Modeling Evolution of Relative Frequencies of Single Nucleotide Polymorphisms

A. Polanski
Department of Automatic Control
Silesian Technical University
Gliwice, Poland

M. Kimmel
Department of Statistics
Rice University
Houston, USA

ABSTRACT
We present new methods for modeling sampling distributions of SNP frequencies in populations with time varying size. Computation based on our methods are more efficient than the Monte Carlo simulations and allow for analysis of large genealogies. We analyze some of available SNP data and we compare our results of demographic parameters estimation to those obtained in previous researches in population genetics. The analyzed data are not inconsistent with the hypothesis of past population growth of modern humans.

KEY WORDS
SNPs, Coalescent, Demography, Hypergeometric summation

1 Introduction
Single Nucleotide Polymorphisms (SNP) are single base changes in DNA. They are very promising genetic markers due to their high density in the human genome. Publicly available SNP databases, like SNP Consortium at http://snp.cshl.org, contain massive amounts of data on positions of SNPs in human genome and constantly increase in number and size. It is believed that eventually SNPs will enable creating fine genetic maps for complex traits analysis [8], [17].

Considering the above, modeling evolution of SNPs is a very important scientific problem. Several interesting studies were already carried out in this area. Studies [3] and [23] 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 in [16], [26]. Using SNPs for estimation of the scaled product parameter \( \theta = 4N_e \mu \) of effective population size \( N_e \) and mutation rate \( \mu \), under assumption of constant population was studied in [9]. They have taken into account various hypotheses of spatial (chromosomal) distributions of SNPs: complete or partial linkage, or linked segments of nonrecombining SNPs, and, based on extensive simulations, evaluated accuracy of estimates and possible sources of bias. Studies [12] and [21] were devoted to detection of signatures of human population growth in SNP data. In the paper [12], the scenario of exponential expansion was fitted to SNP data from [13]. In [21], the model of stepwise change of the population size with population subdivision from [22] was fitted to SNP data from [23], [2] and [1]. Parameter-space regions corresponding to highest likelihoods were not inconsistent with the hypothesis of population growth. Moreover, if the ascertainment bias was not considered, less likely shapes of parameter regions were obtained. Comparison of cases in which population substructure was not considered to those in which it was considered, seemed to support the latter scenario. In order to evaluate SNP frequencies, these studies used the standard coalescent 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 elaborated in the papers [6], [25] and [15]. However, existing approaches suffer from one serious difficulty: numerical instability for larger genealogies. When the analyzed genealogy size is greater than 50 the existing analytical methods are either inapplicable or difficult to apply, due to numerical instabilities. Here we show how techniques of summation of hypergeometric series can be used to remove the described problem. We derive simple recursive expressions to compute SNP frequencies in populations with time varying size. Our method allows performing tasks which 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. We also provide expressions which describe the influence of the discovery procedure on SNP frequencies. Using our methods we analyze some of the available SNP data [13], [20].

2 Methods
We accept standard coalescent assumptions. Our notation is shown in fig. 1, for the sample size \( n = 5 \). Time \( t \) is measured, in number of generations, from the present to the past. Random coalescence times, 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 lower case letters \( t_n, t_{n-1}, \ldots, t_2, 0 < t_n < t_{n-1} \ldots < t_2 \). Times between
coalescence events are denoted by \( S_n, S_{n-1}, \ldots, S_2, \) and \( \rho_n, \rho_{n-1}, \ldots, \rho_2. \) Our notation is not standard since it is more common to denote times between coalescence events by \( T, t. \) In the notation we followed [15]. When referencing times \( T_k \) or \( S_k \) we leave out index \( n \) that would define the sample size. The underlying value of \( n \) is always clear from the context.

