Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 63 (2003) 245–255
Theoretical Population Biology
http://www.elsevier.com/locate/ytpbi

Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium

Rasmus Nielsen* and James Signorovitch

Department of Biometrics, Cornell University, 439 Warren Hall, Ithaca, NY 14853-7801, USA

Received 31 August 2001

Abstract

As large-scale sequencing efforts turn from single genome sequencing to polymorphism discovery, single nucleotide polymorphisms (SNPs) are becoming an increasingly important class of population genetic data. But because of the ascertainment biases introduced by many methods of SNP discovery, most SNP data cannot be analyzed using classical population genetic methods. Statistical methods must instead be developed that can explicitly take into account each method of SNP discovery. Here we review some of the current methods for analyzing SNPs and derive sampling distributions for single SNPs and pairs of SNPs for some common SNP discovery schemes. We also show that the ascertainment scheme has a large effect on the estimation of linkage disequilibrium and recombination, and describe some methods of correcting for ascertainment biases when estimating recombination rates from SNP data.

1. Introduction

Much attention has recently been given to genomic data consisting of variable sites within a species, the so-called single nucleotide polymorphisms (SNPs). SNP data are generated through various protocols. One obvious method for generating SNPs is by direct sequencing of a genomic region (e.g. Zhao et al., 2000). Such SNP data can be analyzed using standard methods applicable to whole DNA sequences. However, most SNP data are not generated by direct sequencing. Often, SNPs are first identified by scanning databases of genomic fragments (e.g. Wang et al., 1998) or expressed sequence tags (ESTs; e.g. Picoult-Newberg et al., 1999) for variable sites. For example, databases generated by shotgun genome sequencing can be used to identify SNPs. The identified SNPs can then be typed in larger samples using high-throughput methods such as denaturing high performance liquid chromatography (DHPLC) or microarrays.

Standard estimators of population genetic parameters cannot be applied to SNP data generated using these protocols. Such methods of SNP discovery will in most cases bias the standard estimators, causing a so-called *ascertainment bias*. However, appropriate population genetic analyses of such data can still be made by modeling the ascertainment method. Methods for considering ascertainment schemes in the analyses of SNP data have been discussed by Nielsen (2000), Kuhner et al. (2000a) and Wakeley et al. (2001). Modeling the method of sampling population genetic markers for the purpose of appropriate statistical analysis dates back at least to Ewens et al. (1981).

In this article we will review some of the available methods for correcting for ascertainment biases in SNP data and we will develop a few new examples and illustrations.

2. The case of no ascertainment bias

Let us first consider the case where the data have been obtained by directly sequencing a region that has been chosen without knowledge regarding the variability of its nucleotides. Let the data be represented by $\mathbf{x}$, where $\mathbf{x}$ is an $n \times s$ matrix of nucleotides from n sequences each consisting of s nucleotides. The columns in $\mathbf{x}$ represent all the sequenced sites, both variable and invariable. The likelihood function, for a vector of relevant population genetic parameters Θ, is any function proportional to

*Corresponding author. Fax: +1-607-255-4698.
E-mail address: rn28@cornell.edu (R. Nielsen).

0040-5809/03/$ - see front matter
doi:10.1016/S0040-5809(03)00005-4

$\Pr(\mathbf{x}|\Theta)$. For multiple linked sites in the same region, this likelihood function cannot be obtained analytically under realistic models, but must instead be obtained by simulation. The likelihood function is written using the representation

$$\Pr(\mathbf{x}|\Theta) = \int_{\Omega} \Pr(\mathbf{x}|\Theta, G) f(G|\Theta)\, dG, \tag{1}$$

where G is a genealogy, Ω is the set of all possible such genealogies, and $f(G|\Theta)$ is the density function of genealogies in Ω. In the absence of recombination, the genealogy is often represented as a labeled history in the sense of Thompson (1975), of which there are $(n!)^2/[n2^{(n-1)}]$, and a set of coalescent times. In the presence of recombination, G represents an ancestral recombination graph (ARG). The ancestral recombination graph is a joint representation of the linked genealogies for multiple sites. The version of the ARG that we will use here represents each site explicitly, i.e. it is not an infinite sites model. For more discussion regarding ARGs, see Hudson (1983, 1985) and Griffiths and Marjoram (1997).

The representation in Eq. (1) suggests that the likelihood function can be evaluated as the expectation of some functional over simulated gene genealogies. For example, in the importance sampling methods by Griffiths and Tavaré (1994a, b) and Stephens and Donnelly (2000), k genealogies, $G_1, G_2, \ldots, G_k$, are simulated from the distribution $h(G; \Theta, \mathbf{x})$, and the likelihood function is evaluated as

$$\Pr(\mathbf{x}|\Theta) \approx \frac{1}{k} \sum_{i=1}^{k} \frac{\Pr(\mathbf{x}|\Theta, G_i) f(G_i|\Theta)}{h(G_i; \Theta, \mathbf{x})}. \tag{2}$$

As shown in Stephens and Donnelly (2000), such simulation schemes are most efficient if the importance sampling function $h(G; \Theta, \mathbf{x})$ closely approximates a density of genealogies proportional to $\Pr(\mathbf{x}|\Theta, G) f(G|\Theta)$.

A related method by Kuhner et al. (1995) uses Markov chain Monte Carlo (MCMC) to simulate genealogies. The advantage of MCMC methods is that it can simulate genealogies directly from the correct distribution, without a need for importance sampling weighting. Their disadvantage is that the sampled genealogies are correlated because they come from the same simulated Markov chain. This correlation often makes it difficult to evaluate the accuracy of the estimate of the likelihood function and to determine the needed simulation time.

Griffiths and Marjoram (1996) and Fearnhead and Donnelly (2001) have developed methods for analyzing recombining sequences using importance sampling schemes. Kuhner et al. (2000b) and Nielsen (2000) have developed similar methods based on MCMC. For nuclear human data, recombination rates appear large enough (e.g. Pritchard and Przeworski, 2001) to invalidate methods that do not explicitly account for recombination over large genomic regions. Some authors (e.g. Harding et al., 1997) have analyzed recombining sequences using methods that do not account for recombination by removing apparent recombinant sequences. The effect of this procedure on any biological inferences made is largely unknown.

