Case-control studies of association in structured or admixed populations

Jonathan K. Pritchard
Peter Donnelly [1]
Department of Statistics
University of Oxford

July 5, 2001

[1]Address for Correspondence: Department of Statistics, University of Oxford, 1 South Parks Rd, Oxford, OX1-3TG, UK. Tel: 44 1865 282852; Fax: 44 1865 272595; Email: pritch@stats.ox.ac.uk, donnelly@stats.ox.ac.uk.

Abstract

Case-control tests for association are an important tool for mapping complex-trait genes. But population structure can invalidate this approach, leading to apparent associations at markers that are unlinked to disease loci. Family-based tests of association can avoid this problem, but such studies are often more expensive, and in some cases—particularly for late-onset diseases—are impractical. In this review article we describe a series of approaches published over the last two years which use multilocus genotype data to enable valid case-control tests of association, even in the presence of population structure. These tests can be classified into two categories. "Genomic control" (GC) methods use the independent marker loci to adjust the distribution of a standard test statistic, while "structured association" (SA) methods infer the details of population structure *en route* to testing for association. We discuss the statistical issues involved in the different approaches, and present results from simulations comparing the relative performance of the methods under a range of models.

Introduction

Association mapping is potentially a very powerful strategy for identifying the loci that contribute to complex diseases (Risch and Merikangas, 1996; Risch, 2000). If a mutation increases disease susceptibility, then we can expect it to be more frequent among affected individuals (cases) than among unaffected individuals (controls). The essential idea behind association mapping is that markers close to the disease mutation may also have allele frequency differences between cases and controls if there is linkage disequilibrium between the marker locus and the susceptibility mutation.

Thus, the simplest form of association mapping involves genotyping a set of markers in a sample of cases and in a sample of unrelated controls, and then testing for allele frequency differences at each marker. For biallelic markers, this is often done using 2x2 tests of association between the alleles and the phenotypes. In a random-mating population, linkage disequilibrium decays quite rapidly with distance, and so if one finds association at a marker locus this implies that the marker is tightly linked to a disease susceptibility mutation.

Unfortunately, this approach may be invalid in the presence of population structure.[1] In structured populations there can be a high rate of significant associations even at markers that are unlinked to any disease loci (e.g., Knowler *et al.*, 1988; Lander and Schork, 1994; Ewens and Spielman, 1995; Risch, 2000). The problem arises because both disease frequencies and allele frequencies can differ among subpopulations. Suppose that a sample of cases and controls are drawn from a population containing a number of subpopulations. If the disease of interest is at high frequency in one subpopulation, then we can expect to find that group over-represented among the cases. Then any marker allele that is at higher frequency in that subpopulation than in the others will appear to be associated with the disease, regardless of where it is in the genome (e.g., Pritchard and Rosenberg, 1999). A similar effect arises in admixed populations, if the risk of disease depends

[1] Here, population structure refers to the presence of subgroups of individuals in the population, for instance different ethnic groups, who differ systematically across loci in their allele frequencies.

on the degree of admixture (Knowler *et al.*, 1988).

One way of tackling this problem is to collect data about ethnicity from the members of the sample, and then stratify the analysis according to reported ethnicity. This sort of approach seems sensible. At the very least, it seems crucial for association studies to collect as much information about ethnicity as possible, to try to detect any mismatching of the cases and controls. However, the use of reported ethnic data may not solve the problem.

First, it is possible that culturally-defined ethnic groups do not adequately reflect the underlying population structure. For example within admixed groups, individuals may differ widely in the proportion of their ancestry from different sources (as, for example, with differing degrees of European ancestry in African Americans; Pfaff *et al.*, 2001). It is also possible that there may be "cryptic" population structure—i.e., population structure at the genetic level that is not well described using ethnic, or national labels (Pritchard *et al.*, 2000a). These concerns mean that many people in the field are sceptical about associations detected by case-control studies.

In response to the problem, the 1990's saw the development of a series of family-based tests of association. Unlike the standard case-control approach, these tests are constructed in such a way that they do not suffer from the problem of false positives due to population structure. That is, they reject the null hypothesis at the nominal rate unless there is *both* association and linkage. The best known of these tests, the "Transmission-Disequilibrium Test" (TDT) uses genetic data from affected individuals and their parents (Spielman *et al.*, 1993; Spielman and Ewens, 1996); other related tests make use of siblings or other relatives (e.g., Boehnke and Langefeld, 1998; Lazzeroni and Lange, 1998; Spielman and Ewens, 1998).

