\gamma+\beta}\, h(\beta, n_J, \gamma+\beta, n_A)}{\sum_{g=0}^{n_D}\sum_{\beta=G}^{n_A-G} q_{n_J g+\beta}\, h(\beta, n_J, g+\beta, n_A)}, \tag{17}$$

$\gamma = 0, 1, \ldots, n_D$. Probabilities q_{nb} are given by (8). The relation $\gamma = b - \beta$ follows from the fact that β chromosomes with mutant bases are removed from the joint set. The numerator in (17) is a sum of contributions to $\pi^A_{n_D\gamma}$ for possible values of β, while the denominator is a normalizing factor. For case ii assume again that a SNP in the joint set has b mutant and $n_J - b$ ancestral bases. The probability that a sample of n_O has β mutant and $n_O - \beta$ ancestral bases is given by (16) with n_A replaced by n_O. For this SNP to be discovered β must satisfy $G \le \beta \le n_O - G$. Consequently, the probability $\pi^O_{n_J b}$ that a discovered SNP in the joint set ii has b mutant and $n_J - b$ ancestral alleles is

$$\pi^{O}_{n_J b} = \frac{q_{n_J b}\sum_{\beta=G}^{n_O-G} h(\beta,\, n_J,\, b,\, n_O)}{\sum_{\sigma=G}^{n_J-G} q_{n_J\sigma}\sum_{\beta=G}^{n_O-G} h(\beta,\, n_J,\, \sigma,\, n_O)}, \qquad (18)$$

$b = G, \ldots, n_J - G$.

If it is not known which of the alleles is mutant and which one is ancestral, we need to symmetrize $\pi^{A}_{n_D\gamma}$ and $\pi^{O}_{n_J b}$ to get probability of data configuration. For case i we have expression

$$P(X^R = \gamma) = p^{A}_{n_D\gamma} = \pi^{A}_{n_D\gamma} + \pi^{A}_{n_D,\, n_D-\gamma}[1 - \delta(\gamma,\, n_D - \gamma)],$$
$$\gamma = 0, 1, \ldots, [n_D/2] \qquad (19)$$

for the probability that the rare allele has γ copies. For case ii the probability that there are b copies of the rare allele is

$$P(X^R = b) = p^{O}_{n_J b} = \pi^{O}_{n_J b} + \pi^{O}_{n_J, n_J - b}[1 - \delta(b,\, n_J - b)],$$
$$b = G,\, G+1, \ldots, [n_J/2]. \qquad (20)$$

In the above $[n/2]$ denotes greatest integer $\le n/2$.

In the sequel, we refer to the models described above as type i and type ii ascertainment, respectively.

Likelihood function of the sample: Data studied are derived from a number of SNP sites. Let us denote the number of SNP loci by M and random variables defined by diallelic data by

$$[X_1, X_2, \ldots, X_M] = [(X^R_1, X^F_1), (X^R_2, X^F_2), \ldots (X^R_m, X^F_m) \ldots, (X^R_M, X^F_M)], \qquad (21)$$

where X^R_m is the number of copies of the less frequent (rare) allele and X^F_m is the number of copies of the more frequent one, in the sample of $n_m = X^R_m + X^F_m$. It is possible that $X^R_m = X^F_m$ for some indices m, in which case both alleles are equally frequent. We assume that the ancestral state is not known. Then, for an SNP (X^R_m, X^F_m), the probability that we observe configuration b_m, $n_m - b_m$, $b_m \le [n_m/2]$ is

$$P(X^R_m = b_m) = p_{n_m b_m} = q_{n_m b_m} + q_{n_m, n_m - b_m}[1 - \delta(b_m,\, n_m - b_m)], \qquad (22)$$

where $\delta(\cdot)$ is the Kronecker delta function and q_{nb} are probabilities defined and evaluated in the previous section.