3. What is an ascertainment bias?

An ascertainment bias arises when data have not been obtained randomly with respect to the observed data patterns. For example, SNPs might initially have been identified in a small sample (panel). After the initial SNP discovery, the SNPs are then typed in a larger sample of chromosomes. By preferentially sampling SNPs at intermediate frequencies, such a protocol will bias the distribution of allelic frequencies compared to the expectation for a random sample. Fig. 1 illustrates this effect. The open bars represent the allelic distribution of SNPs under a neutral equilibrium model in the limit of low mutation rates (Eq. (17)) for a sample size of $n = 100$. This also corresponds to the frequency spectrum under a standard neutral infinite sites model (Tajima, 1989). The solid bars in Fig. 1 represent the allelic distribution under the following ascertainment procedure (see Eqs. (18) and (20)): in a sample of 5 chromosomes SNPs are identified at random positions in the genome. Each of the identified SNPs is then typed in 95 additional chromosomes for a total sample of size $n = 100$. Notice how much of an effect the ascertainment scheme has on the sampling distribution. Obviously, any inferences we would make regarding demographics or mutational processes would be strongly influenced by the ascertainment scheme if we did not appropriately correct for the bias it introduces.

In general, if the sampling distribution of a random sample without any ascertainment bias is given by $\Pr(\mathbf{x}|\Theta)$, and we denote the ascertainment event by A, then the sampling distribution that accounts for the method of ascertainment is given by

$$\Pr(\mathbf{x}|\Theta; A) = \frac{\Pr(\mathbf{x}, A|\Theta)}{\Pr(A|\Theta)}. \tag{3}$$

In the ascertainment scheme used to generate Fig. 1, A would represent the event that the sampled site is variable in the first 5 sampled chromosomes. Eq. (3) gives the sampling distribution that should be used in correct analyses of such SNP data. The phrase "correcting for ascertainment biases" may, therefore, be a bit misleading. The real problem is identifying the correct sampling distribution for use in statistical inferences. Much of the rest of this paper is concentrated on how to define these sampling distributions for different types of SNP data. Once the correct sampling

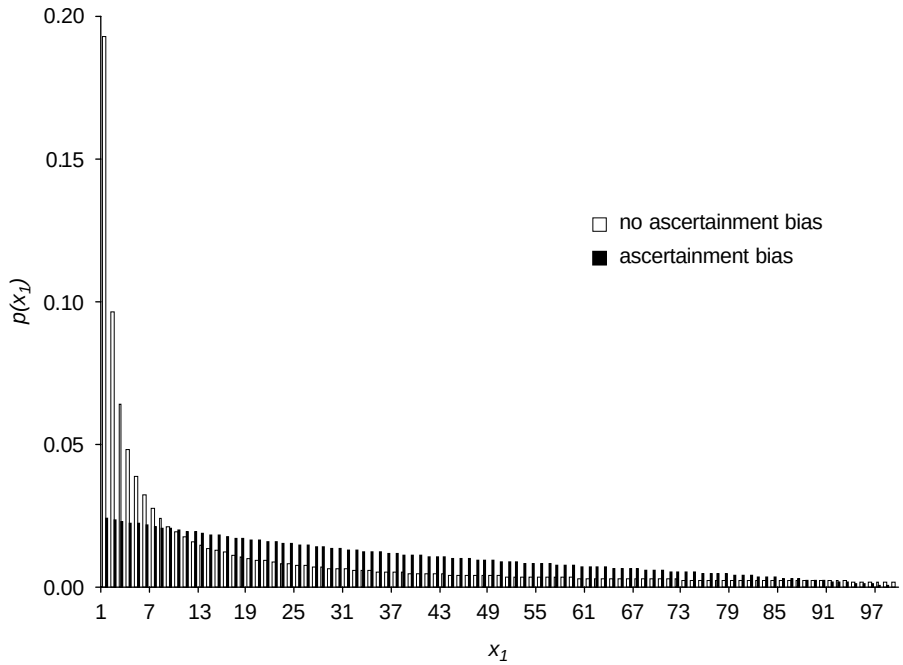

Fig. 1. The distribution of di-allelic locus configurations $\mathbf{x} = (x_1, x_2)$ in a sample size of $n = 100$ under the standard neutral model in the limit of low mutation rates with and without ascertainment bias. These were calculated using Eq. (17) for the case with no ascertainment bias (open boxes) and by Eqs. (18) and (20) for the case in which ascertainment bias is introduced by first identifying SNPs in a panel of 5 chromosomes, then genotyping 95 additional individuals (solid boxes).

distribution has been defined, constructing appropriate likelihood or Bayesian estimators, with associated measures of statistical uncertainty, is a well-trodden path.

4. Information lost regarding invariable sites

A case considered by Nielsen (2000) and Kuhner et al. (2000a) involves data in which SNPs have been identified and typed by direct sequencing, or other methods in which all variable sites are represented in the sample, but where information has been lost regarding invariable sites. In this case, $\mathbf{x}$ contains only the s reported variable sites. However, we still keep track of the distance between sites and the number of invariable sites. The sampling probability for a region of fixed length is then given by

$$\begin{aligned}\Pr(\mathbf{x}|\Theta) &= \int_\Omega \Pr(\mathbf{x}|\Theta, G) f(G|\Theta)\, dG \\ &= \int_\Omega \prod_{i\in S} \Pr(\mathbf{x}_i|\Theta, G) \\ &\quad \times \prod_{i\notin S} \Pr(\text{site } i \text{ is invariable } |G, \Theta) f(G|\Theta)\, dG,\end{aligned} \tag{4}$$

where $\mathbf{x}_i$ is the column of $\mathbf{x}$ corresponding to site i, S is the set of variable sites and G is the ancestral graph for all sites. If sites are independent and identically distributed given the genealogy and there is no recombination, this reduces to

$$\begin{aligned}\Pr(\mathbf{x}|\Theta) = \int_\Omega &\Pr(\text{site is invariable}|G, \Theta)^{L-s} \\ &\times \prod_{i\in S} \Pr(\mathbf{x}_i|\Theta, G) f(G|\Theta)\, dG,\end{aligned} \tag{5}$$

where L is the length of the sequenced region.

Nielsen (2000) considered the case in which information may also be missing for variable sites. That is, among the unobserved sites we have no information regarding which sites are variable or invariable, but all the observed sites are variable. If variable sites are missing at random with respect to their site patterns, inferences can be made by conditioning on the observed sites being variable. Such data may occur, for example, if data from different genomic regions are combined where no information is available regarding polymorphisms in the intervening regions. Let A_S denote the event that all observed sites are variable, then the likelihood function can be defined as the sampling probability

$$\Pr(\mathbf{x}|\Theta; A_S) = \frac{\Pr(\mathbf{x}|\Theta)}{\Pr(A_S|\Theta)}. \tag{6}$$

The numerator can be obtained using standard methods, and the denominator can be obtained analytically for regions without recombination and by simulation for regions with recombination. This is also the appropriate approach to analyzing most allozyme and restriction fragment length polymorphism (RFLP)

data, since invariable loci usually go unreported for these markers.

