Population genetic analysis of ascertained SNP data

Rasmus Nielsen

Department of Biological Statistics and Computational Biology, Cornell University, 439 Warren Hall, Ithaca, NY 14853-7801, USA
Tel: +1 607 255 1643; Fax: +1 607 255 4698; E-mail: rn28@cornell.edu

Date received (in revised form): 16th January 2004

Abstract

The large single nucleotide polymorphism (SNP) typing projects have provided an invaluable data resource for human population geneticists. Almost all of the available SNP loci, however, have been identified through a SNP discovery protocol that will influence the allelic distributions in the sampled loci. Standard methods for population genetic analysis based on the available SNP data will, therefore, be biased. This paper discusses the effect of this ascertainment bias on allelic distributions and on methods for quantifying linkage disequilibrium and estimating demographic parameters. Several recently developed methods for correcting for the ascertainment bias will also be discussed.

Keywords: *single nucleotide polymorphisms, ascertainment bias, statistical analysis, linkage disequilibrium, demographic parameters*

Introduction

Many of the resources previously allocated to genomic sequencing have recently been devoted to the typing and discovery of single nucleotide polymorphisms (SNPs), resulting in a rapid increase in the amount of publicly available SNP data. In August 2003, the public dbSNP database at NCBI[1] contained 268,374 SNPs with allele frequency information (Build 116). In August 2002 (Build 106), it had contained only 47,577 SNPs. In one year, the number of SNPs in dbSNP with frequency information increased more than five-fold.

The major objective of most SNP typing and discovery studies is to develop a resource for genetic mapping studies.[2,3] The large SNP datasets will provide an invaluable resource in both pedigree-based studies and association mapping studies.[4–6] The large SNP datasets also provide a remarkable resource for human population genetic analysis, however. Population geneticists will be interested in estimating recombination rates and levels of linkage disequilibrium,[7–10] as well as parameters relating to the demographics and ancestry of human populations using the available SNP data.[11,12] In addition, large SNP datasets can be used to scan the genome for regions that may have been targeted by selection.[13–15] SNPs targeted by selection are presumably more likely to be disease associated.[16,17]

Unfortunately, most of the standard analytical methods usually used for population genetic inferences are not applicable to the majority of the SNP data. Almost all available population genetic methods assume that the analysed loci have been sampled randomly among the pool of all loci. Most SNP loci, however, were originally identified through an SNP discovery process that tends to select loci with particular allelic distributions.[18–21] This introduces an ascertainment bias which, if uncorrected, will bias parameter estimates and lead to false inferences.[22,23]

The aim of this review is to discuss the effects of the ascertainment bias for some common SNP discovery protocols and also to discuss some recently developed methods for correcting the ascertainment bias problem. If not otherwise stated, it will be assumed that SNPs are effectively unlinked. This will usually be a reasonable assumption for datasets containing multiple SNPs scattered throughout the genome. The case of linked SNPs can similarly be dealt with, however, and is discussed elsewhere.[11,23,24].

Ascertainment schemes

There are probably more different SNP discovery protocols (ascertainment schemes) than there are research groups involved in SNP discovery. It is unlikely that any particular method of addressing the problem of ascertainment bias is appropriate for all ascertainment schemes. Most ascertainment schemes have the common feature that SNPs are originally discovered in a relatively small sample, however. The SNPs are then subsequently typed in a larger sample for the purpose of population genetic inferences. Small samples have a relatively smaller probability of containing rare alleles than large samples. The effect is, therefore, that in the final typed sample there is an excess of SNP loci with common alleles and a deficiency of loci with rare alleles. This deficiency of rare alleles in the typed sample is a common feature in many SNP datasets. Here, the term *ascertainment sample* is used to denote the sample used originally to discover the SNPs, while *typed sample* is used to denote the final sample used for population genetic inferences.

The ascertainment sample usually consists of two or more gene copies from a panel. A panel is a group of individuals whose DNA has been used in the SNP discovery process. SNPs are usually originally identified from an alignment of sequences or a collection of sequences, the SNP discovery alignment. In some cases, all individuals in a panel have been typed in the ascertainment sample and are represented in the SNP discovery alignment. In such cases the depth (d) of the SNP discovery alignment is equal to twice the number of diploid individuals in the panel (n_p). Very often, however, only a subset of the panel haplotypes (gene copies) have been typed for each SNP in the ascertainment sample ($d < n_p$). This may occur, for example, if the SNP discovery process is based on data obtained from shotgun sequencing. Although one may know how many sequences were included in the alignment for each SNP that was discovered, one will not know the true depth of the alignment because the sequences have been sampled with replacement from the panel sequences, ie the alignment in the ascertainment sample for a particular SNP may contain the same sequence more than once. Furthermore, the information regarding the depth of the alignment for each SNP may have been lost through time. This may occur, for example, because the number of sequences in the alignment has increased through time and no records have been kept regarding the number of sequences on which the SNP discovery process was based. SNP discovery protocols may, therefore, differ in the assumptions one can make regarding the depth of the ascertainment sample. Ascertainment schemes may also vary depending on whether singletons or low-frequency SNPs have been eliminated directly, on various aspects relating to the SNP verification process (eg re-sequencing) and on the method used for base-calling. Finally, the effect of the ascertainment bias will differ depending on whether the sequences used for SNP discovery is a subset of the data in the typed sample or if there is no overlap between these two sets of data.

