DOI: 10.1534/genetics.104.031039

Reconstituting the Frequency Spectrum of Ascertained Single-Nucleotide Polymorphism Data

Rasmus Nielsen,*,†,[1] Melissa J. Hubisz* and Andrew G. Clark‡

**Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, ‡Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York and †Center for Bioinformatics, University of Copenhagen, 2100 Copenhagen, Denmark*

Manuscript received May 10, 2004
Accepted for publication September 10, 2004

ABSTRACT

Most of the available SNP data have eluded valid population genetic analysis because most population genetical methods do not correctly accommodate the special discovery process used to identify SNPs. Most of the available SNP data have allele frequency distributions that are biased by the ascertainment protocol. We here show how this problem can be corrected by obtaining maximum-likelihood estimates of the true allele frequency distribution. In simple cases, the ML estimate of the true allele frequency distribution can be obtained analytically, but in other cases computational methods based on numerical optimization or the EM algorithm must be used. We illustrate the new correction method by analyzing some previously published SNP data from the SNP Consortium. Appropriate treatment of SNP ascertainment is vital to our ability to make correct inferences from the data of the International HapMap Project.

THE large-scale single-nucleotide polymorphism (SNP) genotyping projects have generated much interest in population genetic analysis of human polymorphism. SNPs may be used for the estimation of demographic parameters, such as population growth rates, admixture proportions, migration rates, and population divergence times (*e.g.*, Wakeley *et al.* 2001; Cavalli-Sforza and Feldman 2003). In addition, SNPs may be used in studies of the effect of natural selection, for example, for mapping the genomic location of selective sweeps (*e.g.*, Sunyaev *et al.* 2000; Akey *et al.* 2002; Sabeti *et al.* 2002). With the availability of thousands of typed SNPs in multiple human ethnic groups, there is some hope that many questions regarding the human genetic ancestry might soon be resolved. However, the analysis of the SNP data is complicated by the SNP discovery protocols applied in the large SNP genotyping projects. Typically, SNPs are originally identified from the genetic material of a small group of individuals, often called the discovery panel. Thereafter, the SNPs found in this small panel are typed in a larger sample, typically with an ethnic composition similar to that of the discovery panel (*e.g.*, Taillon-Miller *et al.* 1998; Wang *et al.* 1998; Picoult-Newberg *et al.* 1999; Altshuler *et al.* 2000). Basing the SNP discovery protocol on initial identification in a small panel, in contrast to direct sequencing, will bias the composition of the sample to contain more high-frequency alleles (*e.g.*, Nielsen 2000). Most standard population genetic tools for data analysis are, therefore, not applicable to this type of SNP data. Fortunately, it is in many cases possible to correct for the ascertainment bias (*e.g.*, Wakeley *et al.* 2001; Nielsen and Signorovitch 2003; Polanski and Kimmel 2003). For example, Nielsen and Signorovitch (2003) showed how the Hudson (2001) composite-likelihood estimator of the population recombination rate can be modified to provide approximately unbiased estimates.

In this article we focus on methods for estimating the true frequency spectrum from a sample of SNP data. The frequency spectrum is a reduction of the data in which all SNPs are categorized according to the sample allele frequency of the SNP. Assuming no back mutations and assuming that the ancestral state of the SNP is known, there are $n - 1$ possible allele frequencies in a sample of n chromosomes: $x = 1$, $x = 2$, . . . , $x = n - 1$. If the ancestral state is not known, the labeling of alleles is arbitrary, and allele frequencies of type x are identical to allele frequencies of type $n - x$. Consequently, there are only $[n/2]$ possible *folded* configurations, where $[n/2]$ is $n/2$ truncated to the nearest integer. Under the assumption that SNPs are independent and identically distributed (iid), all the information in the data, for example, regarding demographic parameters, is contained in the frequency spectrum. The iid assumption is valid if the SNPs are located far apart and if the evolutionary processes are identical in all regions. If parameters of the evolutionary process vary among regions, the relevant information in the data is then instead contained in the collection of frequency spectra in different regions.

The objective of this article is to show how the true frequency spectrum can be estimated from ascertained

[1]*Corresponding author:* Center for Bioinformatics, Universitetsparken 15, 2100 Kbh Ø, Denmark. E-mail: rasmus@binf.ku.dk

SNP data. We focus on the frequency spectrum for three reasons. First, the methods used to correct the frequency spectrum are conceptually identical to the methods used to correct estimators of any other parameters. We derive formulas for correcting the frequency spectrum that can be applied more or less directly in studies aimed at estimating other parameters. Second, in some cases the frequency spectrum in itself is of interest, for example, for identifying genomic regions with aberrant frequency spectra, possibly due to selection. Third, by correcting the frequency spectrum for the ascertainment bias, while taking into account the inflation of the variance due to the estimation procedure, other parameters, such as demographic parameters, can be estimated.

We show that in simple, but realistic cases, an analytical formula can be used to provide maximum-likelihood estimates of the true frequency spectrum. In the more general cases, fast numerical optimization algorithms can be used to estimate the true frequency spectrum. We use these new methods to analyze a previously published SNP data set from The SNP Consortium (TSC; *e.g.*, Matise *et al.* 2003).

THEORY AND METHODS

We illustrate the methods discussed here on the unfolded frequency spectrum, but the results can trivially be extended to the folded frequency spectrum. Let p_i be the frequency of SNPs with mutant allele frequency i in a sample that has not been subject to any ascertainment bias. Given observed counts of SNP alleles in a sample, we seek the *reconstituted* frequency spectrum, defined here as the maximum-likelihood estimate of $\mathbf{P} = (p_1, p_2, \ldots, p_{n-1})$, where n is the sample size of chromosomes. We assume that some ascertainment condition has been imposed such that only loci fulfilling this condition have been included in the final data set. For example, an ascertainment condition could be that all included SNPs are variable, or that all SNPs were variable in some panel originally used to screen for SNPs. The likelihood function for $\mathbf{P}$ is then given by

$$L(\mathbf{P}) \propto \prod_{i=1}^{S} \Pr(X_i = x_i|\mathbf{P}; \mathrm{Asc}_i) = \prod_{i=1}^{S} \frac{\Pr(X_i = x_i, \mathrm{Asc}_i|\mathbf{P})}{\Pr(\mathrm{Asc}_i|\mathbf{P})}, \quad (1)$$

where S is the number of SNP loci in the sample, X_i is the allele frequency in locus i, and Asc_i is generic notation for the event that the ascertainment condition is met in locus i. Note that we have here assumed independence among loci. In the following we show how this likelihood function can be maximized with respect to $\mathbf{P}$ for a number of different ascertainment (SNP discovery) protocols. Methods for correcting likelihood estimators of demographic parameters have previously been discussed by Kuhner *et al.* (2000), Nielsen (2000), Wakeley *et al.* (2001), Akey *et al.* (2003), Nielsen and Signorovitch (2003), and Polanski and Kimmel (2003).