2.1 Probability that SNP has \( b \) mutant bases

Probability \( q_{nb} \) that a SNP site in a sample of \( n \) chromosomes has \( b \) mutant bases, is given [6] (eq. (1.3)) in the terms of expectations of times in the coalescence tree. In our notation, this expression has the following form

\[
q_{nb} = \frac{1}{[n]!} \sum_{k=2}^{n} k(k-1) E(S_k),
\]

where \( 0 < b < n, S_k = T_k - T_{k+1}, T_{n+1} = 0. \)

Let us assume that the effective population size history is described by a function

\[
N_e(t), \ t \in [0, \infty).
\]

For a random sample of \( n \) DNA sequences, joint probability density function of the distribution of their coalescence times \( T_n, T_{n-1}, \ldots, T_2 \) is given by the expression [7]

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

Marginal distributions \( p_n(t_1), p_{n-1}(t_{n-1}), \ldots, p_2(t_2) \) of times \( T_n, T_{n-1}, \ldots, T_2 \) were computed in [15], as follows

\[
p_k(t_k) = \sum_{j=k}^{n} A_j^k q_j(t_k),
\]

where

\[
q_j(t) = \left( \frac{1}{N_e(t)} \right) \exp \left( - \int_{0}^{t} \left( \frac{1}{N_e(\sigma)} \right) d\sigma \right).
\]

and

\[
A_j^k = \frac{\prod_{i=1}^{n} (i)}{\prod_{i=1}^{n} (i) (i-1) \sigma_j (t_j)} k \leq j \leq n,
\]

\( A_n^n = 1. \) Denote by \( e_j \) the expected value

\[
e_j = \int_{0}^{\infty} t q_j(t) dt.
\]

From (4) and (6) it follows that expected values of times \( E(S_k) \) that appear in equation (1) can be expressed as

\[
E(S_k) = \sum_{j=k}^{n} \frac{j(j-1)}{k(k-1)} A_j^k e_j.
\]

and consequently probabilities \( q_{nb} \) become

\[
q_{nb} = \frac{\frac{1}{[n-b]!} (n-b-1)!}{\sum_{k=2}^{n} \sum_{j=k}^{n} j(j-1) \frac{(n-k)!}{(n-1)!} A_j^k e_j}
\]

\[
\sum_{k=2}^{n} \sum_{j=k}^{n} \frac{j(j-1)}{k(k-1)} A_j^k e_j.
\]

The above is an analytic expression for probabilities \( q_{nb}. \) However, as reported in [15] and [25], it is rather difficult to efficiently apply this formula for genealogies of size \( n > 50 \) because of diverging coefficients with alternating signs.

2.2 Method for efficient computation of \( q_{nb} \) for large genealogies

In order to avoid large numerical errors in summations in (9) one needs to apply computation with precision of hundreds or even thousands of decimal digits [25] which significantly slows down computational process and requires appropriate software.

Below we present a simple method for efficient computing \( q_{nb} \) for genealogies of practically arbitrarily large size. The idea is to change the order of summation in both denominator and numerator in equation (9) and observe that the resulting expressions contain sums of hypergeometric series. Proceeding as described we obtain

\[
q_{nb} = \frac{\sum_{j=2}^{n} e_j \sum_{k=2}^{j} j(j-1) \frac{(n-k)!}{(n-1)!} A_j^k e_j}{\sum_{j=2}^{n} e_j \sum_{k=2}^{j} j(j-1) \frac{1}{k(k-1)} A_j^k e_j}.
\]

\[
q_{nb} = \frac{\sum_{j=2}^{n} e_j W^j}{\sum_{j=2}^{n} e_j V^j}.
\]

In the above we introduced coefficients

\[
W_{bj} = \sum_{k=2}^{j} j(j-1) \frac{(n-k)!}{(n-1)!} A_j^k e_j,
\]

and

\[
V_{bj} = \sum_{k=2}^{j} j(j-1) \frac{1}{k(k-1)} A_j^k e_j.
\]
Expressions in (11) and (12) are sums of hypergeometric series, which can be seen by factoring the denominators in (6), \( \frac{j}{(j)} \frac{j}{(j)} = \frac{1}{2} (l+j+1) \), and then expressing coefficients \( A^k_j \) in (6) as follows:

\[
A^k_j = \frac{n! (n-1)!}{(n+j-1)! (n-j)!} \frac{(2j-1)}{j(j-1)} \frac{(j+k-1)!}{(k-1)! (j+k-1)!}\frac{(-1)^j}{(2j-1)}.
\]

(13)

Substituting (13) in (12) and using Chu-Vandermonde identity [5], we obtain

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

(14)

Expressions (15) and (17) are obtained by using the Chu-Vandermonde identity again, while (16) follows from the fact that (11) reduces to a simple term when \( b = n - 1 \). Using the Pascal triangle property \( \binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \) we can then see that coefficients \( W^n_{b} \) satisfy the following relation

\[
W^n_{b+1} = \frac{n+1}{(n+j)! (n+j+1)!} (n-b) W^n_{b} + (b+1) W^n_{b+1}.
\]

(18)

Expressions (15), (16), (17) and (18) allow computing coefficients recursively, starting from \( n = 2 \), up to the genealogy of the desired size. Provided that multiplications in (14), (15) and (17) are properly numerically organized, the recursive procedure will not suffer from numerical instabilities.