Effect on the frequency spectrum

The frequency spectrum summarises the allelic distribution in a sample. Under the classical neutral coalescence model,[25,26] the probability of observing X copies of a mutant allele in a sample of size n is:[27]

$$\Pr(X = x) = \frac{x^{-1}}{\sum_{i=1}^{n-1} 1/i}, \quad 0 < x < n. \tag{1}$$

The distribution of X in Equation (1) gives the expected frequency spectrum for this model. The particular version shown in Equation (1) assumes that it is known which allelic type is mutant and which allelic type is ancestral.

To illustrate the effect of the ascertainment process on the frequency spectrum, it should be assumed that each SNP was originally discovered in a small sample of known size d, and subsequently typed in a larger sample of size $n - d$, resulting in a final sample of size n, ie, the ascertainment sample is a subset of the typed sample. Only loci that were variable in the sample of size d are included in the analysis. Then:[11,28,29]

$$\Pr(X = x) = \frac{\Pr(\textit{Ascertainment}|X = x)/x}{\sum_{i=1}^{n-1} \Pr(\textit{Ascertainment}|X = i)/i}, \quad 1 \le x \le n - 1, \tag{2}$$

where

$$\Pr(\textit{Ascertainment}|X = i) = 1 - \left[\binom{i}{d} + \binom{n-i}{d}\right]\binom{n}{d}^{-1}.$$

Sherry *et al.*[30] used a similar expression to test 'goodness of fit' of a standard neutral model to data of population frequencies of *Alu* elements, in a case where the *Alu* elements had originally been detected in a single diploid genome.

The frequency spectrum — when it is known which allele is ancestral, the so-called folded frequency spectrum — is given by $\Pr(X = x) + \Pr(X = n - x)$.

Figures 1a and 2a show the unfolded and folded frequency spectra respectively, in a sample of size $n = 20$ when $d = 2$, $d = 5$, $d = 10$ and $d = 20$ (no ascertainment bias). Notice that the effect of the ascertainment bias is quite pronounced, even when d is relatively large ($d = 10$). In an ascertainment sample of size $d = 2$, the folded frequency spectrum becomes uniform on all values from 1 to $n/2$. Clearly, any population genetic inferences based on allele frequencies will be strongly influenced by the ascertainment scheme.

In many cases, it may be more realistic to assume that the ascertainment sample is not a subset of the typed sample. In that case, invariable sites (sites in which the allele frequency is $X = 0$ or $X = n$) may occur in the typed sample. In the following, it should be assumed that such invariable loci in the typed sample have been discarded. This will often be the case because SNPs that are not variable in the typed sample may be assumed to be generated artifactually by sequencing or alignment errors in the ascertainment sample. These loci will be categorised as loci in which the polymorphism could not be verified. The expected frequency spectrum for such samples can be obtained by considering the possible allele frequencies in the sample of size $n + d$ that arises by pooling the ascertainment sample and the typed sample. If the allele frequency in the ascertainment sample is known, this case is identical to the case where the typed sample is a subset of the ascertainment sample, and the expected frequency spectrum in the sample of size $n + d$ can be obtained using Equation (2). If the allele frequency in the ascertainment sample is unknown,