Kuhner et al. (2000a) considered the case in which the probability of sampling a site given that it is variable equals ϕ and is a known quantity. Again, only variable sites (SNPs) are sampled. The sampling probability is then given by (Kuhner et al., 2000a)

$$\int_\Omega \prod_{i\in S} \phi \Pr(\mathbf{x}_i|\Theta, G) \times \prod_{i\notin S}(1-\phi \Pr(\text{site } i \text{ is variable}|G))f(G|\Theta)\, dG. \quad (7)$$

Kuhner et al. (2000a) refer to this as the "reconstituted DNA" method. They contrast this to what they term a "conditional likelihood" method. For fully linked sites, they define the likelihood function as

$$L(\Theta) = \int_\Omega \prod_{i\in S} \frac{\Pr(\mathbf{x}_i|G,\Theta)}{\Pr(\text{site } i \text{ is variable}|G,\Theta)} \times f(G|\Theta)\, dG. \quad (8)$$

We notice that this expression is valid only if the number of SNPs is not a random variable. Such data might arise if sequencing, for all sequences, has proceeded from a random point in the genome and is terminated when a predetermined number of SNPs has been obtained, and information regarding invariable sites is subsequently lost (Mary Kuhner, pers. comm.). If instead SNPs have been obtained by sequencing regions of fixed length, the appropriate method for conditioning on sites being variable is given by Eq. (6). Eq. (6) is a more general version of Eq. (2) in Kuhner et al. (2000a), which is applicable only to unlinked sites. Eq. (6) can obviously also be applied to regions with moderate or no recombination, as in Nielsen (2000). In that case, it takes the form

$$L(\Theta) = \frac{\int_\Omega \prod_{i\in S} \Pr(\mathbf{x}_i|G,\Theta) f(G|\Theta)\, dG}{\int_\Omega \prod_{i\in S} \Pr(\text{site } i \text{ variable}|\Theta,G) f(G|\Theta)\, dG}. \quad (9)$$

For linked sites, it is necessary to use simulation methods to evaluate the likelihood functions in Eqs. (6)–(9). Examples of such simulations are in Nielsen (2000) and Kuhner et al. (2000a). Other simulation methods, such as the methods of Stephens and Donnelly (2000), could also easily be adapted to the sampling schemes considered here. However, all of these methods assume haplotypic data, i.e. that the haplotypic phase is known. Most SNP data is genotypic and not haplotypic. Analysis of genotypic data can in principle easily be incorporated into the MCMC schemes; however, no studies have yet been published documenting the computational feasibility of this approach.

4.1. Independent sites

The case of no linkage is much more computationally and mathematically tractable compared to that of the case of linked sites. In addition, it might be realistic for much SNP data that have been obtained from random locations in the genome, for example by screening EST databases. For linked SNPs, an approximate analysis based on the assumption of independent markers might provide useful heuristic estimators. A similar approach has, for example, been taken by Hudson (2001) for estimating recombination rates in local regions by considering the likelihood function calculated for pairs of sites. In such approaches, confidence intervals and hypothesis tests must be obtained using simulations.

Analytical results can easily be derived for various mutation models assuming independent sites. For example, assuming an infinite alleles model, the sampling distribution in a neutral equilibrium model without ascertainment bias is given by the Ewens (1972) sampling formula. The approach we will take here considers a model of low mutation rates, i.e. the limit of $\theta \to 0$, where $\theta = 4N\mu$, and μ is the base pair mutation rate per generation and $2N$ is the chromosomal population size. The distributions we obtain then correspond to the marginal distribution of a single variable site under the infinite sites model. The use of the low mutation rate limit for human genetic data is supported by the observation that di-allelic SNPs are almost exclusively observed in such data, although some researchers have argued for the presence of multiple mutations in the same site (Templeton et al., 2000).

The data for a single site, for a single population, consist of the counts of each of the two alleles, i.e. $\mathbf{x} = (x_1, x_2)$, and can be represented simply by the counts of the first allele, x_1, suppressing the dependency on the sample size in the notation. Initially, consider the case in which alleles of type 1, of which there are x_1 copies in the sample, are known to be of the mutant type. We will easily be able to derive analytical expressions by considering the underlying genealogy. In the following it is assumed that the lengths of lineages (edges) in the genealogy are measured by the number of generations scaled by the chromosomal population size, and that the genealogy is well behaved, i.e. the expected time until all lineages have coalesced is finite. We will also assume that mutations arise along a lineage according to a Poisson process with rate $\theta/2$. This corresponds to assuming neutrality in the particular site we are looking at; however, we have not made any assumptions regarding neutrality in linked sites or made any assumptions regarding demographic models. Let t be the length of edges in the genealogy which have x_1 descendents, and let T be the total tree length, i.e. the

sum of the lengths of all lineages in the genealogy. Disregarding the possibility of back mutations, we have

$$\Pr(x_1) = E\left((1-e^{-\theta t/2})e^{-\theta(T-t)/2}\right), \tag{10}$$

where the expectation is taken with respect to the joint distribution of t and T. Notice that this expectation depends on the data through the definition of t and T. The distribution of t and T may depend on some unknown parameters that have been suppressed in the notation. Expanding into a Taylor series, we have

$$\Pr(x_1) = \frac{\theta}{2}E(t) + O(\theta^2) \approx \frac{\theta}{2}E(t) \tag{11}$$

for small values of θ. Similarly, if we condition on the event that the site is variable in the sample, A_S, we have (Nielsen, 2000)

$$\begin{aligned}\Pr_0(x_1|A_S) &:= \lim_{\theta\to 0}\Pr(x_1|A_S)\\ &= \lim_{\theta\to 0}\frac{E\left((1-e^{-\theta t/2})e^{-\theta(T-t)/2}\right)}{E(1-e^{-\theta t/2})}\\ &= E(t)/E(T).\end{aligned} \tag{12}$$

This result is implicit in much previous work, such as Griffiths and Tavaré (1998) and references therein. It is useful because it allows easy evaluation of the likelihood function, even for quite complicated models, as long as we can simulate genealogies. For example, if we consider the case of two populations, the data can be represented as $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$, where x_{ij} is the number of alleles of type j in population i. If we now define t as the length of all edges in which a mutation would lead to x_{11} and x_{21} copies of the mutant allele in the two populations, respectively, Eq. (12) is still valid, and can be evaluated quickly using simulations as

$$\Pr(x1|A_S) \approx \frac{\sum_{i=1}^{k} t_i}{\sum_{i=1}^{k} T_i}, \tag{13}$$

where t_i and T_i are the ith values of t and T, respectively, obtained in k simulation replicates of the genealogical process.