Coefficients \( V^n_j \) and \( W^n_{b} \) can be applied to compute probabilities \( q_{nb} \), according to (10), for any history 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.

Numerical accuracy of computations can be verified by connecting (18) with (9). For all computations performed we have checked, with relative accuracy \( 10^{-8} \) that probabilities \( q_{nb} \) sum up to 1 when \( b \) takes values 1, 2, 3, \ldots, \( n - 1 \).

2.3 Exponential history of population size

For an exponential scenario of population growth

\[
N_e(t) = N_{e0} \exp(-rt)
\]

(19)

expectations in (7) become [19]

\[
e_j = e_j(N_{e0}, r) = \frac{- \exp[(\frac{t}{2}) (rN_{e0})^{-1}]}{r} E_i[(\frac{t}{2}) (rN_{e0})^{-1}]
\]

(20)

where \( E_i \) denotes the exponential integral, \( E_i(-\mu) = - \int_{\mu}^{\infty} \exp(-x) / x \, dx \), \( \Re(\mu) > 0 \), [4] (§ 3.351.5). When the argument \( (\frac{t}{2}) (rN_{e0})^{-1} \) in (20) becomes large, computing \( e_j(N_{e0}, r) \) involves solving product of the type \( \infty \cdot 0 \). For \( (\frac{t}{2}) (rN_{e0})^{-1} \) \( > 50 \) we used expansion [4] (§ 8.215)

\[
E_i(-x) = \exp(-x) \sum_{k=0}^{K} \frac{(-1)^k (k-1)!}{x^k} + R_K
\]

(21)

with

\[
|R_K| < \frac{K!}{x^{K+1}},
\]

(22)

which allowed canceling first factor \( \exp[(\frac{t}{2}) (rN_{e0})^{-1}] \).

3 Results

We assume that data under study come from a number of SNP sites. Let us denote number of SNP loci by \( M \), and random variables defined by di allelic data by

\[
[X_1, X_2, \ldots, X_M] = [(X_1^R, X_1^F), \ldots, (X_m^R, X_m^F), \ldots, (X_M^R, X_M^F)]
\]

(23)

where \( X_m^R \) is the number of copies of the less frequent (rare) allele, \( X_m^F \) - number of copies of more frequent one, in the sample of \( n_m \) copies of each. It is possible that \( X_m^R = X_m^F \) for some indices \( m \) in which case both alleles are equally frequent. We assume that ancestral state is not known. Then, for an SNP \( (X_m^R, X_m^F) \), the probability that we observe configuration \( b_m, n_m, b_m = b_m, \quad b_m \leq \lfloor \frac{n_m}{2} \rfloor \) is

\[
P(X_m^R = b_m) = p_{n_m} b_m = q_{n_m} b_m
\]

(24)

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

3.1 Likelihood function

When SNP sites are located far from one another, random variables \( \{X_1, X_2, \ldots, X_M \} \) in (23) are independent. If the observed numbers of copies of rare alleles are \( X_m^R = b_1, X_m^F = b_2, \ldots, X_m^R = b_m, \ldots, X_M^R = b_M \) then the likelihood of the sample (23) is equal to the product

\[
L = \prod_{m=1}^{M} p_{n_m} b_m
\]

(25)
3.2 SNP data from [13] and [20]

There are several population studies in the literature where relative frequencies of SNP alleles are shown. We have chosen to use data from the research [13], and data on SNPs in three human genes: BLM, WRN and RECQL, reported recently in [20]. In our analysis we used the data on Caucasians from both sources. The first reason was because the choice of a mutation to focus on Caucasians was a possibility to compare two results, and the second reason was that discovery samples were from Caucasians. In [13] (Table 4) 44 SNP sites in 8 genes were used in [20] (Table 2) allele frequencies are given for total number of 31 SNPs in samples of chromosomes of sizes varying from 154 to 158.