Case 1—basic model: Let us first consider the case in which all SNPs have been ascertained in an alignment of different sequences of fixed depth (d) and where this ascertainment sample is a subset of the final sample of size n. The depth is the sample size of the ascertainment sample. The ascertainment condition is that the *locus was variable in the ascertainment sample.* Then the probability of ascertainment given an observed allele frequency of x_i is one minus the probability of sampling all d ascertainment gene copies exclusively among either the x_i alleles of one type or the $n - x_i$ alleles of the other type. Also $\Pr(X_i = x_i|\mathbf{P}) = p_{x_i}$ and $\Pr(\mathrm{Asc}_i|X_i = x_i) = \Pr(\mathrm{Asc}_i|X_i = x_i, \mathbf{P})$, so

$$\Pr(X_i = x_i, \mathrm{Asc}_i|\mathbf{P}) = p_{x_i} \Pr(\mathrm{Asc}_i|X_i = x_i),$$

where

$$\Pr(\mathrm{Asc}_i|X_i = x_i) = 1 - \frac{\binom{x_i}{d} + \binom{n-x_i}{d}}{\binom{n}{d}} \quad (2)$$

and

$$\Pr(\mathrm{Asc}_i|\mathbf{P}) = \sum_{j=1}^{n-1} p_j \Pr(\mathrm{Asc}_i|X_i = j).$$

Here and in the following $\binom{i}{j} = 0$ if $j > i$ or $j < 0$. We find the maximum-likelihood estimate by solving a set of equations obtained by setting the partial derivatives of the log-likelihood function with respect to the parameters equal to zero and solving for the parameters. Because of the constraint of $\Sigma_{i=1}^{n} p_i = 1$ we introduce a Lagrange multiplier. After verifying that a global maximum has been found, we find that the maximum-likelihood estimate of $\mathbf{P}$ is simply given by

$$\hat{p}_k = \frac{n_k}{\Pr(\mathrm{Asc}|X = k)} \left[\sum_{j=1}^{n-1} \frac{n_j}{\Pr(\mathrm{Asc}|X = j)} \right]^{-1}, \quad k = 1, 2, \ldots, n - 1, \quad (3)$$

where n_j is the observed number of loci with allele frequency j. An example is given in Figure 1. Ten thousand independent SNP sites were simulated under an infinite-sites model in a sample of size $n = 20$. Note that when $d = 5$, the true and the observed frequency spectra differ dramatically. In particular, in the observed data there is an excess of loci with intermediate frequency alleles. Also note how the maximum-likelihood correction method accurately recovers the true frequency spectrum.

In some cases, the SNP selection criterion used in the SNP discovery process might include a cutoff in the SNP frequency. Such cases can easily be incorporated into the current scheme. If the cutoff is set at a value C, then the likelihood function is just modified using

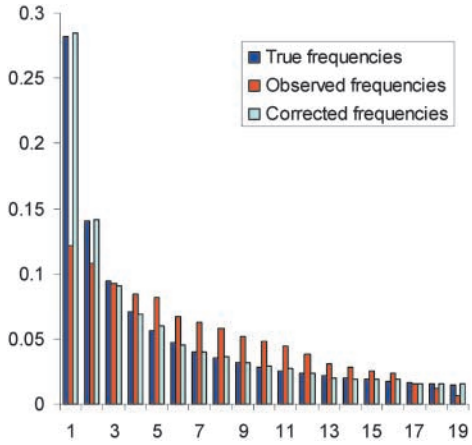

FIGURE 1.—The frequency spectrum in a sample of $n = 20$ gene copies and 10,000 SNPs assuming $d = 5$ and the ascertainment scheme in Equation 1. The data were simulated assuming the standard neutral coalescent model and independence among sites.

$$\Pr(\text{Asc}_i | X_i = x_i) = \frac{\sum_{j=C}^{n-C} \binom{x_i}{j}\binom{n-x_i}{d-j}}{\binom{n}{d}}, \qquad (4)$$

and the maximum-likelihood estimate of **P** can be obtained by substituting this expression into Equation 3.

Case 2—variation in *d*: This case is similar to case 1, but we assume the discovery depth (d) varies among loci. Consider first the case where information regarding d in each locus has been lost, but information is available regarding the distribution of d among loci, $f(d)$. Then the likelihood function must be modified by summing over all possible (unknown) alignment depths when calculating the ascertainment probability,

$$\Pr(\text{Asc}_i | X_i = x_i) = \sum_{d=2}^{d_{\max}} f(d)\left(1 - \frac{\left[\binom{x_i}{d} + \binom{n-x_i}{d}\right]}{\binom{n}{d}}\right), \qquad (5)$$

where $d_{\max}$ is the maximum value d can take. The maximum-likelihood estimate of **P** is then given by Equation 3, replacing the definition of $\Pr(\text{Asc}_i | X_i = x_i)$ by Equation 5.

In the case where information regarding d is available for each locus, but d varies among loci, the likelihood function is given by Equation 2, replacing d with d_i, where d_i is the value of d in SNP locus i. Numerical optimization of this likelihood function (Equation 2) is necessary, but can be done very fast and efficiently using standard algorithms.