When SNP sites are located far from one another, random variables $\{X_1, X_2, \ldots, X_M\}$ in (21) are independent. If the observed numbers of copies of rare alleles are $X^R_1 = b_1$, $X^R_2 = b_2, \ldots, X^R_m = b_m, \ldots, X^R_M = b_M$, then the log likelihood of the sample (21) is

$$l = \sum_{m=1}^{M} \log(p_{n_m b_m}) \qquad (23)$$

(Nielsen 2000; Wooding and Rogers 2002). If sample sizes are equal for all SNP loci, $n_1 = n_2 = \ldots = n_M = n$, the above expression can be written as

$$l = \sum_{b=1}^{[n/2]} c_b \log(p_{nb}), \qquad (24)$$

where c_b denotes number of SNP loci in the sample, which have configuration of b copies of the rare allele *vs.* $n - b$ copies of the frequent allele. Subsequently, we use expressions (23) and (24) to compute likelihoods of SNP samples with different ascertainment models. To specify the ascertainment model we substitute in (23) or (24), $p_{nb} = p_{nb}$ [expression (22), no ascertainment step], $p_{nb} = p^{A}_{nb}$ [expression (19), ascertainment model type i], or $p_{nb} = p^{O}_{nb}$ [expression (20), ascertainment model type ii].

RESULTS

Exponential history of population size: In our computations we assume an exponential history of effective population size. In previous publications devoted to SNP and demography, Nielsen (2000) assumed an exponential history of $N_e(t)$. However, his analysis is very brief and restricted only to simulations. Others (Wakeley *et al.* 2001; Wooding and Rogers 2002) analyzed stepwise histories of effective population size changes.

For an exponential scenario of population growth

$$N_e(t) = N_{e0}\exp(-rt), \qquad (25)$$

expectations in (3) become

$$e_j = e_j(N_{e0},\, r) = -\frac{\exp\left[\binom{j}{2}(rN_{e0})^{-1}\right]}{r}\mathrm{Ei}\left[-\binom{j}{2}(rN_{e0})^{-1}\right] \qquad (26)$$

(Slatkin and Hudson 1991), where Ei denotes the exponential integral, $\mathrm{Ei}(-\mu) = -\int_1^\infty \exp(-\mu x)/x\, dx$, $\mathrm{Re}(\mu) > 0$ (Gradshteyn and Ryzhik 1980, Sect. 3.351.5). When the argument $\binom{l}{2}(rN_{e0})^{-1}$ in (26) becomes large, computing $e_l(N_{e0}, r)$ involves solving product of the type $\infty \cdot 0$. For $\binom{l}{2}(rN_{e0})^{-1} > 50$, we used expansion,

$$\mathrm{Ei}(-x) = \exp(-x)\sum_{k=1}^{K}(-1)^k\frac{(k-1)!}{x^k} + R_K \qquad (27)$$

(Gradshteyn and Ryzhik 1980, Sect. 8.215) with

$$|R_K| < \frac{K!}{|x|^{K+1}}, \qquad (28)$$

which allowed canceling $\exp\left[\binom{l}{2}(rN_{e0})^{-1}\right]$ in (26).

It turns out that sampling frequencies of SNPs depend only on the product parameter $\kappa = rN_{e0}$ of initial effective population size and exponential factor.

Distributions of SNP frequencies: Figure 3 provides examples of probabilities of different configurations of SNP sites, for sample size $n = 30$, and different values of the parameter κ (0, 1, 10), under the assumption that data collection did not include an ascertainment step [expression (22)] or under the ascertainment model of type ii [expression (20)] with $n_O = 10$, $G = 1$, or $G = 2$. As already reported in many articles, increasing κ results in higher proportions of rare alleles in the sample. Plots