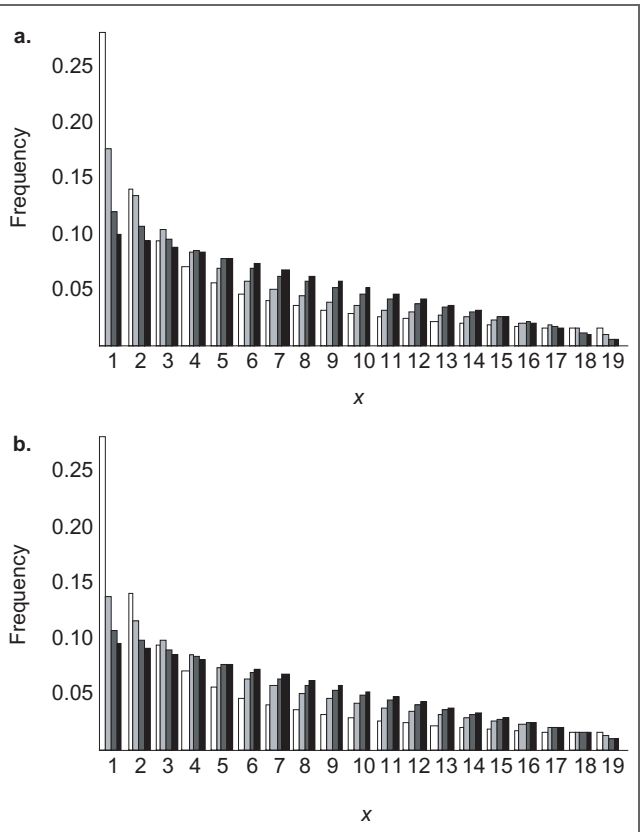

Figure 1. The expected frequency of mutant alleles (unfolded frequency spectrum) of allele frequency X in the standard neutral model assuming a sample of size $n = 20$ chromosomes and an ascertainment sample size of $d = 2$ (black), $d = 5$ (dark grey), $d = 10$ (light grey) and $d = 20$ (white; no ascertainment bias). In (a) it is assumed that the ascertainment sample is a subset of the typed sample and in (b) it is assumed that it is not.

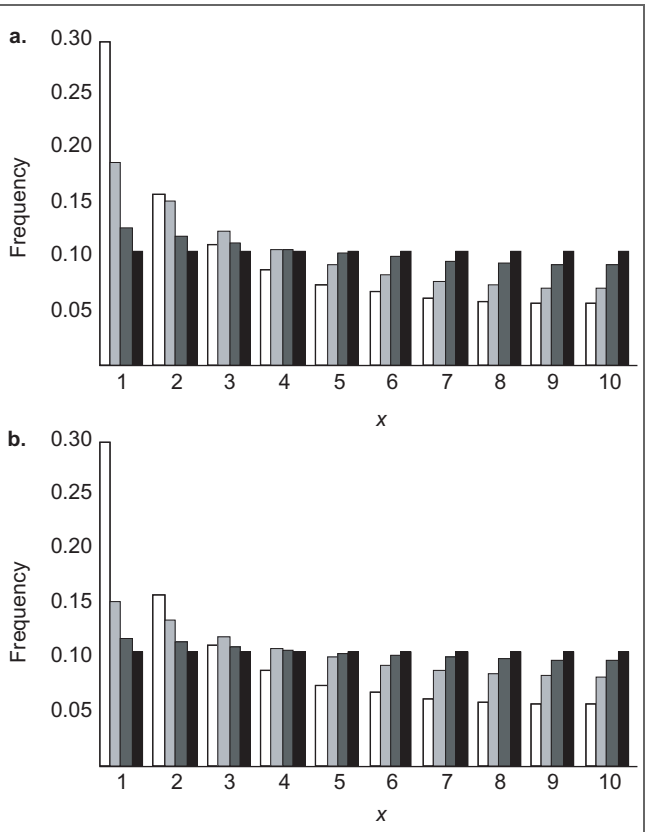

Figure 2. The expected folded frequency spectrum in the standard neutral model assuming a sample of size $n = 20$ chromosomes and an ascertainment sample size of $d = 2$ (black), $d = 5$ (dark grey), $d = 10$ (light grey) and $d = 20$ (white; no ascertainment bias). In (a) it is assumed that the ascertainment sample is a subset of the typed sample and in (b) it is assumed that it is not.

the expected frequency spectrum in a sample of size n is given by:[11,28,29]

$$\Pr(X=x)=\frac{\sum_{j=1}^{d-1}\Pr(X=x, Y=j|Z=x+j)/(x+j)}{\sum_{i=1}^{n-1}\sum_{j=1}^{d-1}\Pr(X=i, Y=j|Z=i+j)/(i+j)}, \quad 1<x<n-1, \tag{3}$$

where

$$\Pr(X=x, Y=j|Z=x+j)=\frac{\binom{x+j}{j}\binom{n+d-x-j}{d-j}}{\binom{n+d}{d}},$$

and Y and Z are the number of mutant alleles in the ascertainment sample and the pooled sample, respectively.

Figures 1b and 2b show the unfolded and folded frequency spectrum, respectively, in a sample of size $n = 20$. Notice that the effect of the ascertainment bias is even stronger than in the case where the ascertainment sample sequences were a subset of the typed sample sequences.