Returning to the case of a single population, we can make some further progress if we assume that only bifurcations occur in the genealogy and that all lineages are exchangeable. These assumptions are valid, for example, under Kingman's (1982) coalescent and its extensions to the case of varying population size. Griffiths and Tavaré (1998) write Eq. (12) in the form

$$\Pr_0(x_1|A_1) = \frac{\sum_{i=2}^{n} i p_{n,i}(x_1)E(T_i)}{\sum_{i=2}^{n} iE(T_i)}, \quad 0<x_1<n, \tag{14}$$

where T_i, is the length of time in the genealogy in which there are i lineages, and $p_{n,i}(x_1)$ is the probability that a mutation arising while there are i lineages leaves x_1 descendants among the n chromosomes in the sample. Under the previously mentioned assumptions,

$$p_{n,i}(x_1) = \frac{\binom{n-x_1-1}{i-2}}{\binom{n-1}{i-2}}, \quad 0<x_1<n, \tag{15}$$

where we define $\binom{i}{j} = 0$ if $j>i$. Together with Eq. (14), Eq. (15) provides an analytical result for the sampling probability expressed only as a function of expected coalescence times (Griffiths and Tavaré, 1998).

So far we have assumed that it is known which nucleotide is the mutant and which nucleotide is ancestral. This might be a reasonable assumption for some data if the mutation rate is sufficiently low and the identity of the SNP in an outgroup species is known. In cases where the ancestral state of the SNP is not known, results can easily be obtained from the ancestral-state-known case. For example, from Eqs. (14) and (15), we have for the ancestral-state-unknown case Nielsen (2000)

$$\begin{aligned}\Pr_0(x_1|A1) = &\frac{1}{\sum_{i=2}^{n} iE(T_i)}\\ &\times\sum_{i=2}^{n}\left(E(T_i)i\frac{\binom{x_1-1}{i-2}+(1-\delta_{x_1,x_2})\binom{n-x_1-1}{i-2}}{\binom{n-1}{i-1}}\right),\\ &0<x_1\leqslant n/2,\end{aligned} \tag{16}$$

where δ_{ij} is the Kronecker delta function.

Under Kingman's coalescent (Kingman, 1982) T_i is exponentially distributed with mean $2/(i(i-1))$, so Eq. (14) reduces to the well-known result for the frequency spectrum under the infinite sites model (e.g. Tajima, 1989)

$$\Pr_0(x_1|A_1) = \frac{x_1^{-1}}{\left(\sum_{i-1}^{n-1} 1/i\right)}, \quad 0<x_1<n, \tag{17}$$

the expression used to generate the open bars in Fig. 1.

5. More realistic ascertainment schemes: panel SNPS

Many ascertainment schemes involve identifying SNPs in a small sample, a panel, and then subsequently typing them in a larger sample. The final data set may then contain all of the panel chromosomes, some of the panel chromosomes or none of the panel chromosomes, in addition to the sampled chromosomes that are not members of the panel. Here we will assume that the

sample size of panel chromosomes is m, the number of chromosomes in the final sample is n and that there are o (for overlap) chromosomes that were used in the panel and also included in the final sample. We will first consider the case in which $o = m$. Let A_p be the event that a polymorphism is detected among the m panel chromosomes, then

$$\Pr(x_1|A_p) = \frac{\Pr(A_p|x_1)\Pr(x_1)}{\Pr(A_p)} = \left\{1 - \left[\binom{x_1}{m} + \binom{x_2}{m}\right]\binom{n}{m}^{-1}\right\}\frac{\Pr(x_1)}{\Pr(A_p)}, \quad x_1 < n. \tag{18}$$

The combinatorial expression in the braces on the right-hand side of the equation is 1 minus the probability that all of the panel SNPs are of type 1 minus the probability that all of the panel SNPs are of type 2. Both $\Pr(x_1)$ and $\Pr(A_p)$ may depend on some unknown parameters that, as previously, have been suppressed in the notation. Sampling distributions can then be developed as in the previous chapter, for example using

$$\lim_{\theta\to 0}\frac{\Pr(x_1)}{\Pr(A_p)} = \frac{E(t)}{\sum_{i=2}^{m} E(T_i)}, \tag{19}$$

i.e. the total tree length of a sample of size m is now in the denominator. Under Kingman's (1982) coalescent, we have

$$\frac{E(t)}{\sum_{i=2}^{m} E(T_i)} = \frac{x_1^{-1}}{\sum_{i=1}^{m-1} 1/i}, \quad 0 < x_1 < n, \tag{20}$$

which combined with Eq. (18) provides the sampling distribution plotted in solid boxes in Fig. 1.

We consider next the case where $o = 0$. Again, A_p is defined as the event that a polymorphism is detected among the m panel SNPs. The sampling distribution is now given by

$$\Pr(x_1|A_p) = \sum_{i=1}^{m-1}\binom{n}{x_1}\binom{m}{i}\binom{n+m}{x_1+i}^{-1} \times \frac{\Pr(\mathbf{y}^{(i)})}{\Pr(A_p)}, \quad 0 \leqslant x_1 \leqslant n, \tag{21}$$

where $\mathbf{y}^{(i)}$ is the sample augmented by the panel SNPs which, for the sake of generality, is defined as $\mathbf{y}^{(i)} = (x_1 + i, x_2 + m - o - i)$.

In the general case in which $0 \leqslant o \leqslant m$:

$$\Pr(x_1|A_p) = \sum_{i=1}^{m-1}\sum_{v=0}^{i}\binom{o}{v}\binom{n-o}{x_1-v}\binom{m-o}{i-v}\binom{n+m-o}{x_1+i-v}^{-1} \times \frac{\Pr(\mathbf{y}^{(i)})}{\Pr(A_p)}, \quad 0 \leqslant x_1 \leqslant n. \tag{22}$$

In some cases, SNPs are reported only if they had some minimum frequency in the panel. Typically, singletons are not reported because of the large probability that they are caused by a sequencing error. The sampling probability when singletons are not reported in the panel sample becomes

$$\Pr(x_1|A_{p2}) = \sum_{i=2}^{m-2}\sum_{v=0}^{i}\binom{o}{v}\binom{n-o}{x_1-v}\binom{m-o}{i-v}\binom{n+m-o}{x_1+i-v}^{-1} \times \frac{\Pr(\mathbf{y}^{(i)})}{\Pr(A_{p2})}, \quad 0 \leqslant x_1 \leqslant n, \tag{23}$$

where A_{p2} is the event that there are at least two copies of the rarest allele in the panel sample. It is worth noting that Eqs. (21)–(23) allow SNPs identified as variable in the panel, but found to be invariable in the final sample, to be used for inference. We see that for selection of single SNPs, it is possible to define sampling distributions, and thereby likelihood functions, for almost any imaginable ascertainment scheme. Extensions to the case of population subdivision can also be derived similarly to the expressions derived here. In general, as long as we can calculate the sampling distribution without an ascertainment bias, it is relatively easy to obtain appropriate sampling distributions, even for complicated ascertainment schemes, for independent SNPs.