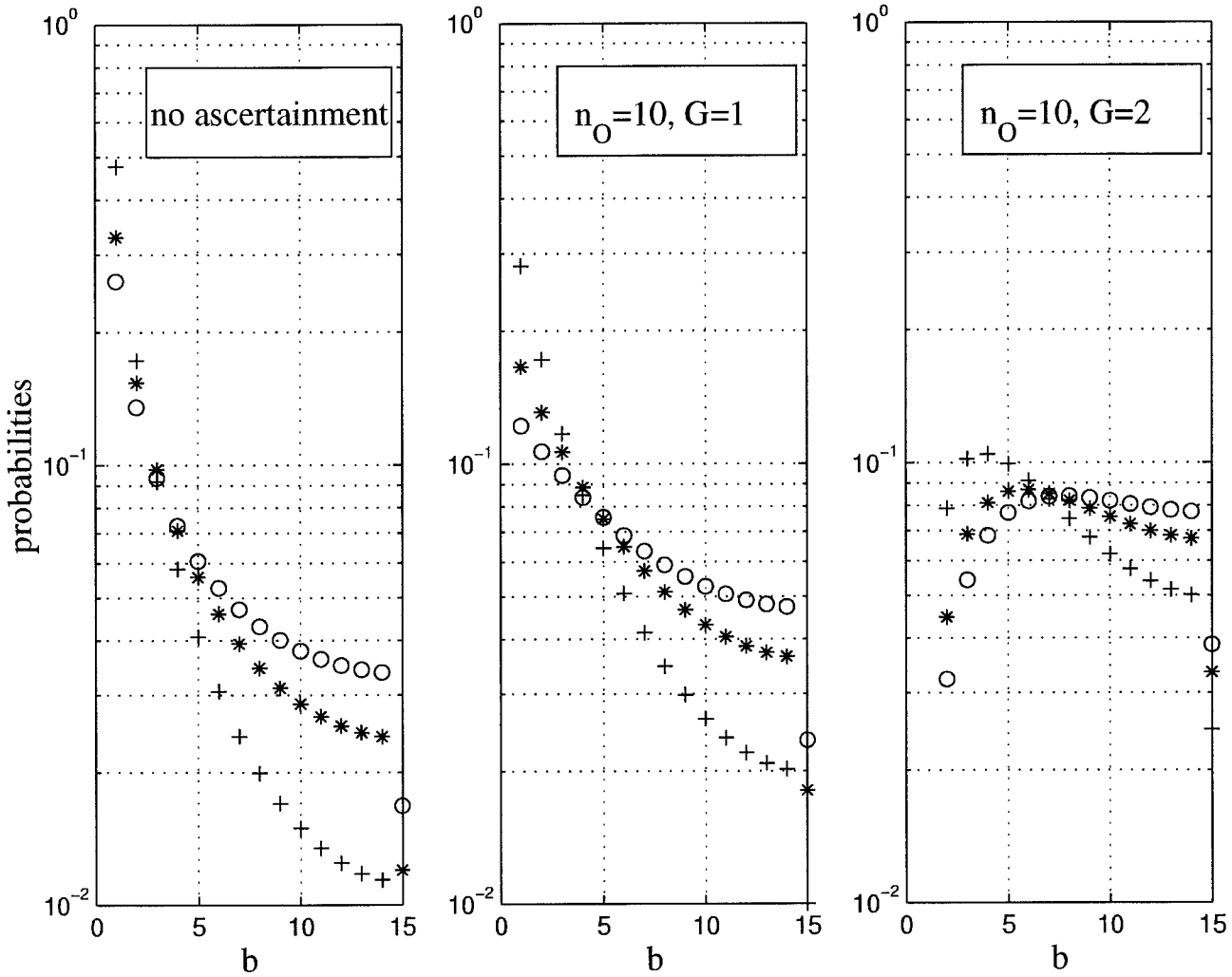

FIGURE 3.—Probabilities of different configurations of SNP sites, for sample size $n = 30$; different values of the parameter κ, $\kappa = 0$ (○), $\kappa = 1$ (*), and $\kappa = 10$ (+); under the assumption that data collection was without the ascertainment step [expression (22)] or with ascertainment model type ii [expression (20)] with $n_O = 10$, $G = 1$, or $G = 2$.

in Figure 3 also show how ascertainment modifies the distribution of SNP frequencies. Increasing the threshold value G flattens the distribution of frequencies. Both types of ascertainment (i and ii) have similar effects on SNP frequency distributions (results not shown).