An example is shown in Figure 2. Ten thousand independent SNPs were simulated assuming $n = 20$ and a mixture of ascertainment sample sizes of $d = 2$, 3, 5, and 10 with equal probability. Again, the simulated data have an excess of loci with alleles of intermediate frequency compared to the true distribution. Three different correction schemes are considered. First, the likelihood function based on Equation 2 is used to correct

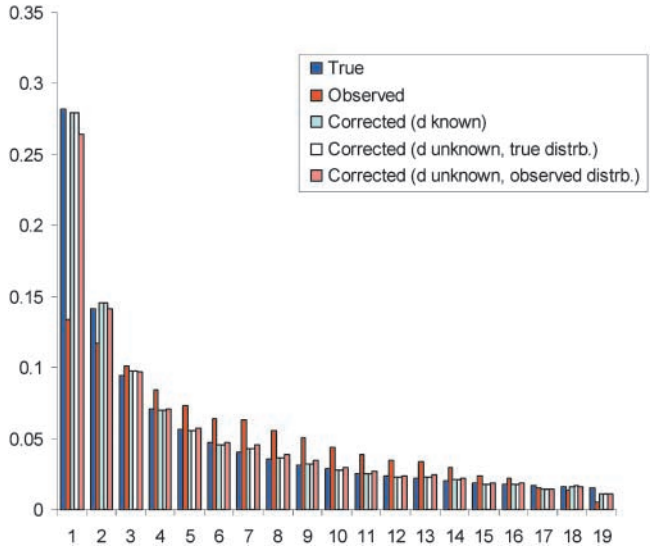

FIGURE 2.—The true and estimated frequency spectra assuming $n = 20$ and $d = 2$, 3, 5, and 10 with equal probability for 10,000 SNPs. Three different corrections schemes are assumed: d known for each locus (light blue), d unknown for each locus but the true distribution of d known (white), and d unknown for each locus but the distribution of d inferred from the distribution in the typed SNP data (pink).

the distribution, assuming d_i is known for all loci. This procedure accurately recovers the true frequency spectrum. Two different correction schemes based on Equation 5 are then considered. In the first case it is assumed that the true distribution of ascertainment sample sizes is known. Using this distribution, the correct frequency spectrum is again recovered. In the second procedure, the observed distribution of ascertainment sample sizes is used in combination with Equation 5. The observed distribution is obtained by simply counting the number of typed SNPs for which $d = 2, 3, \ldots$, etc. Using this procedure leads to a small bias and a deficiency of rare alleles. The reason is that the observed distribution of ascertainment sample sizes is in itself biased, because samples in which no SNPs occurred have been eliminated.

So far we have assumed that the ascertainment sample consists of an alignment of different sequences. However, in more realistic cases the ascertainment sample has been obtained by sampling with replacement from a panel of chromosomes of size m. For example, National Human Genome Research Institute sponsored a SNP discovery effort in which a SNP discovery panel of $m = 24$ individuals was used by many groups to find SNPs. In the reduced representation shotgun scheme (ALTSHULER *et al.* 2000), multiple overlapping sequences were aligned for SNP discovery. In these overlaps, not all 24 individuals were represented, and some individuals were represented by more than one read. In this case, we need to distinguish between the observed depth of the alignment (A_i) in locus i and the true number of differ-

ent sequences in the alignment in locus i (d_i). Consider, for example, the case in which the alignment depth was known for each locus. Then

$$\Pr(\text{Asc}_i|X_i = x_i, A_i = a) = \sum_{d=1}^{m} \Pr(\text{Asc}_i|X_i = x_i, d_i = d)\Pr(d_i = d|A_i = a),$$

and

$$\Pr(d_i = d|A_i = a) = S(a, d)\, d! \binom{m}{d} m^{-a}, \qquad (6)$$

where $S(a, d)$ is a Stirling number of the second kind. The combinatorial expression in Equation 6 gives the probability of sampling exactly d different chromosomes when sampling a chromosomes, with replacement, from a panel of m chromosomes. m^a is the number of possible (ordered) ways we can sample a chromosomes with replacement from a panel of m chromosomes. $S(a, d)\, d! \binom{m}{d}$ is the number of ways we can sample, with replacement, a chromosomes among a panel of m chromosomes such that there are exactly d different chromosomes in the sample. There are $\binom{m}{d}$ different sets of chromosomes of size d to sample and $S(a, d)$ ways to partition the a draws into d nonempty sets, which again can be ordered in $d!$ different ways.

How important is it to model sampling with replacement, in contrast to assuming sampling without replacement ($d = a$) as previously assumed? In general, the effect is not large. For example, most of the available data from TSC (http://snp.cshl.org/) have values of $a < 5$, but $m = 20$ (see DATA ANALYSIS). In such cases correction with or without replacement gives almost identical results, because $E(d_i|A_i = a)$ is close to a; *e.g.*, if $m = 20$, then $E(d_i|A_i = 4) = 3.71$. We also explore cases where d is not much smaller than m and for the purpose of illustration show the case of $a = 5$ and $m = 7$ in Figure 3. In this case, $\Pr(d_i = 5|A_i = 5) \approx 0.15$, and we would expect relatively large differences between sampling with and without replacement. However, the difference between correcting with and without replacement is very minor compared to the effect of not correcting for the ascertainment bias. Corrections performed without taking into account the possibility that the same sequence has been sampled more than once from the panel sequences may perform reasonably well as long as $a < m$. Most of the TSC data can probably be modeled reasonably well without taking sampling with replacement into account.

Case 3—allele frequencies in the ascertainment sample unknown: In many cases the ascertainment sample may not have been included in the final typed sample. In this case we redefine the ascertainment condition as *variability in the ascertainment sample and variability in the typed sample*, since invariable loci in the typed sample in most cases will be discarded (Figure 4). If the information regarding the allele frequency in the ascertainment sample has been preserved, the previous methods can easily be adapted to deal with this case. However, if the information regarding allele frequencies in the ascertainment sample is not available, this introduces quite a bit more complexity. In the following, we illustrate how case 1 can be expanded to include this type of ascertainment scheme. The basic idea is to calculate the likelihood function by summing over all the possible values of the allele frequency in the unobserved ascertainment sample. First, redefine $\Pr(X_i = x_i, \text{Asc}_i|\mathbf{P})$ in an alignment of depth d as