Effect on inferences of demographic parameters

With the exception of population growth parameters, the effect of the ascertainment bias on inferences regarding demographic parameters has not been extensively analysed in the literature.[29] Population growth has the effect of skewing the frequency spectrum towards an excess of rare alleles.[31] Because the effect on the frequency spectrum of most ascertainment schemes is in the opposite direction — towards an increase in the number of intermediate frequency alleles — the effect of

the ascertainment bias will be to reduce or eliminate the evidence for population growth. For example, Nielsen[23] found that there was little or no evidence for population growth in a dataset of 39 SNP loci. The lack of evidence for population growth was probably caused by the effects of the ascertainment scheme originally used to discover the SNPs.[29] Polanski and Kimmel found that estimates of population growth rates are extremely sensitive to the exact details of the ascertainment scheme.[29] They noted that even if just a small number of SNPs with low frequency polymorphisms have been eliminated due to presumed sequencing errors, this can substantially alter estimates of population growth rates.

The ascertainment scheme also has a profound effect on inferences regarding population structure. Wakeley *et al.*[11] showed that under a model of human demographics, the effect of an ascertainment bias would be to overestimate the rate of migration between populations. The complex ascertainment scheme considered by Wakeley *et al.* would preferentially select SNPs in genomic fragments with very old coalescent times. Because of the older coalescent times, these fragments of the genome have had an increased opportunity for migration in their genealogical history than fragments with very recent coalescent times, leading to ascertainment bias towards estimates of lower population subdivision.

If only one population is represented in the ascertainment sample, the effective population size of this population will be overestimated relative to the population size of other populations included in the typed sample. This is an issue that has been explored extensively for restriction fragment length polymorphism (RFLP) data.[32–35] Most of the available human RFLP data are based on polymorphisms that were originally identified in European populations. Initial analysis of these data led to the conclusion that the effective population size of Europeans is as large, or larger, than the effective population size of Africans. Most other data, however, such as mitochondrial DNA data, have shown that the effective population size of Africans in fact is much larger than the effective population size of Europeans.[32,36] Once discovered, the higher African heterozygosity, which had been obscured by ascertainment bias, became an important feature of 'out of Africa' theories. To date, this is probably the best practical example of how ascertainment bias may lead to erroneous conclusions in human genetics.

Figure 3 shows the expected unfolded frequency spectra from two populations simulated under different ascertainment schemes, assuming a standard coalescent model with migration,[37] with the number of migrants exchanged between the populations per generation set to $M = 1$.

In all cases, it is assumed that the ascertainment sample is a subset of the typed sample and that a locus is included in the analysis if there is variability in the ascertainment sample pooled from the two populations. Notice (see Figure 3a) that the frequency spectrum, when the typed sample is used as the ascertainment sample, is not very different from the frequency spectrum expected from a panmictic population, although there is a slight shift towards fewer ancestral alleles of low frequency. Because the ascertainment scheme is based on variability in any of the populations, one of the populations may now be invariable ($X = 0$ or $X = 20$), while the other population is variable. In this case, the expected value of the