Likelihood-ratio tests to detect signatures of population growth: An interesting issue is our power to test hypothesis H_0 of evolution with constant population size, $\kappa = \kappa_0 = 0$, against the alternative hypothesis H_1 of population expansion, $\kappa = \kappa_1 > 0$, on the basis of SNP data. It is also of interest to determine how this power is affected by the ascertainment step of data collection. From previous computations it follows that SNP data can be seen as samples from multinomial distributions given by expressions (22), (19), or (20). Assuming that the number of SNP sites is always large enough to allow asymptotic approximation (BICKEL and DOKSUM 2001, p. 227) we computed powers of single-value *vs.* single-value likelihood-ratio tests of statistical null hypothesis H_0 (constant population size $\kappa = \kappa_0 = 0$) *vs.* the alternative H_1 (population expansion with $\kappa = \kappa_1 > 0$). We assumed significance level $\alpha = 0.05$ and values of κ_1, $\kappa_1 = 0.1$, $\kappa_1 = 1$, $\kappa_1 = 10$, $\kappa_1 = 100$. Table 1, A and B, gives powers of likelihood-ratio tests for sample size $n = 50$, for different models of ascertainment: no ascertainment [probabilities given by expression (22)] or ascertainment model type ii [expression (19)] with parameters n_O and G. Table 1A is for the number of SNP loci $M = 30$, and Table 1B is for $M = 100$. From values of powers of tests depicted in Table 1, A and B, one can see that the cases $\kappa_0 = 0$, $\kappa_1 = 0.1$ are practically indistinguishable; $\kappa_0 = 0$, $\kappa_1 = 1$ may be distinguished only for a large enough number of SNP sites, while $\kappa_0 = 0$, $\kappa_1 = 10$, or $\kappa_1 = 100$ are rather easily distinguishable even for small numbers of SNPs. The ascertainment step in data collection can deteriorate the power to detect signatures of population growth. Increasing the threshold value of G, the aim of which typically is filter-

TABLE 1

Powers of likelihood-ratio tests

	No ascertainment	$n_O = 10$, $G = 1$	$n_O = 10$, $G = 2$	$n_O = 10$, $G = 3$	$n_O = 10$, $G = 4$
	A. No. of SNP loci $M = 30$				
$\kappa_1 = 0.1$	0.0737	0.0736	0.0670	0.0620	0.0586
$\kappa_1 = 1$	0.3072	0.3058	0.2129	0.1534	0.1174
$\kappa_1 = 10$	0.9654	0.9573	0.7635	0.5160	0.3353
$\kappa_1 = 100$	1.0000	1.0000	0.9733	0.7678	0.5061
	B. No. of SNP loci $M = 100$				
$\kappa_1 = 0.1$	0.1004	0.0973	0.0817	0.0714	0.0647
$\kappa_1 = 1$	0.6919	0.6534	0.4290	0.2755	0.1866
$\kappa_1 = 10$	1.0000	1.0000	0.9929	0.8848	0.6448
$\kappa_1 = 100$	1.0000	1.0000	1.0000	0.9905	0.8615

Null hypothesis H_0 (constant population size $\kappa = \kappa_0 = 0$) *vs.* the alternative H_1 (population expansion with $\kappa = \kappa_1 > 0$) is shown. Significance level is $\alpha = 0.05$; sample size is $n = 50$. Models of ascertainment are no ascertainment [probabilities given by expression (22)] or ascertainment model ii [expression (19)] with parameters n_O and G.

ing out sequencing errors in the data, also progressively lowers the probability of rejecting the hypothesis H_0 of constant population size; *i.e.*, it increases the probability of committing type II error. This results from the flattening effect of increasing G observed in Figure 3.

Data analysis: *Data on segregating sites in mitochondrial DNA from* Cann *et al. (1987):* First, we apply our method to the data on segregating sites in mitochondrial DNA from the article by Cann *et al.* (1987). We fit the exponential scenario (25) to these data by treating each segregating site as an independent SNP. Technically, we estimate the product parameter $\kappa = rN_{e0}$ in (25).

Data in Cann *et al.* (1987) include 195 segregating sites in 148 individuals. Table 2 shows the statistics of segregating sites in these data. Elements in the first column (b) are possible numbers of copies of the rare allele, and elements in the second column (c_b) are numbers of segregating sites in the sample that have the number of copies of the rare allele equal to b. Figure 4 shows the plot of log-likelihood function obtained by using expressions (8–15) and (26). Maximum of the log-likelihood function is attained at $\hat{\kappa} = 80$. The 95% confidence interval for this estimate, obtained with the use of likelihood-ratio statistics (Bickel and Doksum 2001), is $\kappa \in [40, 166]$.