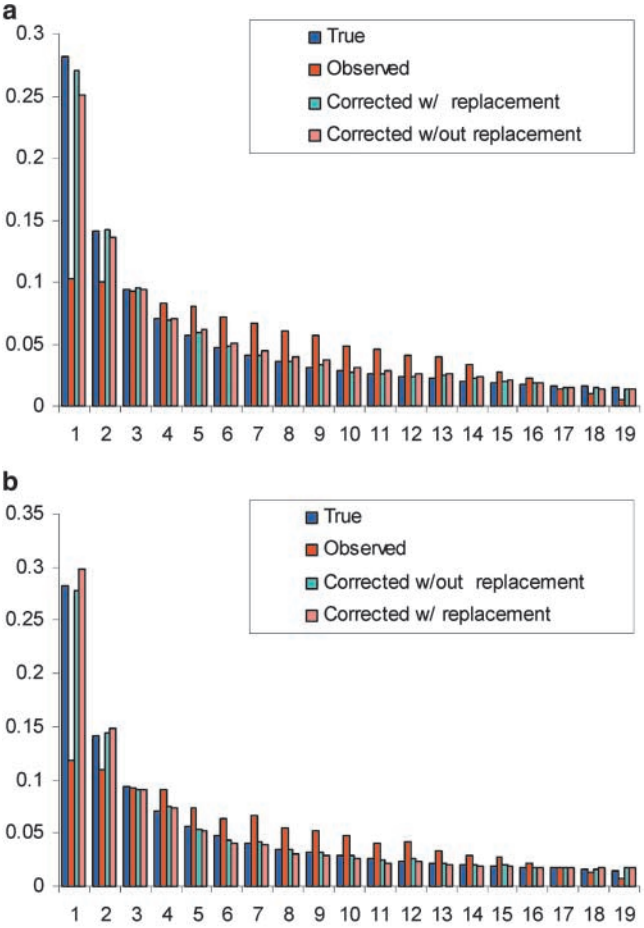

FIGURE 3.—The true and estimated frequency spectra assuming $n = 20$ and $a = 5$ and $m = 7$ for 10,000 SNPs, in a SNP discovery process where ascertainment sequences have been sampled with replacement from the panel sequences (a) and without replacement (b).

$$\Pr(X_i = x_i, \text{Asc}_i|\mathbf{P}) = \sum_{j=x_i+1}^{x_i+d-1} p_j \Pr(X_i = x_i|Y_i + X_i = j), \qquad (7)$$

where Y_i is the unknown allele frequency in the ascertainment sample of size d and $\mathbf{P} = (p_2, p_3, \ldots, p_{n+d-2})$. Also,

$$\Pr(X_i = x_i|Y_i + X_i = j) = \frac{\binom{j}{x_i}\binom{n+d-j}{n-x_i}}{\binom{n+d}{n}}. \qquad (8)$$

Similarly, redefine

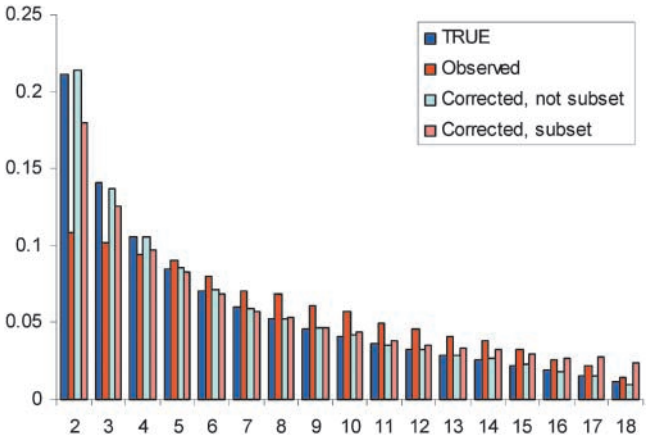

FIGURE 4.—The true and estimated frequency spectra assuming $n = 20$ and $d = 5$ for 10,000 SNPs in which the ascertainment sample is disjoint from (not a subset of) the typed sample. Two different corrections are performed, one in which it is correctly assumed that the ascertainment sample is disjoint from the typed sample (light blue), estimated using the EM algorithm described in the APPENDIX, and one in which it was incorrectly assumed that the ascertainment sample was a subset of the typed sample. For fair comparison, all distributions are calculated conditionally on $1 < x < 19$.

$$\Pr(\text{Asc}_i|\mathbf{P}) = \sum_{j=2}^{x_i+d-1} p_j \Pr(\text{Asc}_i|X_i + Y_i = j), \quad (9)$$

where

$$\Pr(\text{Asc}_i|X_i + Y_i = j) = 1 - \frac{\binom{j}{n} + \binom{n+d-j}{n} + \binom{j}{d} + \binom{n+d-j}{d} - I_{d=j} - I_{n=j}}{\binom{n+d}{n}}, \quad (10)$$

if $1 < j < n + d - 1$ and 0 otherwise. Note that the model is now parameterized in terms of the allele frequencies in a pooled sample of size $n + d$. Because both the ascertainment sample and the typed sample are required to be variable for a locus to be ascertained, only $p_2, p_3, \ldots, p_{n+d-2}$ can be estimated. The frequency of singletons in the sample cannot be consistently estimated without making more model assumptions, because the pooled sample contains no singletons.

The likelihood function is now of an algebraic form where the maximum likelihood cannot easily be obtained analytically. Instead, we can develop a fast EM algorithm for maximizing the likelihood function. When d varies among loci, the EM algorithm is no longer easily applicable and other numerical optimization methods must be used. The APPENDIX describes the EM algorithm and the necessary alterations of the likelihood function when d varies among loci.

Case 4—the "double-hit" ascertainment scheme: The International HapMap project is the largest SNP genotyping project ever conceived—currently planned to include a minimum of 600,000 SNPs genotyped in 270 individuals. Prior to this genotyping, SNPs are selected for the study on the basis of prior knowledge that they are variable sites and of their position in the genome. Recently, the criterion that has been selected for ascertainment is the "double-hit" scheme, meaning that both allelic states were observed in two separate studies (www.hapmap.org).