6. Genomic fragments

In the previous section we treated cases where a SNP is selected (ascertained) based on properties of the SNP itself. However, in some cases SNPs may be selected on the basis of the genomic fragment (e.g. ESTs) to which they belong. For example in some of the data of Ardlie et al. (2001), genomic fragments that contained at least two segregating SNPs in the panel were sequenced directly in the final sample. However, direct sequencing might recover more SNPs than originally identified in the panel. An approach for analyzing such data, under the infinite sites model, was described in Wakeley et al. (2001). They assumed that there is no recombination within fragments but free recombination between fragments. The likelihood function for such data can then be calculated using a method based on simulating

coalescence genealogies. The data in a single fragment is, for computational reasons, reduced to the allelic frequencies, i.e. $\mathbf{x} = \{(x_{11}, x_{12}), (x_{22}, x_{22}), \ldots, (x_{s1}, x_{s2})\}$, where s is a random variable representing the number of SNPs in the fragment. Reducing the data to allelic frequencies ignores haplotype information, but greatly reduces the computational complexity of the problem and allows direct analysis of genotypic data. Let the total number of SNPs variable only in the sample be s_D, in both the sample and the panel be s_O, and only in the panel be s_A. The only known quantity is the observed number of SNPs in the sample $s = s_O + s_D$. The ascertainment event we consider here is $\{s_A + s_O > 1\}$, but the method easily generalizes to other ascertainment schemes. The likelihood function is then given by

$$\Pr(\mathbf{x}|s_A + s_O > 1) = \frac{\Pr(\mathbf{x}, s_A + s_O > 1)}{\Pr(s_A + s_O > 1)}. \tag{24}$$

This expression can be evaluated, even for complex demographic models, using simulation. To do this, we will consider the joint genealogy of the sample and the panel chromosomes. Let T_D be the total tree length of lineages in the joint genalogy in which all descendent chromosomes belong only to the sample, T_A be the total tree length of lineages in which all the descendent chromosomes belong only to the panel, and T_O be the total tree length of lineages that have both panel and sample chromosomes as descendents. Also, let t_i be the length of all lineages in the genealogy in which a single mutation would cause data pattern (x_{i1}, x_{i2}), among the lineages contributing to T_D and T_O. Given these times, the total number of SNPs variable only in the sample (s_D), in both the sample and the panel (s_O), and only in the panel (s_A) are independent Poisson random variables with means $\theta T_D/2$, $\theta T_O/2$, and $\theta T_A/2$, respectively. Then

$$\Pr(s_A + s_D > 1) = 1 - E\left[\left(1 + \frac{\theta(T_O + T_A)}{2}\right) e^{-\frac{\theta(T_O+T_A)}{2}}\right], \tag{25}$$

where the expectation is taken with respect to the joint distribution of T_O and T_A. Likewise, assuming that it is known which of the sample SNPs were also variable in the panel

$$\Pr(\mathbf{x}, s_A + s_O > 1) = E\left[\sum_{s_o=0}^{s} \Pr(\mathbf{x}|s_O, s_D, T_O, T_D, t_i) \times \Pr(s_D|T_D)\Pr(s_A + s_O > 1|s_O, T_A)\right]. \tag{26}$$

The expectation is over the joint density of T_O, T_D, T_A, and t_i, $i = 1, 2, \ldots, s$. $\Pr(s_O|T_O)$ and $\Pr(s_D|T_D)$ are Poisson random variables with means $\theta T_0/2$ and $\theta T_D/2$, respectively. Assuming a Poisson process of mutation, mutations are distributed uniformly along the lineages of the genealogy. Therefore,

$$\Pr(\mathbf{x}|s_O, s_D, T_O, T_D, t_i) = \left(\prod_{i=1}^{s} \frac{t_i}{T_O + T_D}\right). \tag{27}$$

Also,

$$h(s_O, T_A) := \Pr(s_A + s_O > 1|s_O, T_A) = \begin{cases} 1 & \text{if } s_O = 0, \\ 1 - e^{-\theta T_A/2} & \text{if } s_O = 1, \\ 1 - (1 + \theta T_A/2)e^{-\theta T_A/2} & \text{if } s_O \geqslant 2. \end{cases} \tag{28}$$

So Eq. (26) reduces to

$$\Pr(\mathbf{x}, s_A + s_O > 1) = E\left[\left(\prod_{i=1}^{s} \frac{t_i}{T_O + T_D}\right)\left(\frac{\theta(T_O + T_D)}{2}\right)^{s-s_O} \times e^{-\frac{\theta(T_O+T_D)}{2}} \sum_{s_O=0}^{s} \frac{h(s_O, T_A)}{s_O!(s - s_O)!}\right]. \tag{29}$$

This expression differs in some details from the expressions in Wakeley et al. (2001). For example it was assumed in Wakeley et al. (2001) that s_O was known. Wakeley et al. (2001) also consider more complicated demographic models and other ascertainment schemes. This, and similar expressions can easily be evaluated by directly simulating coalescence genealogies as in Eq. (13). This approach was in Wakeley et al. (2001) found to be computationally feasible for large data sets containing multiple fragments with 2–10 SNPs on each fragment with models involving multiple populations. It was used to estimate population growth rates and migration rates from the data by Ardlie et al. (2001). The method provides a general inference framework for SNP fragment data, such as ESTs, under the assumption that no recombination occurs within fragments. However, the framework could easily be generalized to the case of recombination by evaluating the expectation in Eq. (29) using simulations of models that include recombination, as in Hudson (1983).

7. Linkage disequilibrium

There has recently been considerable interest in estimating the level of linkage disequilibrium in human SNP data (e.g. Reich et al., 2001; Ardlie et al., 2001). The main reason for the interest is the potential utility of SNPs in linkage disequilibrium mapping. There has been particular interest in the rate at which linkage disequilibrium decays with distance. One of the interesting observations is that there seems to be a shortage of linkage disequlibrium at short distances and an excess of linkage disequilibrium over long genomic distances

(e.g. Pritchard and Przeworski (2001); Ardlie et al., 2001).