Segregating sites collected by Cann *et al.* (1987) are obtained from nonrecombining DNA and the independence assumption is clearly not satisfied. To explore whether a violation of the assumption that SNPs are independent significantly affects the estimate of the parameter κ, we have performed 100 coalescent simulations of genealogies representing ancestries for 148 mtDNA sequences. We added mutations along branches of coalescence trees according to the infinite-sites model with intensity μ. In the simulations we assumed mutational time scale $\tau = 2\mu t$ and exponential change of $\theta(\tau) = 2N_e(\tau)\mu$, $\theta(\tau) = \theta_0 \exp(-\rho\tau)$, with parameters $\theta_0 = 400$, $\rho = 0.2$. So, the true value of the product parameter κ was $\kappa = 80$. For each of these 100 simulation experiments we treated segregating sites as independent SNPs and we estimated the parameter κ by maximizing likelihood (24). We obtained the mean of estimates equal to 86.8 and standard deviation equal to 29.7. This confirms that our approach, at least for these specific

TABLE 2

Statistics of segregating sites in mtDNA data

b	c_b
1	98
2	31
3	21
4	6
5	9
6	5
7	4
8	1
10	3
11	2
12	2
13	3
14	2
16	1
26	2
35	1
43	1
58	1
62	1
67	1

Based on Cann *et al.* (1987), elements in b are possible numbers of copies of the rare allele, and elements in c_b are numbers of segregating sites in the sample that have the number of copies of the rare allele equal to b.

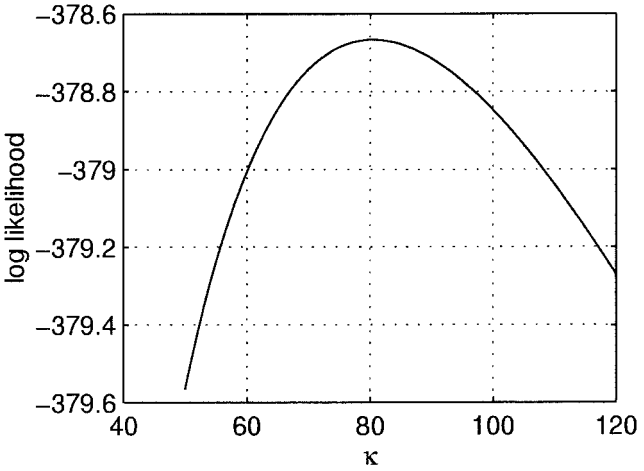

FIGURE 4.—Log-likelihood curve for the exponential model of population growth for data on segregating sites in mtDNA from CANN *et al.* (1987). Each segregating site was treated as a separate SNP. The maximum is attained at $\hat{\kappa} = 80$.

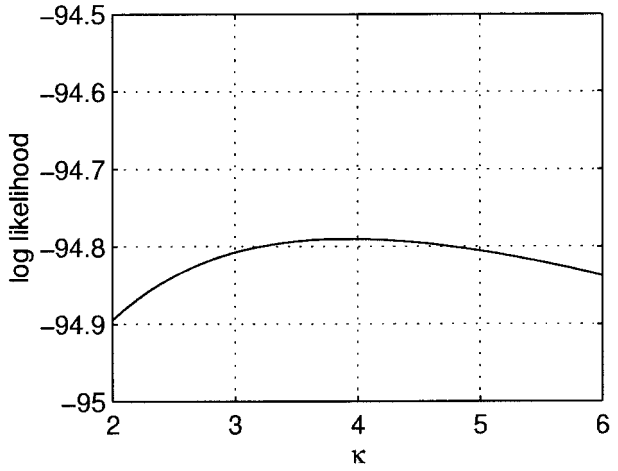

FIGURE 5.—Log-likelihood curve for the exponential model of population growth for SNP data from PICOULT-NEWBERG *et al.* (1999). The maximum is attained at $\hat{\kappa} = 3.7$.

values, will allow us to obtain a reasonable estimate of κ. From these simulations follows the estimate of 95% confidence interval for the parameter κ, when the sample consists of 148 mtDNA sequences and the demography is as shown above. This estimate, [mean − 2 standard deviations, mean + 2 standard deviations], equals $\kappa \in [27, 147]$. This estimate is quite consistent with the 95% confidence interval obtained from likelihood-ratio statistics, $\kappa \in [40, 166]$. The shift toward the left of the confidence interval based on simulations results from the asymmetric shape of the distribution of the estimate of κ. By applying a logarithmic transformation to simulation results (estimates of κ) we were able to obtain almost perfect agreement of the two confidence intervals, [40, 166] and [45, 175].