Assume that we know the panel depth for both ascertainment experiments and that the discovery panel is part of the sample used to obtain the frequency spectrum (as in case 1). Further assume that the two discovery samples were drawn from the same population. Let Asc1_i refer to a SNP satisfying ascertainment condition 1 [*i.e.*, it was discovered in an alignment of sequences of depth $d^{(1)}$] and Asc2_i implies that the SNP was discovered in another alignment of sequences of depth $d^{(2)}$. Similarly to case 1, assume that the ascertainment samples are subsets of the typed sample and further assume that the intersection between these two subsets is empty. Then,

$$\Pr(\text{Asc1}_i, \text{Asc2}_i|X_i = x_i) = \frac{\Sigma_\Omega \binom{x_i}{x_i-j-k,j,k}\binom{n-x_i}{n-x_i-d^{(1)}-d^{(2)}+j+k,\ d^{(1)}-j,\ d^{(2)}-k}}{\binom{n}{n-d^{(1)}-d^{(2)},d^{(1)},d^{(2)}}}, \quad (11)$$

where X_i is the frequency of the mutant allele in the ith locus of a sample of size $n + d^{(1)} + d^{(2)}$ and $\Omega = \{(j, k)|0 < j < d^{(1)}, 0 < k < d^{(2)}\}$. The maximum-likelihood estimate of $\mathbf{P} = (p_2, p_3, \ldots, p_{n-2})$ is then simply given by Equation 3 using (Asc1, Asc2) as the ascertainment condition. Unknown allele frequencies in the ascertainment sample and varying ascertainment sample size can also be incorporated in this ascertainment scheme.

An example of this ascertainment scheme is shown in Figure 5. Note the magnitude of the ascertainment bias under this selection scheme. In a sample of size $n = 20$, SNPs with allele frequencies in the range of 4/20–10/20 are now the most common SNPs. Again, the maximum-likelihood correction accurately recovers the true allele frequencies; however, incorrectly assuming a single-hit correction and $d = 2$ does not fully recover the true frequency spectrum. Clearly, under the double-hit ascertainment scheme ascertainment corrections based on a single-hit scheme are not appropriate.

Hypothesis testing and confidence intervals: The previous discussion has illustrated how the frequency spectrum can be corrected for a variety of different ascertainment schemes. However, it has not addressed the fundamental problem of how to apply estimates of the frequency spectrum for further population genetic analysis. It is important to stress that ascertainment-corrected frequency spectra cannot be directly applied in further data analysis without taking the uncertainty in the parameter estimates into account. Fortunately, it is relatively easy to obtain measures of statistical uncertainty in these models. For example, consider ascertainment schemes where the likelihood function has the same functional form as in case 1. Then the approximate

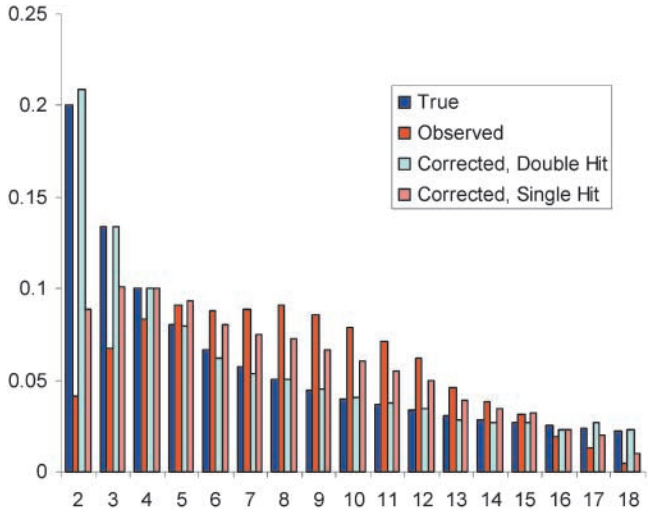

FIGURE 5.—The true and estimated frequency spectra assuming $n = 20$, $d^{(1)} = 2$, and $d^{(2)} = 5$ for 10,000 SNPs for the double-hit ascertainment scheme. Two different corrections are performed, one in which it is correctly assumed that the double-hit ascertainment scheme has been used (light blue) and one in which it was erroneously assumed that the data were obtained using a single-hit ascertainment scheme with $d = 2$.

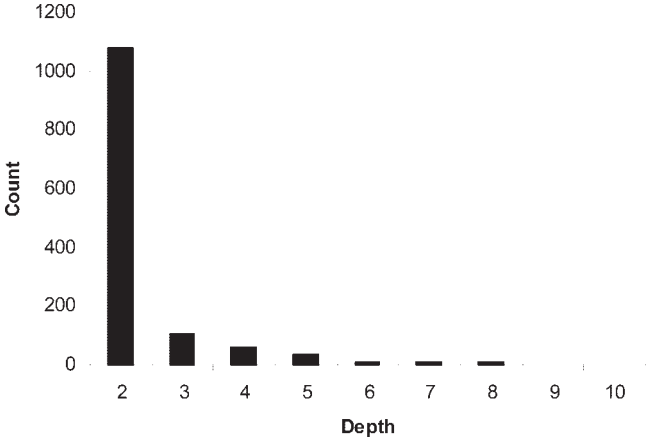

FIGURE 6.—The observed distribution of alignment depths for 1308 SNPs from The SNP Consortium.

variances of the estimates can be obtained using asymptotic likelihood theory. The observed Fisher information matrix $\mathbf{I_P} = \{I_{ij}\}$ for $0 < i < n - 1$ is given by the negative of the matrix of second derivatives of the log-likelihood function,

$$I_{ij} = \begin{cases} \frac{n_{n-1}}{\hat{p}_{n-1}^2} - \frac{S(\Pr(\text{Asc}|X = i) - \Pr(\text{Asc}|X = n-1))(\Pr(\text{Asc}|X = j) - \Pr(\text{Asc}|X = n-1))}{(\sum_{v=1}^{n-1} \hat{p}_v \Pr(\text{Asc}|X = v))^2}, \\ \qquad i \neq j \\ \frac{n_{n-1}}{\hat{p}_{n-1}^2} + \frac{n_i}{\hat{p}_i^2} - \frac{S(\Pr(\text{Asc}|X = i) - \Pr(\text{Asc}|X = n-1))^2}{(\sum_{v=1}^{n-1} \hat{p}_v \Pr(\text{Asc}|X = v))^2}, \\ \qquad i = j \end{cases} \tag{12}$$

and the approximate variance-covariance matrix can be found as $\mathbf{I}_\mathbf{P}^{-1}$, which can be obtained analytically, but is messy. For models in which the ascertainment scheme varies among loci, the observed Fisher information matrix is given by