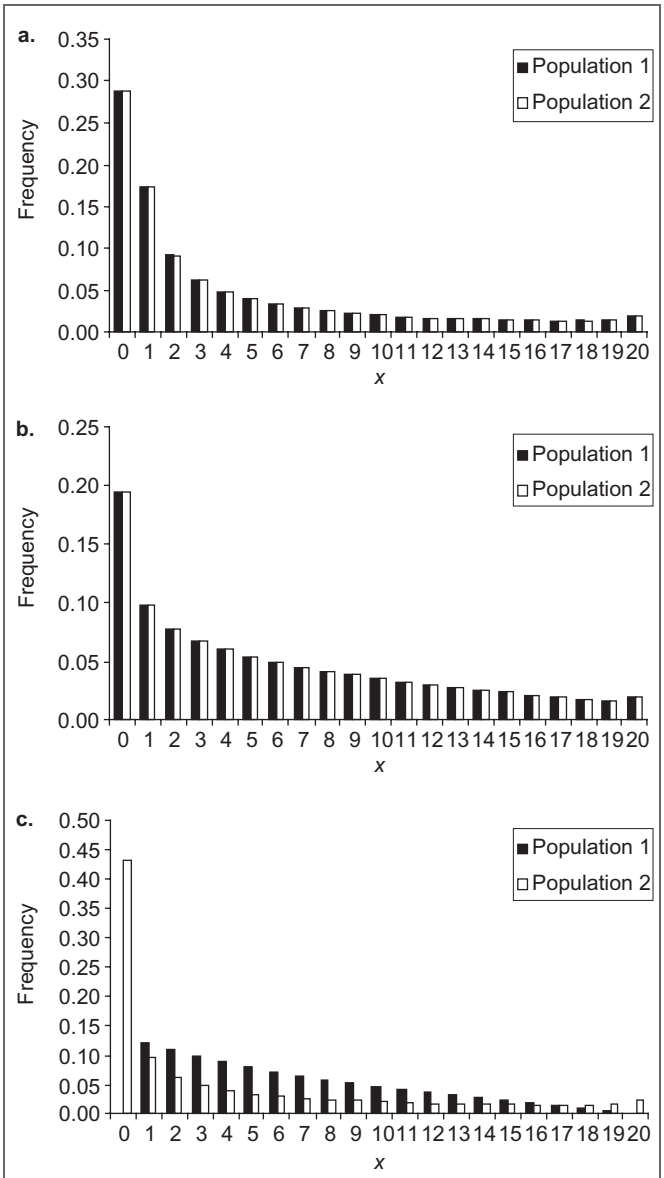

Figure 3. The expected frequency of mutant alleles (unfolded frequency spectrum) of allele frequency X in a neutral model of two populations exchanging $M = 1$ migrants per generation, assuming: (a) a sample of size $n = 20$ chromosomes and no ascertainment bias; (b) an ascertainment sample of two chromosomes from each population; and (c) an ascertainment sample of four chromosomes from Population 1.

popular measure of population subdivision, F_{ST}, is approximately 0.12.

If the ascertainment sample consists of two chromosomes from each population (Figure 3b), the frequency spectrum is further skewed towards including more high frequency ancestral alleles. In this case, the expected value of F_{ST} is approximately 0.17. The ascertainment scheme has increased the level of expected heterozygosity both within and between populations, but the combined effect is to increase the value of F_{ST} in this case. If ascertainment is based on a sample from only one population, the ascertainment population (Figure 3c), the frequency spectra of the two populations will differ because invariable loci may not exist in the population from which the data have been ascertained. The expected value of F_{ST} is now 0.13. Furthermore, among the variable site patterns, the ascertainment population has relatively more mutant alleles of intermediate frequency. The expected heterozygosity will be higher in the ascertainment population than in the other population.

Effect on linkage disequilibrium

The effect of the SNP discovery protocol on measures of linkage disequilibrium (LD) was examined by Nielsen and Signorovitch[28] and Clark *et al.*[10] The effect depends on the measure of LD and the exact details of the ascertainment protocol. For a protocol in which the ascertainment sample is a subset of the typed sample and identical for all loci, the standardised linkage disequilibrium coefficient, D', will be underestimated.[28,38] Another measure of LD, the squared allele frequency correlation coefficient, r^2, however, will be overestimated in the presence of this type of ascertainment bias.[28] In both cases, the effect is quite strong. For example, if the population recombination rate ($\rho = 2N_eR$, where N_e = effective population size and R = recombination rate) equals 1, r^2 is increased 2.5 times if $n = 100$ and $d = 5$. Akey *et al.*[38] also showed that in the case of population subdivision, where only one or a subset of a population are represented in the ascertainment sample, the ascertainment bias may be even more pronounced for the populations not represented in the ascertainment sample.

By contrast, the ascertainment protocol has much less effect on Hudson's composite likelihood estimator of ρ.[39] Typically, the bias in the estimate is less than 20 per cent and can be almost negligible, depending on the exact details of the ascertainment scheme.[28] In general, if ascertainment for all loci is based on the same set of chromosomes, the ascertainment bias will be towards lower values of ρ.

Correcting ascertainment bias

It should by now be clear that appropriate population genetic analysis of ascertained SNP data is problematic in the absence of methods for correcting for ascertainment bias. Fortunately, it is possible in many cases to correct the ascertainment bias relatively easily, if reliable information is available regarding the details of the ascertainment scheme.

The true frequency spectrum can be estimated from the observed frequency spectrum in an ascertainment biased sample. For example, consider the ascertainment scheme modelled in Equation (2). For this ascertainment scheme, the maximum likelihood estimates ($\hat{p}_k$) of the probabilities $\Pr(X = k)$ are given by:

$$\hat{p}_k = \frac{n_k}{\Pr(Ascertainment|X=k)}\left[\sum_{j=1}^{n-1}\frac{n_j}{\Pr(Ascertainment|X=j)}\right]^{-1}, \quad k = 1,\ldots,n-1 \tag{4}$$

where n_j is the observed number of loci with allele frequency j.