SNP data from PICOULT-NEWBERG et al. (1999) and TRIKKA et al. (2002): There are several population studies in the literature where relative frequencies of SNP alleles are shown. We have chosen data from the research by PICOULT-NEWBERG *et al.* (1999) and data on SNPs in three human genes: BLM, WRN, and RECQL, reported recently by TRIKKA *et al.* (2002). In our analysis, we used the data on Caucasians from both sources. The first reason to focus on Caucasians was the possibility of comparing two results, and the second reason was that discovery samples were from Caucasians. PICOULT-NEWBERG *et al.* (1999, Table 4) have 44 SNP sites in 8 Caucasians (16 chromosomes), while TRIKKA *et al.* (2002, Table 2) show allele frequencies of a total number of 31 SNPs in samples of chromosomes of sizes varying from 154 to 158.

When analyzing SNP data we followed remarks given in the source articles (PICOULT-NEWBERG *et al.* 1999; TRIKKA *et al.* 2002) to adjust parameters n_A, n_O, and G of the model of ascertainment procedure. We modeled the ascertainment procedure for collecting data from PICOULT-NEWBERG *et al.* (1999) by using expression (17) with $n_A = 10$ and $G = 2$. A plot of log-likelihood function for the data on Caucasians from the articles by PICOULT-NEWBERG *et al.* (1999) is shown in Figure 5. It attains a maximum at $\hat{\kappa} = 3.9$, with the 95% confidence interval, obtained with the use of likelihood-ratio statistics, $\kappa \in [0, 105.3]$. Log-likelihood function for the data on Caucasians from TRIKKA *et al.* (2002) is plotted in Figure 6. Ascertainment was modeled by expression (18) with $n_O = 10$ and $G = 1$. The maximum-likelihood estimate of the product parameter, from the plot in Figure 6, is $\hat{\kappa} = 0.78$, with the 95% confidence interval, obtained with the use of likelihood-ratio statistics, $\kappa \in [0, 6.1]$.

Sensitivity of estimates to ascertainment model parameters: A question arises: How sensitive are the estimates of parameter κ to changes of the model of the ascertainment? We studied this question by increasing or decreasing the value of the threshold G in expressions (17) and (18). Indexing the estimated parameter with n_A, n_O, and G, we can denote our estimates from the previous section as

$$\hat{\kappa}_{[n_A=10,G=2]} = 3.9 \quad \text{(PICOULT-NEWBERG } \textit{et al.} \text{ 1999)} \tag{29}$$

and

$$\hat{\kappa}_{[n_O=10,G=1]} = 0.78 \quad \text{(TRIKKA } \textit{et al.} \text{ 2002)}. \tag{30}$$

Here we compute estimates $\hat{\kappa}_{[n_A=10,G=1]}$, $\hat{\kappa}_{[n_A=10,G=3]}$ on the basis of data from PICOULT-NEWBERG *et al.* (1999) and $\hat{\kappa}_{[n_O=10,G=0]}$, $\hat{\kappa}_{[n_O=10,G=2]}$ on the basis of data from TRIKKA *et al.* (2002). Analysis of data from TRIKKA *et al.* (2002) requires more comment. The model to estimate $\hat{\kappa}_{[n_O=10,G=0]}$ assumes that no ascertainment procedure is taken into account. The model to estimate $\hat{\kappa}_{[n_O=10,G=2]}$ is inconsistent with complete data of TRIKKA *et al.* (2002) in the sense that the data contain one SNP locus with $b = 1$.

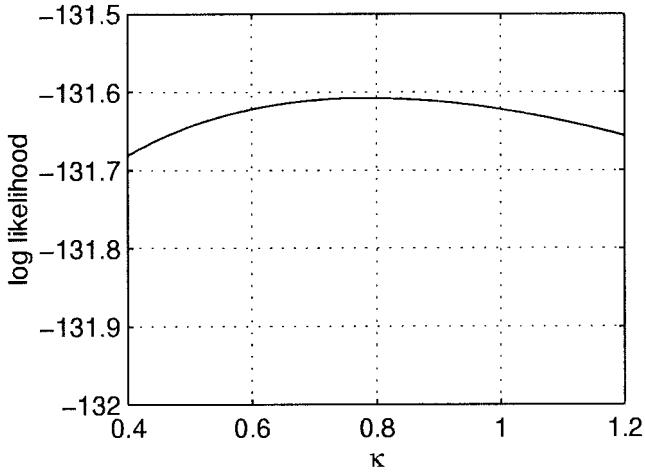

FIGURE 6.—Log-likelihood curve for the exponential model of population growth for SNP data from the article by TRIKKA *et al.* (2002). The maximum is attained at $\hat{\kappa} = 0.78$.