$$I_{ij} = \begin{cases} \frac{n_{n-1}}{\hat{p}_{n-1}^2} - \sum_{k=1}^{S} \frac{(\Pr(\text{Asc}_k|X_k = i) - \Pr(\text{Asc}_k|X_k = n-1))(\Pr(\text{Asc}_k|X_k = j) - \Pr(\text{Asc}_k|X_k = n-1))}{(\sum_{v=1}^{n-1} \hat{p}_v \Pr(\text{Asc}_k|X_k = v))^2}, \\ \qquad i \neq j \\ \frac{n_{n-1}}{\hat{p}_{n-1}^2} + \frac{n_i}{\hat{p}_i^2} - \sum_{k=1}^{S} \frac{(\Pr(\text{Asc}_k|X_k = i) - \Pr(\text{Asc}_k|X_k = n-1))^2}{(\sum_{v=1}^{n-1} \hat{p}_v \Pr(\text{Asc}_k|X_k = v))^2}. \\ \qquad i = j \end{cases} \tag{13}$$

After obtaining the variance-covariance matrix, confidence intervals for the parameters can be obtained using standard methods. Likewise, approximate confidence intervals for any function that has been calculated on the basis of the frequency spectrum (*e.g.*, an estimator of growth rates or other demographic parameters) can be obtained, if this function is differentiable. The approximate variance of the function is obtained by applying the delta method (see, *e.g.*, CASELLA and BERGER 1990, p. 326).

An alternative would be to bootstrap the SNPs, and for each bootstrap sample estimate the reconstituted frequency spectrum. By taking into account the increased variance due to the estimation of the frequency spectrum, such a bootstrap would accurately represent the true variance in the estimates. However, it should be noted that when numerical optimization is necessary, such a bootstrap approach can be quite computationally intensive.

Hypothesis testing can be performed using similar methods. For example, we might be interested in testing if the frequency spectrum conforms to a specific model such as the standard neutral equilibrium model, *i.e.*, KINGMAN's (1982) coalescent model. In this model

$$p_i = \frac{1/i}{\sum_{j=1}^{n-1} 1/j}, \qquad 0 < i < n. \tag{14}$$

We may now calculate a likelihood-ratio test statistic as $\text{Log}(L(\hat{\mathbf{P}})/L(\mathbf{P}_c))$, where $\mathbf{P}_c$ is the value of $\mathbf{P}$ under Kingman's coalescent. Two times this statistic is asymptotically χ^2_{n-2} distributed. However, for most data sets, the observations in the categories of the high-frequency-derived alleles will be so low that the asymptotic result may not apply. In such cases, the distribution of the test statistic must be evaluated by simulations.

DATA ANALYSIS

To illustrate the utility of the method we analyzed 1308 SNPs from The SNP Consortium (*e.g.*, MATISE *et al.* 2003; THORISSON and STEIN 2003), for which chimpanzee outgroup information was available (allowing consideration of the full rather than the folded frequency spectrum). The typed sample consisted of 90 individuals, and the discovery (ascertainment) panel consists of 24 individuals, except for some cases in which $m = d = 2$. We assume that the sets of typed and ascertainment individuals are disjoint. The distribution of alignment depths for the 1308 SNPs is shown in Figure 6.

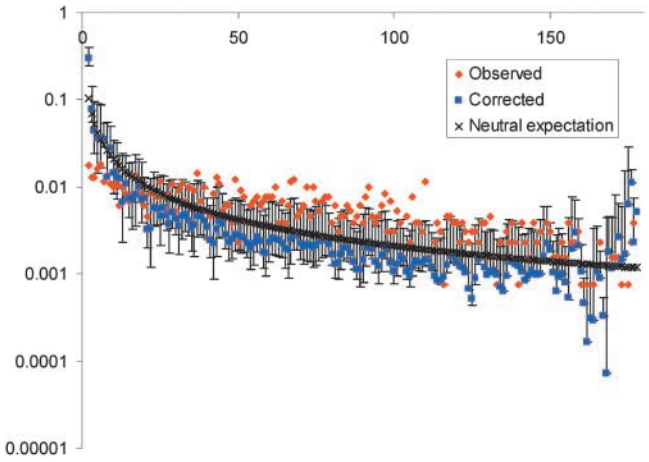

FIGURE 7.—The observed, expected, and estimated frequency spectra for a data set from The SNP Consortium containing 90 individuals for 1308 SNPs. Note the logarithmic scale on the *y*-axis. Error bars show plus or minus two times the standard deviation estimated from the observed Fisher information.

The vast majority of SNPs are obtained from alignment depths of only two sequences and only 32 SNPs have alignment depths >5. No SNPs have alignment depths >10. Because the alignment depths in general are much smaller than the panel size, except for the case of $m = d = 2$, we model the ascertainment process using Equations 6–9, ignoring the possibility that the same sequence occurs twice in an alignment. However, we do take variation in d among loci into account.

The estimated and observed frequency spectra for these data are shown in Figure 7. Approximate 95% confidence intervals were obtained as plus or minus two times the standard deviation. Standard deviations were approximated as the square root of the asymptotic variances obtained using Equation 13. The frequency of singletons cannot be estimated consistently using this approach, because of the assumption of variability in both the ascertainment and the typed sample. Note that the observed distribution is quite uniform compared to the distribution expected under neutrality. There is a deficiency of rare new mutants and an excess of common alleles. In contrast, the corrected frequency spectrum shows an excess of rare alleles, as is observed in much of the available human data obtained by direct sequencing (STEPHENS *et al.* 2001). The excess of rare alleles may most likely be caused by population growth and/or by selection against slightly deleterious mutations.

We tested the fit of Kingman's coalescent model to this data using the previously described likelihood-ratio test. The observed value of the test statistic was 100.3. To evaluate the distribution of this test statistic, 100 data sets of 1308 independent SNPs were simulated under Kingman's coalescent (*i.e.*, from Equation 14), while imposing the same ascertainment conditions as observed in the real data. The simulated distribution is shown in Figure 8. In this case the data do not fit Kingman's coalescent, due to an excess of rare derived alleles.