Figure 4 shows an example of uncorrected and corrected estimates of the frequency spectrum for a simulated dataset. Although analytical formulae such as Equation (4) may not be obtainable for all possible ascertainment schemes, it will in general be possible to obtain maximum likelihood estimates of the sample allele frequencies (the frequency spectrum). Such estimates may be useful for exploratory data analysis and to correct various parameter estimates based on the frequency spectrum. Correcting the frequency spectrum, however, is, in many cases, not the optimum method for correcting population genetic estimates based on ascertained SNP data, because this complicates the construction of valid confidence

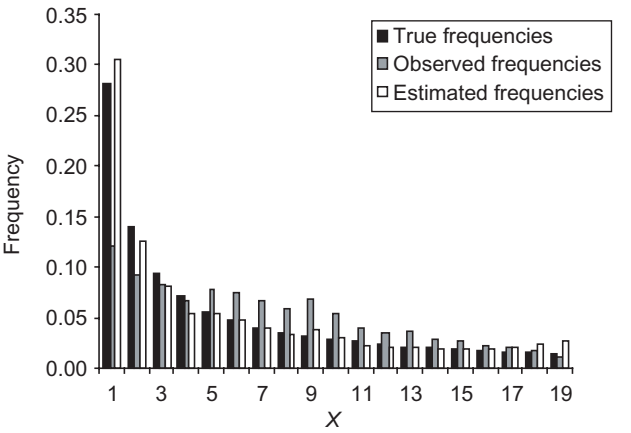

Figure 4. The observed frequency (grey) of a mutant allele in a sample of 1,000 SNPs simulated under the standard neutral model with the ascertainment scheme of Equation 2 and assuming $n = 20$ and $d = 4$. The black bars show the corresponding expected frequencies assuming no ascertainment bias and the white bars show the estimated frequencies using the correction given in Equation 4.

intervals or other measures of statistical uncertainty. In many cases, it might be preferable to correct the estimators directly. Such direct methods for correcting estimators are often mathematically easy to devise, especially when the estimators are based on likelihood functions. Correcting for ascertainment bias then simply becomes a question of defining the correct likelihood function. The likelihood function can be modified by conditioning on ascertainment, ie the corrected likelihood function should be defined as:

$$\Pr(Data|\Theta;\ Ascertainment) = \frac{\Pr(Data,\ Ascertainment|\Theta)}{\Pr(Ascertainment|\Theta)}, \quad (5)$$

where Θ is the vector of parameters. For example, in the case where the ascertainment sample is a subset of the typed sample, the corrected likelihood function calculated for a single locus with X mutant gene copies will be:[28,29]

$$\Pr(X = x|\Theta;\ Ascertainment) = \frac{\Pr(Ascertainment|X = x)\Pr(X = x|\Theta)}{\sum_{i=1}^{n-1}\Pr(Ascertainment|X = i)\Pr(X = i|\Theta)}, \quad 1 \le x \le n-1. \quad (6)$$

Such an approach has been used to correct estimates of ρ based on Hudson's (2001) estimator.[28] Application of this procedure produces approximately unbiased estimates of ρ. Polanski and Kimmel[29] used this method to estimate population growth rates from SNP data. They showed that, under a model of exponential population growth, as the effect of the ascertainment bias increases, the power to reject the hypothesis of no population growth decreases, even after correction of the ascertainment bias. In this case, an SNP discovery protocol in which loci with high frequency alleles have been chosen preferentially has caused a reduction in power compared with randomly sampled SNPs. The reason is that most of the information regarding population growth comes from rare alleles. In general, using ascertainment protocols that enrich the data with respect to common alleles can lead both to a decrease and an increase in power, depending on the specifics of the models and parameters being estimated.

The methods for correcting the likelihood function and estimating the frequency spectrum can easily be extended to the case where low frequency alleles have been eliminated directly[29] and to more complicated ascertainment schemes involving linked SNPs.[11] Wakeley *et al.*[11] considered data in which SNPs have been ascertained on the basis of the number of SNPs occurring on the genomic fragment on which they are located. They used methods similar to Equation (5) to estimate population growth rates and migration rates between human populations.

In theory, if detailed records regarding the SNP discovery protocols are being kept, corrections of the ascertainment bias are always possible. Even in the case where some information regarding the ascertainment scheme has been lost, such as the allele frequencies in the ascertainment samples, it may be possible to recover approximately unbiased parameter estimates and valid confidence intervals by statistical modelling of the ascertainment process. In cases where there is little or no overlap between the ethnicities of the individuals included in the typed sample and the ascertainment samples, however, corrections can only be made in parametric models describing the genetic relationship between the populations. In such cases, it will typically be difficult or impossible to use classical non-parametric methods for statistical inference.

Conclusions and recommendations

The SNP discovery protocol has a clear and pronounced effect on almost any population genetic inferences. Estimation of population genetic parameters based on information in the frequency spectrum is particularly sensitive to the applied ascertainment scheme. Even the elimination of relatively few SNPs with rare alleles, due to presumed sequencing errors, can have a pronounced effect on estimates of parameters such as the population growth rate.[29] If the exact protocol used for ascertainment is known, however, appropriate corrections can be performed. The information needed includes:

1. The size of the panel and the ethnicity of the panel members.
2. The details of the protocol used to sample sequences from the panel, ie full sampling or sampling with or without replacement; the ascertainment sample sizes; and, for linked SNPs, information regarding independent or correlated sampling of SNPs in the ascertainment sample.
3. Details regarding base-calling and elimination of rare alleles.