In the following, we will extend some results derived in the previous sections to the case of two loci, and show how an ascertainment scheme influences the decay of linkage disequilibrium with distance. Furthermore, we will demonstrate how the Hudson (2001) estimator of recombination rate can easily be adapted to account for the method of ascertainment.

We will first consider the effect of sampling on measures of linkage disequilibrium. For two loci, SNP data can be represented by the vector $\mathbf{x} = (x_{00}, x_{10}, x_{01}, x_{11})^{\mathrm{T}}$, where x_{ij} is the number of haplotypes with allele i in the first locus and allele j in the second locus. The two most common linkage disequilibrium statistics are

$$r^2(\mathbf{x}) = \frac{(p_{1*}p_{*1} - p_{11})^2}{p_{1*}(1-p_{1*})p_{*1}(1-p_{*1})} \tag{30}$$

and

$$|D'(\mathbf{x})| = \begin{cases} \dfrac{p_{1*}p_{*1} - p_{11}}{\min(p_{1*}p_{*1}, 1-p_{1*})(1-p_{*1})}, & p_{1*}p_{*1} \geqslant p_{11}, \\ \dfrac{p_{11} - p_{1*}p_{*1}}{\min(p_{1*}, 1-p_{1*})(1-p_{*1})}, & p_{1*}p_{*1} < p_{11}, \end{cases} \tag{31}$$

where

$$p_{11} = \frac{x_{11}}{x_{00}+x_{01}+x_{10}+x_{11}},$$
$$p_{1*} = \frac{x_{11}+x_{10}}{x_{00}+x_{01}+x_{10}+x_{11}},$$
$$p_{*1} = \frac{x_{11}+x_{01}}{x_{00}+x_{01}+x_{10}+x_{11}}.$$

The solid lines in Figs. 2a and b show the expectations of $|D'|$ and r^2, respectively, over a range of recombination rates for independent pairs of variable loci identified in a sample of 100 chromosomes (see Eq. (33)). The dotted lines in Figs. 2a and b show the expectations of $|D'|$ and r^2, respectively, when independent pairs of variable loci are first identified in a panel of 5 chromosomes, then typed in 95 more chromosomes for a total sample size of 100 (see Eq. (36)). Notice that for both statistics the decay of linkage disequilibrium is much faster under the panel ascertainment scheme. An intuitive explanation is that the panel ascertainment scheme has preferentially selected loci containing high frequency alleles. Such loci tend to have deeper than average genealogies, providing more opportunity for recombination in the ancestry of the sample, and thereby less linkage disequilibrium. Also notice that r^2 takes higher values under the panel ascertainment scheme and that $|D'|$ exhibits the opposite pattern. An intuitive understanding of this effect can be gained by considering how each statistic treats low-frequency SNPs. As an extreme example, consider the case of

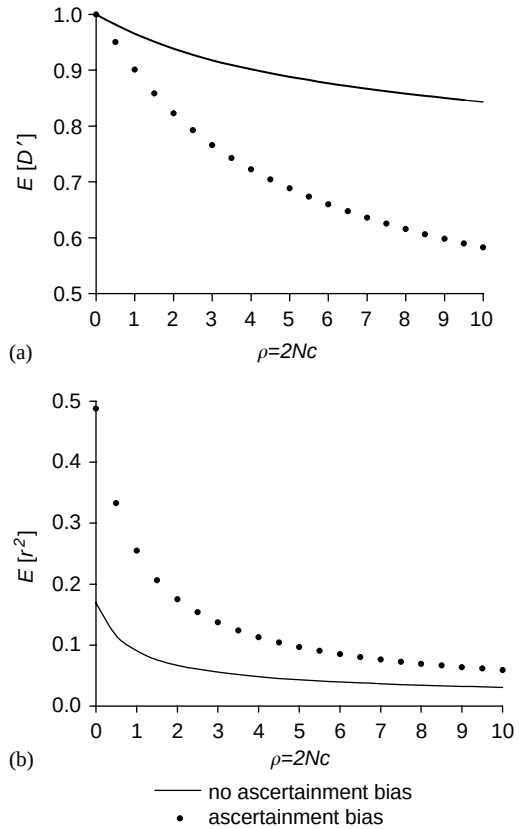

Fig. 2. Expected values of the summary statistics for linkage disequilibrium, $|D'|$ and r^2, with and without ascertainment bias. For each value of $\rho = 2Nc$ in Hudson's (2001) table of two-locus sampling probabilities, we used Eq. (33) to calculate the expectation without sampling bias (solid lines), and Eq. (36) to calculate the expectations under an ascertainment scheme in which a panel of 5 chromosomes is first used to identify SNPs (dotted line). The curves were interpolated with cubic splines.

two singleton SNPs observed in a large sample. Whether the loci are completely linked or unlinked, it is most likely that three haplotypes will be observed. This 3-haplotype configuration reveals little correlation between the loci, so r^2, which can be thought of as the squared correlation coefficient for the two-locus allele patterns, returns low values. $|D'|$, on the other hand, has the extreme value of $p_{11} - p_{1*}p_{*1}p_{11} - p_{1*}p_{*1}$ for the given marginal allele frequencies in the denominator. Since the case of two singletons only allows two possible two-locus configurations, $|D'|$ can only return 1, its highest value. For example, given the data $\mathbf{x}_{OBS} = \{x_{00} = 98, x_{10} = 1, x_{01} = 1, x_{11} = 0\}$, $r^2(\mathbf{x}_{OBS}) = 0.012$ and $|D'(\mathbf{x}_{OBS})| = 1.0$. The data $\mathbf{x}_{OBS}$ suggest extremely low linkage disequlibrium, in the r^2 sense, and complete linkage disequilibrium in the $|D'|$ sense. Since the panel ascertainment scheme greatly decreases the probability of sampling low-frequency SNPs, these opposing behaviors of r^2 and $|D'|$ become apparent.

We realize that the ascertainment scheme has a large effect on the observed values of the linkage

disequilibrium statistics. Pritchard and Przeworski (2001) have argued that instead of focusing on measures of linkage disequilibrium such as r^2 and $|D'|$, estimates of $\rho = 2Nc$ might be more useful as a standard for comparing levels of linkage and recombination rates among different genomic regions and among different studies. Here, c is the recombination rate per generation and N is the chromosomal population size. One of the reasons not to use r^2 and $|D'|$ is that the variances of these linkage disequilibrium statistics are quite high. The fact that the linkage disequilibrium statistics are highly sensitive to the ascertainment schemes also lends further support to the argument of Pritchard and Przeworski (2001).