DISCUSSION

In this article we present a set of methods for correcting the frequency spectrum in ascertained SNP data. The methods are nonparametric in the sense that they make no assumptions regarding the processes that generate the data. There is no need for a prior distribution of allele frequencies or a population genetical model. Inferences regarding population level processes can be based on the reconstituted frequency spectrum. When doing so, methods for taking uncertainty in the estimate of the frequency into account should be used. This can be done either by using the bootstrap or by using asymptotic likelihood theory.

Given the simplicity of implementation of these methods, and the growing prevalence of SNPs ascertained

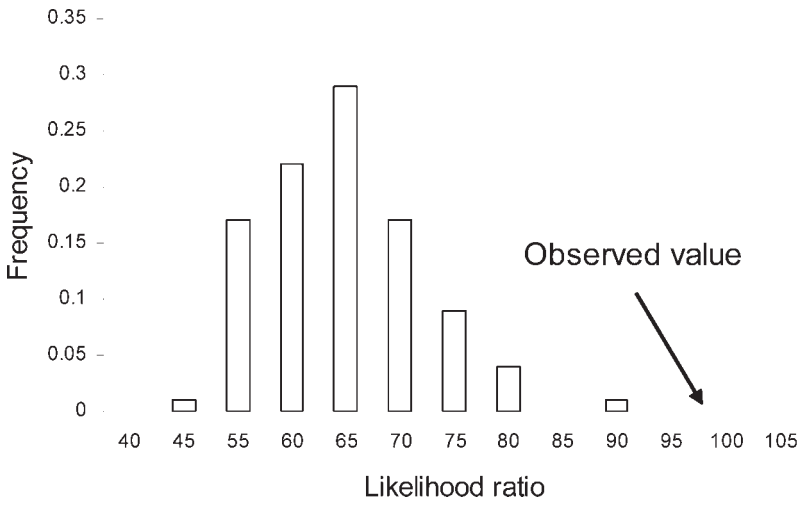

FIGURE 8.—The distribution of the likelihood-ratio test statistic of fit to the standard neutral model. The observed value is calculated from The SNP Consortium data containing 90 individuals for 1308 SNPs.

through a small panel, we emphasize the importance of considering and correcting ascertainment in the analysis of human SNP data. Many of the primary inferences to be drawn from SNP data about demographic history, such as allele age, rely on an accurate assessment of the frequency spectrum. Some methods of inference of association between SNPs and risk of complex diseases also rely on inference of allele frequency spectrum, and for this application we need the most accurate statistical procedures available.

The methods described in this article assume that all SNPs are independent. While this may be true for some data sets, many data sets will contain SNPs that are correlated due to linkage. We expect that the estimate of the frequency spectrum is approximately unbiased also in such cases. For example, even in the case of linkage, several of the maximum-likelihood estimators, such as the estimator derived in case 1, can be shown also to be method-of-moments estimators. However, the measures of statistical uncertainty obtained using asymptotic likelihood theory or the bootstrap are no longer valid in the presence of linkage.

Throughout this article we have also assumed that all sequences are exchangeable, *i.e.*, that there is no population subdivision. This is a rather strong assumption given the moderate amount of population structure observed in most human SNP data. If the ascertainment sample and the typed sample have the same ethnic makeup the effects on the estimates will probably be minor. However, the methods discussed here should not be applied to data for which the ethnic makeup is radically different between ascertainment sample and typed sample. In such cases ascertainment correction methods that explicitly take population subdivision into account should be applied.

We thank the associate editor and two anonymous reviewers for their useful comments. This work was supported by Human Frontier in Science Program grant RGY0055/2001-M, National Institutes of Health (NIH) grant HG03229, and National Science Foundation/NIH grant 0201037.

LITERATURE CITED

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12:** 1805–1814.

Akey, J. M., K. Zhang, M. Xiong and L. Jin, 2003 The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. Mol. Biol. Evol. **20:** 232–242.

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:** 513–516.

Casella, G., and R. L. Berger, 1990 *Statistical Inference*. Duxbury Press, Belmont, CA.

Cavalli-Sforza, L. L., and M. W. Feldman, 2003 The application of molecular genetic approaches to the study of human evolution. Nat. Genet. **33:** 266–275.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

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.

Matise, T. C., R. Sachidanandam, A. G. Clark, L. Kruglyak, E. Wijsman *et al.*, 2003 A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am. J. Hum. Genet. **73:** 271–284.

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

Nielsen, R., and J. Signorovitch, 2003 Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theor. Popul. Biol. **63:** 245–255.

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., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics **165:** 427–436.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1992 *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

Sunyaev, S. R., W. C. Lathe, III, V. E. Ramensky and P. Bork, 2000 SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. Trends Genet. **16:** 335–337.

Taillon-Miller, P., Z. Gu, Q. Li, L. Hillier and P. Y. Kwok, 1998 Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Res. **8:** 748–754.

Thorisson, G. A., and L. D. Stein, 2003 The SNP Consortium website: past, present and future. Nucleic Acids Res. **31:** 124–127.

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.

Communicating editor: L. Excoffier

APPENDIX

The basic idea in the EM algorithm is to augment the data with some additional hypothetical unobserved data. In the E-step of the EM algorithm the expected value of this augmented log-likelihood function is calculated, conditional on a current guess of the parameter values. In the M-step, this expectation is maximized with respect to the parameters, leading to a new set of parameter values. This algorithm will under general conditions converge to a (local) optimum when the two steps of the algorithm are iterated. We augment the data with the unknown allele frequencies in the ascertainment sample, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_S)$. The full-data likelihood becomes

$$L(\mathbf{P};\mathbf{Y}) = \prod_{i=1}^{S}\Pr(X_i = x_i,\ Y_i = y_i|\mathbf{P};\ \mathrm{Asc}_i) = \prod_{i=1}^{S}\frac{\Pr(X_i = x_i,\ Y_i = y_i,\ \mathrm{Asc}_i|\mathbf{P})}{\Pr(\mathrm{Asc}_i|\mathbf{P})}, \tag{A1}$$

where

$$\Pr(X_i = x_i,\ Y_i = y_i,\ \mathrm{Asc}_i|\mathbf{P}) = I_{(0<y_i<d)}\,I_{(0<x_i<n)}\,p_{x_i+y_i}\frac{\binom{y_i + x_i}{x_i}\binom{n + d - (x_i + y_i)}{n - x_i}}{\binom{n + d}{n}}. \tag{A2}$$