To apply the model $\hat{\kappa}_{[n_O=10,G=2]}$ we have removed this one locus.

The results of computations show an extreme sensitivity of estimates to the ascertainment model. Notably,

$$\hat{\kappa}_{[n_A=10,G=1]} = 0,$$

$$\hat{\kappa}_{[n_A=10,G=3]} = \infty \quad \text{(PICOULT-NEWBERG } \textit{et al.} \text{ 1999)} \quad (31)$$

and

$$\hat{\kappa}_{[n_O=10,G=0]} = 0,$$

$$\hat{\kappa}_{[n_O=10,G=2]} = 683 \quad \text{(TRIKKA } \textit{et al.} \text{ 2002)}. \quad (32)$$

In (31), by $\hat{\kappa}_{[n_A=10,G=3]} = \infty$ we meant that the likelihood function was increasing for values of κ up to 10^8.

The fact that the ascertainment model strongly affects estimates of parameters is also confirmed in the previous articles on SNPs. WAKELEY *et al.* (2001) in their Figure 3 show a large bias in SNP frequencies resulting from ascertainment. Similarly, NIELSEN (2000) uses a rather oversimplified model, $n_O = 2$, $G = 1$, for data (PICOULT-NEWBERG *et al.* 1999) and obtains $\hat{\kappa} = 0$. The need for careful modeling of ascertainment is also stressed by KUHNER *et al.* (2000).

DISCUSSION

The methods developed in this article allow us to analyze large data sets and carry out computations for different parameter values, which helps us draw more conclusions from data. We have shown examples of applying our methodology to the study of several issues arising in SNP data analysis.

We are particularly interested in the problem of what are reasonable values of the exponential growth product parameter $\kappa = rN_{e0}$ obtained on the basis of DNA data. Insight into this problem can be gained by comparing estimates obtained using different approaches.

Our aim when estimating κ from relative frequencies of segregating sites in the article by CANN *et al.* (1987) was to confirm that different methodologies used for the same data will still lead to comparable results. Therefore, we compared our estimate, obtained by treating nonrecombining segregating sites as SNPs, to those previously obtained on the basis of the same or similar data, but with the use of different methods. Studies that we compared to ours were those by SLATKIN and HUDSON (1991), ROGERS and HARPENDING (1992), and POLANSKI *et al.* (1998), who used pairwise difference statistics, and by WEISS and VON HAESELER (1998), who applied the maximum-likelihood approach. Data in these articles originate from different sources, but considering estimations of the authors, reasonable ranges of the product parameter κ, for both the worldwide population and Caucasians, fit into the interval from $\kappa = 50$ to $\kappa = 500$. Our estimate of $\hat{\kappa} = 80$ is consistent with the above ranges.

Mutation intensity (per site) at autosomal loci is approximately one order of magnitude lower than that in mtDNA (LI 1997). However, the estimate of the product parameter $\kappa = rN_{e0}$ is invariant with respect to timescale changes and therefore does not depend on the value of the mutation intensity. We can assume that mutation intensity is used only to scale the time axis. The effective population size for autosomal loci is four times the effective population size for loci at mtDNA. So, the estimate of κ from mtDNA should be one-fourth the estimate of κ from nuclear DNA. Taking into account the large stochastic variation, the estimates of κ coming from SNP data should then be comparable (of the same order of magnitude) to those obtained from mtDNA.

However, our estimates of the parameter κ based on SNP data, $\hat{\kappa} = 3.9$ and $\hat{\kappa} = 0.78$, are markedly smaller than values coming from mtDNA, which runs counter to the expected tendency. Differences between our estimates and the above ranges can be, probably, attributed to two factors. The first one, mentioned by WOODING and ROGERS (2002), is that some fraction of SNPs in the data could be under balancing selection, which would shift their frequencies toward higher values and move the estimate of κ toward lower values. The second factor, which comes from our analysis, is the sensitivity to the parameters of the ascertainment model, shown in (31) and (32). With this high sensitivity, even a small unmodeled factor resulting from eliminating some low-frequency SNPs by assuming that they were sequencing errors can lead to estimates substantially lower than the true value of κ.