Nielsen (2000) and Hudson (2001) have considered extensions of Eq. (12) to more than one locus. Let t_{ij} be the length of lineage j in locus i. Also, let I_{jk}, be an indicator function that takes on the value 1 if one mutation in lineage j of locus 1 and one mutation in lineage k of locus 2, with no other mutations occurring in the history of the genealogies, would generate exactly the data pattern $\mathbf{x} = (x_{00}, x_{10}, x_{01}, x_{11})^{\mathrm{T}}$. Also, let A_S be the event that both locus 1 and locus 2 are variable in the sample, and $T^{(i)}$ be the total tree length in locus i. Then (Hudson, 2001)

$$\begin{aligned}\Pr(\mathbf{x}|A_s) &= \lim_{\theta\to 0}\frac{E\left(\sum_{j,k} I_{ij}(1-e^{-\theta t_{1j}/2})(1-e^{-\theta t_{2k}/2})e^{-\theta(T^{(1)}-t_{1j})/2}e^{-\theta(T^{(2)}-t_{2k})/2}\right)}{E\left((1-e^{-\theta T^{(1)}/2})(1-e^{-\theta T^{(2)}/2})\right)}\\ &= E\left(\sum_{j,k} I_{ij}t_{1i}t_{2k}\right)/E\left(T^{(1)}T^{(2)}\right). \qquad (32)\end{aligned}$$

Now we can express the equation used to produce the solid lines in Fig. 2 as

$$E_0[r^2|A_S] = \sum_{\mathbf{x}\in\psi} r^2(\mathbf{x})\Pr_0(\mathbf{x}|A_s) \qquad (33)$$

and likewise for $|D'|$, where ψ is the set of all two-locus configurations in which both sites are variable.

Hudson (2001) has tabulated the values of $E(\sum_{j,k} I_{ij}t_{1i}t_{2k})$for multiple values of ρ, for all possible sample configurations of size $n = 100$, under the neutral equilibrium model (available from http://pondside.uchicago.edu/ecol-evol/faculty/hudson_r.html). When first these values are known, likelihood estimation of ρ can be done using Eq. (32). For multiple markers, a pseudo-likelihood estimate can be obtained by combining the likelihood function for loci serially along the chromosome (Hudson 2001). In Hudson (2001) $E(\sum_{j,k} I_{ij}t_{1i}t_{2k})$ is estimated on a grid for 14 values of ρ. In our implementation, we fit a cubic spline (e.g. Lange, 2000, pp. 104–106) to the 14 values to obtain a smooth likelihood function (as in Fig. 4).

To find the distribution of this estimator under a simple model, we simulated 100,000 data sets, each containing 200 pairs of variable SNPs, by selecting variable pairs of sites from independent simulations of ARGs for 100 samples of a 300 bp fragment with $\theta = 0.001$ per site per generation and $\rho = 2Nc = 5.0$, where c is the per generation recombination rate between the outer-most sites in the fragment. For each data set we estimated ρ by maximizing the product of Eq. (32) over all 200 pairs. The distribution of these estimates is shown in Fig. 3a. To model a panel ascertainment scheme, we performed the simulations as above, but sampled only pairs of sites that were variable in the first five sequences. The distribution of estimates of ρ obtained from these simulations is shown by the dotted line in Fig. 3b. The estimates in the panel ascertainment scheme have less variance because the panel data have a much larger proportion of high frequency alleles, which contain more information. However, the panel ascertainment scheme produces estimates that are biased towards small values. We can correct for this bias by using the correct sampling distribution as the likelihood function. Under the panel ascertainment scheme, the sampling distribution is given by

$$\Pr_0(\mathbf{x}|A_p) = \Pr(A_p|\mathbf{x})\frac{E(\sum_{j,k} I_{ij}t_{1i}t_{2k})}{E(T_m^{(1)}T_m^{(2)})}, \qquad (34)$$

where $T_m^{(j)}$ is the total tree length of site i in a sample of size m, and

$$\Pr(A_p|\mathbf{x}) = 1 - \frac{\binom{x_{10}+x_{11}}{m} + \binom{x_{01}+x_{11}}{m} + \binom{x_{10}+x_{00}}{m} - \sum_{k\in\{x_{11},x_{10},x_{01},x_{00}\}}\binom{k}{m}}{\binom{n}{m}}. \qquad (35)$$

We can use this expression to define the likelihood function for ρ for the SNP data obtained form the panel ascertainment scheme. As shown by the solid line in Fig. 3b, appropriate use of this sampling distribution produces an approximately unbiased estimator of ρ. We also used this sampling distribution to compute the expectations of the linkage disequilibrium statistics under the panel ascertainment scheme (dotted lines in Fig. 2) as

$$E[r^2|A_p] = \sum_{x\in\psi} r^2(x)\Pr_0(x|A_p) \qquad (36)$$

and likewise for $|D'|$, where ψ is all two-locus configurations in which both sites are variable.

The standardized likelihood functions using Eqs. (32) and (34) for a sample data set are plotted in Fig. 4. Obviously, appropriately modeling the method of ascertainment is of great importance when defining the likelihood function.

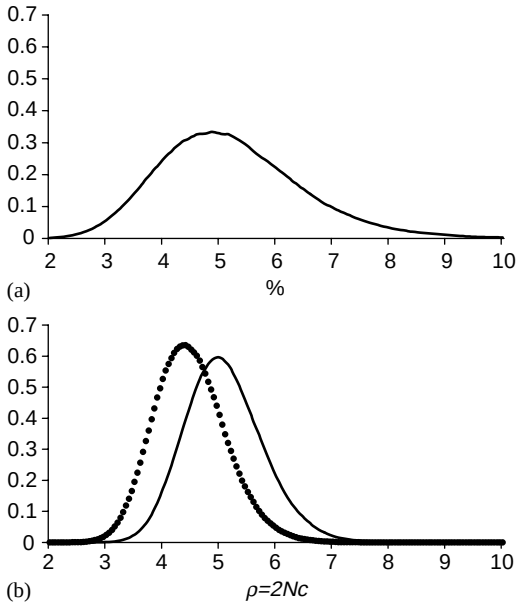

Fig. 3. Histograms of the Hudson (2001) estimator of ρ from 100,000 simulated data sets, each containing 200 pairs of variable SNPs, where pairs were identified in independently simulated ARGs of 100 chromosome fragments of 300 bp each with $\theta = 0.001$ and $\rho = 2Nc = 5$, where c is the per-generation recombination rate between the outermost sites. In Fig. 3a, all variable pairs are included. The dotted line in Fig. 3b is generated under an ascertainment scheme that consider only pairs of sites that are variable in a sub-sample of 5 chromosomes. The solid line in Fig. 3b results from the same ascertainment scheme, but the estimator is now based on the correct sampling distribution corresponding to the ascertainment scheme (Eq. (34)). The mean estimates for ρ are 5.19, 4.45 and 5.06, respectively. The distributions were smoothed using a sliding window of 10 bins.