In many cases, this information is available or can be reconstructed. If not, this will in most cases preclude valid population genetic inferences based on the SNP data.

Much work is still needed on SNP ascertainment bias corrections. For example, there is still a need for standard methods for estimating levels of population subdivision from SNP data corresponding to the classical F_{ST} estimator.[40] Such an estimator would be useful, for example, in studies aimed at detecting (possibly selected) genomic regions with extreme F_{ST} values.[14] Researchers may also find corrections to estimates of the linkage disequilibrium coefficient (D') and its derivatives useful.

The large SNP datasets provide an unrivalled population genetic resource that, most likely, for many years will not be rivalled by data obtained using direct sequencing. Much effort is being devoted in the human genetics and population genetics communities to estimate ancestral and demographic parameters and parameters relating to recombination and mutation from human population genetic data. There is no reason why the large SNP datasets should not be used in

this effort. Before this can happen, however, details regarding ascertainment schemes must be publicly available in greater detail. For example, information regarding the ethnicity of a panel is not sufficient without detailed information regarding how ascertainment samples were constructed from the panel when not all panel members have been sequenced for each SNP. Information regarding base-calling, used to assess the probability of unintentionally eliminating a low frequency allele, should also be available. Making this type of information available in databases, in a fashion that facilitates proper statistical analysis, provides a major bioinformatics task that should be given a very high priority by the human population genetics community.