As described in previous studies, a very important aspect of SNP data analysis is taking into account their ascertainment scheme. Here we use the following notation from [21]: 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. In order to determine how ascertainment modifies probability distribution (24) we merge ascertainment and data sets to obtain the joint set of size $n_J = n_D + n_O + n_A$. We treat ascertainment procedure as sampling without replacement from the joint set. Since the joint set contains elements of two types (two alleles) then the number of copies of alleles in 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 $n_O = 0$, $n = n_J = n_D + n_O$. 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 $< G$. The case where both overlap and ascertainment only samples are present is obtained by combining (i) and (ii).

Let us assume that a SNP has $b$ mutant and $n_I - b$ ancestral bases. We analyze first the case (i). Probability that a sample of $n_A$ has $b$ mutant and $n_A - b$ ancestral bases is

$$h(\beta, n_J, b, n_A) = \binom{b}{\beta} \binom{n_J - b}{n_A - \beta}$$

(26)

Moreover, the following inequalities must hold: $G \leq \beta \leq b, G \leq n_A - \beta \leq n_J - b$. Consequently, the probability $\pi_{n_D, \gamma}^A$ that a SNP in the data only set (i) has $\gamma = b - \beta$ mutant and $n_D - \gamma$ ancestral alleles is

$$\pi_{n_D, \gamma}^A = \frac{\sum_{\gamma = 0}^{n_O} \sum_{\beta = G}^{n_A - \beta} h(\beta, n_J + \beta, \gamma + \beta, n_A)}{\sum_{\gamma = 0}^{n_O} \sum_{\beta = G}^{n_A - \beta} h(\beta, n_J + \beta, n_A)}$$

(27)

$\gamma = 0, \ldots, n_D$. For the case (ii) probability that a sample of $n_O$ has $\beta$ mutant and $n_O - \beta$ ancestral bases is given by (26) with $n_A$ replaced by $n_O$, and the probability $\pi_{n_J, b}^O$ that a SNP in the joint set (ii) has $b$ mutant and $n_J - b$ ancestral alleles is

$$\pi_{n_J, b}^O = \frac{\sum_{\beta = G}^{n_O - b} \sum_{\gamma = 0}^{n_J - b} h(\beta, n_J - \beta, \gamma, n_A)}{\sum_{\beta = G}^{n_O - b} \sum_{\gamma = 0}^{n_J - b} h(\beta, n_J - \beta, \gamma, n_A)}$$

(28)

$b = G, \ldots, n_J - G$. Since it is not known which of the alleles is mutant and which one is ancestral, we need to symmetrize $\pi_{n_D, \gamma}^A$ and $\pi_{n_J, b}^O$ as in expression (24) to get probability of data configuration. For the case (i) we have expression

$$P \left( X^R = \gamma \right) = \pi_{n_D, \gamma}^A$$

$$\pi_{n_D, \gamma}^A = \pi_{n_D, \gamma}^A + \pi_{n_D, n_D - \gamma}^A \left[ 1 - \delta(\gamma, n_D - \gamma) \right]$$

(29)

$\gamma = 0, 1, \ldots, [n_D/2]$, for probability that the rare allele has $\gamma$ copies, and for (ii) probability that there are $b$ copies of the rare allele is:

$$P \left( X^R = b \right) = \pi_{n_J, b}^O$$

$$\pi_{n_J, b}^O = \pi_{n_J, b}^O + \pi_{n_J, n_J - b}^O \left[ 1 - \delta(h, n_J - b) \right]$$

(30)

$b = G, G + 1, \ldots, [n_J/2]$.

![Figure 2. Log likelihood curve for exponential model of population growth for SNP data from [13].](image)

![Figure 3. Log likelihood curve for exponential model of population growth for SNP data from the paper [20].](image)

When analyzing SNP data we followed remarks given in the source papers [13] and [20] to adjust parameters $n_A$, $n_D$, $n_O$, and $n_J$. The likelihood function for the exponential model of population growth can be used to estimate the parameters of the model from the SNP data.
n_o and G of the model of ascertainment procedure. We have modeled ascertainment procedure for collecting data from [13] by using expression (27) with n_A = 10 and G = 2. A plot of log likelihood function for the data on Caucasians from the paper [13] is shown in fig. 2. It attains its maximum at \( \hat{\kappa} = 3.9 \). Log likelihood function for the data on Caucasians from [20] is plotted in fig. 3. Ascertainment was modeled by expression (28) with n_o = 10 and G = 1. Maximum likelihood estimate of the product parameter, from the plot in fig. 3, is \( \hat{\kappa} = 0.78 \).