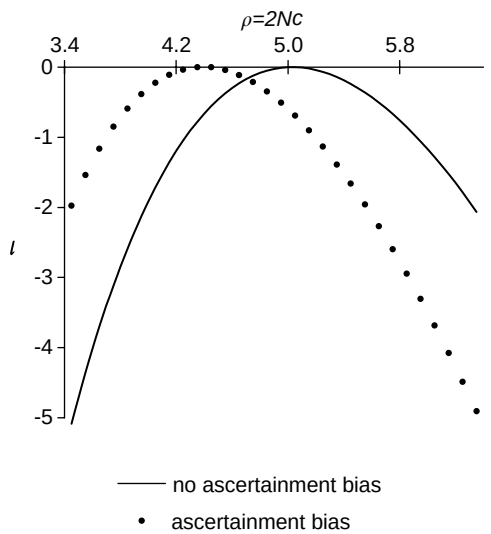

Fig. 4. Re-scaled log-likelihood curves for ρ calculated from a single simulated data set of 200 pairs of variable SNPs generated as in Fig. 3b. The dotted line shows the likelihood function when the ascertainment bias is not taken into account. The solid line is the correct likelihood function appropriately accounting for the ascertainment scheme (Eq. (34)). The true value of ρ is 5 and the maxima occur at $\rho = 4.35$ and 4.99, respectively.

8. Conclusion

The method of ascertainment has a large effect on the allelic distributions of samples from natural populations. The SNP data currently being generated provide an excellent opportunity to address population genetic questions. However, valid inferences based on such data require appropriate modeling of the ascertainment process. This is a point that has been previously emphasized in Nielsen (2000), Kuhner et al. (2000a) and Wakeley et al. (2001). Here we have shown how appropriate modeling of realistic ascertainment schemes can be done in a variety of situations. In particular, it is quite easy to correct the sampling distribution for SNPs at one or two loci with a combinatorial expression that models the ascertainment scheme. It is computationally very difficult to evaluate the full likelihood function based on multiple linked loci (Griffiths and Marjoram, 1996; Nielsen, 2000; Kuhner et al., 2000b; Fearnhead and Donnelly, 2001). An attractive alternative is to develop pseudo-likelihood estimators, such as in Hudson (2001), based on single loci or pairs of loci. Such estimators will be particularly attractive under very complicated ascertainment schemes, where it would otherwise be difficult or impossible to evaluate the correct sampling distribution.

Acknowledgments

This research was supported by NSF Grant DEB-0089487 to RN. We thank John Wakeley, Mary Kuhner, Joe Felsenstein and Peter Beerli for many stimulating discussions on the topic and two anonymous reviewers for helpful comments on the manuscript.

References

Ardlie, K., Liu-Cordero, S.N., Eberle, M.A., Daly, M., Barrett, J., Winchester, E., Lander, E.S., Kruglyak, L., 2001. Linkage disequilibrium and gene conversion. Am. J. Hum. Genet. 69, 582–589.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Pop. Biol. 3, 87–112.

Ewens, W.J., Speilman, R.S., Harris, H., 1981. Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc. Natl. Acad. Sci. USA 78, 3748–3750.

Fearnhead, P., Donnelly, P., 2001. Estimating recombination rates from population genetic data. Technical report available from http://www.stats.ox.ac.uk/~fhead/

Griffiths, R.C., Marjoram, P., 1996. Ancestral inference from samples of DNA sequences with recombination. J. Comp. Biol. 3, 479–502.

Griffiths, R.C., Marjoram, P., 1997. An ancestral recombination graph. In: Tavare, S., Donnelly, P. (Eds.), Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications. Spriger, Berlin, pp. 257–270.

Griffiths, R.C., Tavaré, S., 1994a. Simulating probability distributions in the coalescent. Theor. Popul. Biol. 46, 131–159.

Griffiths, R.C., Tavaré, S., 1994b. Ancestral inference in population genetics. Stat. Sci. 9, 307–319.

Griffiths, R.C., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. Stochastic Models 14, 273–295.

Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S., Clegg, J.B., 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. 60, 772–789.

Hudson, R.R., 1983. Properties of the neutral allele model with intergenic recombination. Theor. Pop. Biol. 23, 183–201.

Hudson, R.R., 1985. The sampling distribution of linkage disequilibrium under an infinite alleles model without selection. Genetics 109, 566–631.

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

Kingman, J.F.C., 1982. The coalescent. Stochastic Proc. Appl. 13, 235–248.

Kuhner, M.K., Yamato, J., Felsenstein, J., 1995. Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. Genetics 140, 1421–1430.

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

Kuhner, M.K., Yamato, J., Felsenstein, J., 2000b. Maximum likelihood estimation of recombination rates from population data. Genetics 156, 1393–1401.

Lange, K., 2000. Numerical Analysis for Statisticians, Springer Series in Statistics and Computing. Springer, Berlin.

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

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

Pritchard, J.K., Przeworski, M., 2001. Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69, 1–14.

Stephens, M., Donnelly, P., 2000. Inference in molecular population genetics. J. Roy. Statist. Soc. Ser. B 62, 605–655.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Templeton, A.R., Clark, A.G., Weiss, K.M., Nickerson, D.A., Boerwinkle, E., Sing, C.F., 2000. Recombinational and mutational hotspots within the human lipoprotein lipase gene. Am. J. Hum. Genet. 66, 69–83.

Thompson, E.A., 1975. Evolutionary Trees. Cambridge University Press, Cambridge, UK.

Wakeley, J., Nielsen, R., Ardlie, K., Liu-Cordero, S.N., 2001. The discovery of single nucleotide polymorphisms and inferences about human demographic history. Am. J. Hum. Genet., in review.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lander, E.S., et al., 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280, 1077–1082.

Zhao, Z., Jin, L., Fu, Y.-X., Ramsay, M., Jenkins, T., Leskinen, E., Pamilo, P., Trexler, M., Patthy, L., Jorde, L.B., Ramos-Onsins, S., Yu, N., Li, W-H., 2000. Worldwide dna sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA 97, 11354–11358.