The authors are grateful to Peter Paule and Markus Schorn for making their program, implementing Zeilberger's algorithm in Mathematica, available to the scientific community. The authors were supported by National Institutes of Health grants GM58545 and CA75432, Polish Scientific Committee (KBN) research projects PBZ/KBN/040/

P04/2001 and 4T11F 01824, and NATO collaborative linkage grant LST.CLG.977845.

LITERATURE CITED

Altshuler, D., V. J. Pollar, C. R. Cowles, W. J. Van Etten, J. Baldwin *et al.*, 2000 A SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407:** 582–589.

Bickel, P. J., and K. A. Docksum, 2001 *Mathematical Statistics: Basic Ideas and Selected Topics.* Prentice Hall, Upper Saddle River, NJ.

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

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

Cann, R. L., M. Stoneking and A. C. Wilson, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231–238.

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

Durrett, R., and V. Limic, 2001 On the quantity and quality of single nucleotide polymorphisms in the human genome. Stoch. Proc. Appl. **93:** 1–24.

Eberle, M. A., and L. Kruglyak, 2000 An analysis of strategies for discovery of single nucleotide polymorphisms. Genet. Epidemiol. **19** (Suppl 1): S29–S35.

Graham, R. L., D. E. Knuth and O. Patashnik, 1998 *Concrete Mathematics. A Foundation for Computer Science*, Ed. 2. Addison-Wesley, Reading, MA.

Gradshteyn, I. S., and I. M. Ryzhik, 1980 *Table of Integrals, Series and Products*, Ed. 2. Academic Press, San Diego.

Griffiths, R. C., and S. Tavare, 1998 The age of a mutation in the general coalescent tree. Stoch. Models **14:** 273–295.

Fu, X.-Y., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172–197.

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood pressure homeostasis. Nat. Genet. **22:** 239–247.

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

Kuhner, M. K., P. Beerli, J. Yamamoto and J. Felsenstein, 2000 Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics **156:** 439–447.

Li, W. H., 1997 *Molecular Evolution.* Sinauer Associates, Sunderland, MA.

Marth, G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu *et al.*, 1999 A general approach to single-nucleotide polymorphism discovery. Nat. Genet. **23:** 452–456.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Paule, P., and M. Schorn, 1994 A Mathematica version of Zeilberger's algorithm for proving binomial coefficients identities. J. Symbol. Comput. **11:** 1–25.

Petkovsek, M., H. S. Wilf and D. Zeilberger, 1996 *A=B.* A. K. Peters, Wellesley, MA (http://www.cis.upenn.edu/~wilf/AeqB.html).

Picoult-Newberg, L., T. E. Ideker, M. G. Pohl, S. L. Taylor, M. A. Donaldson *et al.*, 1999 Mining SNPs from EST databases. Genome Res. **9:** 167–174.

Polanski, A., M. Kimmel and 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 **95:** 5456–5461.

Polanski, A., A. Bobrowski and M. Kimmel, 2003 A note on distributions of times to coalescence under time-dependent population size. Theor. Popul. Biol. **63:** 33–40.

Renwick, A., P. Bonnen, D. Trikka, D. Nelson, R. Chakraborty *et al.*, 2003 Sampling properties of estimators of nucleotide diversity at discovered SNP sites. Appl. Math. Comp. Sci. (in press).

Rish, N. J., 2000 Searching for genetic determination in the new millenium. Nature **405:** 847–856.

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

Sherry, S. T., H. C. Harpending, M. A. Batzer and M. Stoneking, 1997 Alu evolution in human populations: using the coalescent to estimate effective population size. Genetics **147:** 1977–1982.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA in stable and exponentialy growing populations. Genetics **129:** 555–562.

Trikka, D., Z. Fang, A. Renwick, S. H. Jones, R. Chakraborty *et al.*, 2002 Complex SNP-based haplotypes in three human helicases: implication for cancer association studies. Genome Res. **12:** 627–639.

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

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

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.

Weiss, G., and A. von Haeseler, 1998 Inference on population history using a likelihood approach. Genetics **149:** 1539–1546.

Wooding, S., and A. Rogers, 2002 The matrix coalescent and an application to human single-nucleotide polymorphisms. Genetics **161:** 1641–1650.

Yang, Z., G. Wong, M. A. Eberle, M. Kibukawa, D. A. Passey *et al.*, 2000 Sampling SNPs. Nat. Genet. **26:** 13–14.

Communicating editor: N. Takahata