At the rth iteration of the algorithm the E-step consists of finding

$$E\{\log L(\mathbf{P};\mathbf{Z},\ Y)|\mathbf{X},\mathbf{P}^r\} = \sum_{i=1}^{S}\sum_{j=x_i+1}^{m} E[I_{(X_i=x_i,Y_i=j-x_i)}|\mathbf{X},\mathbf{P}^r;\ \mathrm{Asc}_i]\log\Pr(X_i = x_i,\ Y_i = j - x_i|\mathbf{P};\ \mathrm{Asc}_i), \tag{A3}$$

where

$$E[I_{(X_i=x_i,Y_i=j-x_i)}|\mathbf{X},\mathbf{P}^r;\ \mathrm{Asc}_i] = \frac{\Pr(X_i = x_i,\ Y_i = y_i,\ \mathrm{Asc}_i|\mathbf{P}^r)}{\sum_{y=1}^{d-1}\Pr(X_i = x_i,\ Y_i = y_i,\ \mathrm{Asc}_i|\mathbf{P}^r)}. \tag{A4}$$

The M-step of the algorithm can be completed by noting that Equation A3 can be optimized with respect to **P** using Equation 3. The algorithm then proceeds as follows:

1. Set $r = 0$ and $p_k^0 = (n + d - 3)^{-1}$, $k = 2, 3, \ldots, n + d - 2$.
2. Set $\hat{p}_{ki} = E[I_{(X_i+Y_i=j)}|\mathbf{X},\mathbf{P}^r;\ \mathrm{Asc}_i]$, $k = 2, 3, \ldots, n + d - 2$.
3. Set

$$p_k^{r+1} = \sum_{i=1}^{S}\frac{\hat{p}_{ki}}{\Pr(\mathrm{Asc}_i|X_i + Y_i = k)}\left[\sum_{i=1}^{S}\sum_{j=2}^{n+d-2}\frac{\hat{p}_{ji}}{\Pr(\mathrm{Asc}_i|X_i + Y_i = j)}\right]^{-1}, \qquad k = 2, 3, \ldots, n + d - 2.$$

4. Repeat steps 2 and 3 until convergence.

After convergence at the rth step of the algorithm, the reconstituted frequency spectrum in a sample of size $n + d$ is then given by p_j^{r+1}, $j = 2, \ldots, n + d - 2$. The reconstituted frequency spectrum in a sample of size n is then given by

$$\hat{p}_i = \frac{h_i}{1 - h_o - h_1 - h_n - h_{n-1}}, \qquad i = 2, \ldots, n - 2, \tag{A5}$$

where

$$h_i = \sum_{j=\max\{i,2\}}^{n+d-2} p_j^{r+1}\frac{\binom{j}{i}\binom{n + d - j}{n - i}}{\binom{n + d}{n}}. \tag{A6}$$

When the ascertainment sample is not contained in the observed sample and d varies among loci similarly to case 2, the EM-algorithm can no longer be applied, but standard numerical optimization algorithms must be used instead. However, this is the case relevant to data analysis of much of the available SNP data such as the data from TSC. First, redefine $\Pr(X_i = x_i, \mathrm{Asc}_i|\mathbf{P})$ in an alignment of depth d_i as

$$\Pr(X_i = x_i, \mathrm{Asc}_i|\mathbf{P}) = \sum_{j=x_i+1}^{x_i+d_i-1} p_j\Pr(\mathrm{Asc}_i,\ X_i = x_i|Y_i + X_i + Z_i = j), \tag{A7}$$

where Y_i is the unknown allele frequency in the ascertainment sample, Z_i is the allele frequency in a hypothetical sample of size $nd_{\max} - n - d_i$, $nd_{\max} = n + \max_j\{d_j\}$, $m = \max_j\{x_j + d_j\} - 2$, and $\mathbf{P} = (p_2, p_3, \ldots, p_m)$. Also

$$\Pr(\mathrm{Asc}_i,\ X_i = x_i|Y_i + X_i + Z_i = j) = \Pr(X_i = x_i|Y_i + X_i + Z_i = j) \times \Pr(\mathrm{Asc}_i|Y_i + Z_i = j - x_i), \tag{A8}$$

$$\Pr(X_i = x_i|Y_i + X_i + Z_i = j) = \frac{\binom{j}{x_i}\binom{nd_{\max} - j}{n - x_i}}{\binom{nd_{\max}}{n}}, \tag{A9}$$

and

$$\Pr(\mathrm{Asc}_i|Y_i + Z_i = j - x_i) = 1 - \frac{\binom{nd_{\max} - n - j + x_i}{d_i} + \binom{j - x_i}{d_i}}{\binom{nd_{\max} - n}{d_i}}. \tag{A10}$$

Similarly, redefine

$$\Pr(\mathrm{Asc}_i|\mathbf{P}) = \sum_{j=2}^{d_i+n-2} p_j \Pr(\mathrm{Asc}_i|X_i + Y_i + Z_i = j), \tag{A11}$$

where

$$\Pr(\mathrm{Asc}_i|X_i + Y_i + Z_i = j) = 1 - \frac{\binom{j}{d_i} + \binom{nd_{\max} - j}{d_i}}{\binom{nd_{\max}}{d_i}} - \frac{\binom{j}{n} + \binom{nd_{\max} - j}{n}}{\binom{nd_{\max}}{n}} + \frac{\binom{nd_{\max} - j}{n,\, d_i,\, nd_{\max} - j - n - d_i} + \binom{j}{0,\, d_i,\, j - d_i}\binom{nd_{\max} - j}{n,\, 0,\, nd_{\max} - n - j}}{\binom{nd_{\max}}{n,\, d_i,\, nd_{\max} - n - d_i}}$$

$$+ \frac{\binom{j}{n,\, 0,\, j - n}\binom{nd_{\max} - j}{0,\, d_i,\, nd_{\max} - j - d_i} + \binom{j}{n,\, d_i,\, j - n - d_i}}{\binom{nd_{\max}}{n,\, d_i,\, nd_{\max} - n - d_i}}. \tag{A12}$$

The likelihood function can then be optimized using standard algorithms. In this case we used a version of the BFGS algorithm (*e.g.*, Press *et al.* 1992, pp. 425–430) modified to include constraints on the parameters.