References

1. Smigielski, E.M., Sirotkin, K., Minghong, W. *et al.* (2000), 'dbSNP: A database of single nucleotide polymorphisms', *Nucl. Acids Res.* Vol. 28, pp. 352–355.
2. Collins, F.S., Guyer, M.S. and Chakravarti, A. (1997), 'Variations on a theme: Cataloging human DNA sequence variation', *Science* Vol. 278, pp. 1580–1581.
3. Collins, F.S., Brooks, L.D. and Chakravarti, A. (1998), 'A DNA polymorphism discovery resource for research on human genetic variation', *Genome Res.* Vol. 8, pp. 1229–1231.
4. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* Vol. 273, pp. 1516–1517.
5. Brookes, A.J. (1999), 'The essence of SNPs', *Gene* Vol. 234, pp. 177–186.
6. Kruglyak, L. (1999), 'Prospects for whole-genome linkage disequilibrium mapping of common disease genes', *Nat. Genet.* Vol. 22, pp. 139–144.
7. Reich, D.E., Cargill, M., Bolk, S. *et al.* (2001), 'Linkage disequilibrium in the human genome', *Nature* Vol. 411, pp. 199–204.
8. Dawson, E., Abecasis, G.R., Bumpstead, S. *et al.* (2002), 'A first-generation linkage disequilibrium map of human chromosome 22', *Nature* Vol. 418, pp. 544–548.
9. Gabriel, S.B., Schaffner, S.F., Nguyen, H. *et al.* (2002), 'The structure of haplotype blocks in the human genome', *Science* Vol. 296, pp. 2225–2229.
10. Clark, A.G., Nielsen, R., Signorovitch, J. *et al.* (2003), 'Linkage disequilibrium and inference of ancestral recombination in 538 single nucleotide polymorphism clusters across the human genome', *Am. J. Hum. Genet.* Vol. 73, pp. 285–300.
11. Wakeley, J., Nielsen, R., Liu-Cordero, S.N. *et al.* (2001), 'The discovery of single nucleotide polymorphisms—and inferences about human demographic history', *Am. J. Hum. Genet.* Vol. 69, pp. 1332–1347.
12. Cavalli-Sforza, L.L. and Feldman, M.W. (2003), 'The application of molecular genetic approaches to the study of human evolution', *Nat. Genet.* Vol. 33, pp. 266–275.
13. Sunyaev, S.R., Lathe III, W.C., Ramensky, V.E. *et al.* (2000), 'SNP frequencies in human genes: An excess of rare alleles and differing modes of selection', *Trends Genet.* Vol. 16, pp. 335–337.
14. Akey, J.M., Zhang, G., Zhang, K. *et al.* (2002), 'Interrogating a high-density SNP map for signatures of natural selection', *Genome Res.* Vol. 12, pp. 1805–1814.
15. Sabeti, P.C., Reich, D.E., Higgins, J.M. *et al.* (2002), 'Detecting recent positive selection in the human genome from haplotype structure', *Nature* Vol. 419, pp. 832–837.
16. Reich, D.E., Cargill, M., Bolk, S. *et al.* (2001), 'Linkage disequilibrium in the human genome', *Nature* Vol. 411, pp. 199–204.
17. Sunyaev, S.R., Ramensky, V.E., Koch, I. *et al.* (2001), 'Prediction of deleterious human alleles', *Hum. Mol. Genet.* Vol. 10, pp. 591–597.
18. Taillon-Miller, P., Gu, Z., Li, Q. *et al.* (1998), 'Overlaping genomic sequences: a treasure trove of single-nucleotide polymorphisms', *Genome Res.* Vol. 8, pp. 748–754.
19. Wang, D.G., Fan, J.B., Siao, C.J. *et al.* (1998), 'Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome', *Science* Vol. 280, pp. 1077–1082.
20. Picoult-Newberg, L., Ideker, T.E., Pohl, M.G. *et al.* (1999), 'Mining SNPs from EST databases', *Genome Res.* Vol. 9, pp. 167–174.
21. Altshuler, D., Pollar, V.J., Cowles, C.R. *et al.* (2000), 'A SNP map of the human genome generated by reduced representation shotgun sequencing', *Nature* Vol. 407, pp. 513–516.
22. Eberle, M.A. and Kruglyak, L. (2000), 'An analysis of strategies for discovery of single nucleotide polymorphisms', *Genet. Epidemiol.* Vol. 19, pp. S29–S35.
23. Nielsen, R. (2000), 'Estimation of population parameters and recombination rates using single nucleotide polymorphisms', *Genetics* Vol. 154, pp. 931–942.
24. Kuhner, M.K., Beerli, P., Yamamoto, J. *et al.* (2000), 'Usefulness of single nucleotide polymorphism data for estimating population parameters', *Genetics* Vol. 156, pp. 439–447.
25. Kingman, J.F.C. (1982), 'The coalescent', *Stochast. Proc. Appl.* Vol. 13, pp. 235–248.
26. Hudson, R.R. (1983), 'Testing the constant-rate neutral model with protein sequence data', *Evolution* Vol. 37, pp. 203–217.
27. Tajima, F. (1989), 'Statistical method for testing the neutral mutation hypothesis by DNA polymorphism', *Genetics* Vol. 123, pp. 585–595.
28. Nielsen, R. and Signorovitch, J. (2003), 'Correcting for ascertainment biases when analyzing SNP data: Applications to the estimation of linkage disequilibrium', *Theor. Pop. Biol.* Vol. 63, pp. 245–255.
29. Polanski, A. and Kimmel, M. (2003), 'New explicit expressions for relative frequencies of single nucleotide polymorphisms with application to statistical inference on population growth', *Genetics* Vol. 165, pp. 427–436.
30. Sherry, S.T., Harpending, H.C., Batzer, M.A. *et al.* (1997), '*Alu* evolution in human populations: Using the coalescent to estimate effective population size', *Genetics* Vol. 147, pp. 1977–1982.
31. Slatkin, M. and Hudson, R.R. (1991), 'Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations', *Genetics* Vol. 129, pp. 555–562.
32. Mountain, J.L. and Cavalli-Sforza, L.L. (1994), 'Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms', *Proc. Natl. Acad. Sci. USA* Vol. 91, pp. 6515–6519.
33. Rogers, A.R. and Jorde, L.B. (1996), 'Ascertainment bias in estimates of average heterozygosity', *Am. J. Hum. Genet.* Vol. 58, pp. 1033–1041.
34. Urbanek, M., Goldman, D. and Long, J.C. (1996), 'The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: A new approach to measuring gene identity reveals asymmetric patterns of divergence', *Mol. Biol. Evol.* Vol. 13, pp. 943–953.
35. Eller, E. (2001), 'Effects of ascertainment bias on recovering human demographic history', *Hum. Biol.* Vol. 73, pp. 411–428.
36. Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J. *et al.* (1994), 'High resolution of human evolutionary trees with polymorphic microsatellites despite a constraint in allele length', *Nature* Vol. 368, pp. 455–457.
37. Hudson, R.R. (1990), 'Gene genealogies and the coalescent process', in P.H. Harvey and L Partridge (eds.), *Oxford Surveys in Evolutionary Biology*, Vol. 7, Oxford University Press, New York, NY, pp. 1–44.
38. Akey, J.M., Zhang, K., Xiong, M. *et al.* (2003), 'The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium', *Mol. Biol. Evol.* Vol. 20, pp. 232–242.
39. Hudson, R.R. (2001), 'Two-locus sampling distributions and their application', *Genetics* Vol. 159, pp. 1805–1817.
40. Weir, B.S. and Cockerham, C.C. (1984), 'Estimating F-statistics for the analysis of population structure', *Evolution* Vol. 38, pp. 1358–1370.