### 3.3 Sensitivity to ascertainment model parameters

A question arises: “How sensitive are the estimates of parameter \( \kappa \) to changes of the model of the ascertainment?”.

We studied this question by increasing or decreasing the value of the threshold \( G \) in expressions (27) and (28). Indexing estimated parameter with \( n_A \), \( n_O \) and \( G \), we can denote our estimates from previous section as

\[
\hat{\kappa}_{n_A=10, G=2} = 3.9, [13] \tag{31}
\]

and

\[
\hat{\kappa}_{n_O=10, G=1} = 0.78, [20]. \tag{32}
\]

Here we compute estimates \( \hat{\kappa}_{n_A=10, G=3} \), \( \hat{\kappa}_{n_A=10, G=0} \), \( \hat{\kappa}_{n_O=10, G=0} \) and \( \hat{\kappa}_{n_O=10, G=2} \) for data [20]. Analysis of data (Trikka et al. 2002) needs some more comments. The model to estimate \( \hat{\kappa}_{n_O=10, G=0} \) means that no ascertainment procedure is taken into account. The model \( \hat{\kappa}_{n_O=10, G=2} \) is inconsistent with [20] in the sense that the data contain one SNP locus with \( b = 1 \). To apply model \( \hat{\kappa}_{n_O=10, G=2} \) we have removed this one locus.

The results of computations show an extreme sensitivity of the ascertainment model. Namely

\[
\hat{\kappa}_{n_A=10, G=1} = 0, \hat{\kappa}_{n_A=10, G=3} = \infty, [13] \tag{33}
\]

and

\[
\hat{\kappa}_{n_O=10, G=0} = 0, \hat{\kappa}_{n_O=10, G=2} = 683, [20]. \tag{34}
\]

In (33), by \( \hat{\kappa}_{n_A=10, G=3} = \infty \) we meant that likelihood function was increasing for values of \( \kappa \) up to \( 10^8 \).

### 4 Discussion

The presented methods allow us to analyze large data sets and to carry out computations for different parameter values, which helps drawing more conclusions from data analysis.

We were interested in the problem: “What are reasonable values of the exponential growth product parameter \( \kappa = rN_q \) that can be obtained from DNA data?” and in comparing estimates obtained using different approaches. Estimates of populations histories from mtDNA were obtained e.g., in [18], [14], [24] and [19]. Data in the referenced papers come from different sources, but considering estimations of the authors, reasonable ranges of the product parameter \( \kappa \), for both world - wide population and Caucasians seem \( \kappa = 50 \div 500 \).

Mutation intensity (per site) at the autosomal loci is approximately one order of magnitude lower than that in mtDNA [10]. However, since the product parameter \( \kappa = rN_q \) is invariant with respect to time scale changes, the estimates of \( \kappa \) from SNP data should be also comparable to the above described ranges. Our estimates \( \hat{\kappa} = 3.9 \), and \( \hat{\kappa} = 0.78 \) are markedly smaller then values coming from mtDNA. Differences between our estimates and the above ranges can, in our opinion, be attributed to the sensitivity to the parameters of the ascertainment model, shown in (33), (34). With this extremely high sensitivity even a small unmodeled bias, e.g., resulting from eliminating of some low - frequency SNPs by assuming that they resulted from sequencing errors, can lead to values of \( \hat{\kappa} \) substantially lower than expected.

The fact that the ascertainment model strongly affects estimates of parameters is also confirmed in the previous papers on SNPs. In [21] in their fig. 3 one can see a large bias in SNP frequencies resulting from ascertainment. Similarly, in [12] a rather oversimplified model was used, \( n_O = 2, G = 1 \), for data from [13] and the obtained estimate was \( \hat{\kappa} = 0 \).

### 5 Acknowledgments

The authors were supported by NIH grants GM58545, CA75432, Polish Scientific Committee (KBN) research project 4 T 11F 018 24 and NATO collaborative linkage grant LST.CLG.977845.

### References