By avoiding the population structure problem, the family-based tests have had a major impact on the use of association studies in genetics. However, in many contexts there are important practical advantages to case-control approaches. It can be much more time-consuming and expensive to assemble family-based samples, and for late-onset diseases it may be impractical. Moreover, family-based studies require more genotyping: for instance the standard TDT requires genotypes from the affected offspring plus two

parents for the rough equivalent of one case and one control. Finally, if genome screens for association become viable (Risch and Merikangas, 1996), there is the possibility that databases of control genotypes could be established to allow more efficient case-control studies.

For all these reasons, case-control studies would often be an attractive study design for association mapping, were it not for the problem of false positives due to population structure. In response, there has been a series of papers in the last two years on methods that aim to construct valid case-control tests of association. We review these here. All of these methods use the idea that demographic effects such as population structure will affect loci across the genome similarly, whereas associations due to linkage disequilibrium with a disease marker will only be seen over short scales. This means that by using multilocus data, we can hope to detect and correct for the effects of population structure.

The impact of population structure in practice

Surprisingly perhaps, in view of the level of concern about false positives due to population structure, there are few clear examples of studies in which population structure has led to "spurious associations" (Thomas and Witte, 2001). In part, this probably reflects the fact that until recently, there has been no direct way of assessing the impact of population structure on association studies; however, there are now several statistical methods, reviewed below, for detecting the effects of population structure in this context.

Probably the best-known example in which population structure has impacted an association study is from a study of non-insulin-dependent diabetes in the Pima and Papago tribes (Native Americans), who suffer from an extremely high rate of diabetes (Knowler *et al.*, 1988). The data indicated a strong negative association between diabetes and a haplotype at the immunoglobulin G locus. However, many of the sampled individuals had recent European ancestry, and it was found that the average proportion of European ancestry in the controls was higher than in affected individuals. The haplotype in question was shown to be at much higher frequency

in Europeans in general, regardless of phenotype, and the authors showed that the protective effect of this haplotype disappeared if the analysis was stratified according to reported ancestry.

Another example of the potential impact of population structure has been observed in a recent study in the northwest of England, a region that has extensive immigration of Irish, Scots and Welsh. Examination of HLA-DPB1 genotype data in a birth cohort of healthy caucasian controls suggested that allele frequencies varied according to social class defined by parental occupation (G.M. Taylor, personal communication). A possible explanation for this result is that there are subtle differences in the ethnic makeup of social strata in this region. In this example, it seems that mismatching of cases and controls by social class could lead to invalid tests of association.

In general it is the case that there are frequently difficulties in replicating reported associations (Terwilliger and Weiss, 1998), but at this time it is difficult to know how much of this is due to the effects of population structure, and how much is simply due to a publication bias in favor of reporting significant results (Terwilliger and Weiss, 1998; Cardon and Bell, 2001).

Devlin and Roeder (1999) provide a convenient framework for quantifying the impact of population structure on association studies of biallelic markers. Consider the situation of performing 2x2 tests of association between the alleles and the phenotypes. Under the null hypothesis of no association, the expected value of the standard Chi-square test statistic (or trend test (Devlin and Roeder, 1999)) is 1.0. Suppose that we collect a sample of R cases and R controls from a population consisting of K subpopulations. Let the fraction of the cases (controls) drawn from subpopulation k be f_k (g_k). Then, conditional on the realized values of the f_k and g_k, the null distribution of the trend test statistic is inflated by a multiplicative factor λ, which can be written

$$\lambda = 1 + \frac{RF \sum_k \left[(f_k - g_k)^2\right] - 2F}{1 + F}, \tag{1}$$

under certain modelling assumptions (modified from Eq. 5 of Devlin and Roeder, 1999). Here, F, also known as F_{ST}, denotes Wright's coefficient of

inbreeding (Wright, 1951). Then

$$\lambda \approx 1 + RF \sum_k (f_k - g_k)^2 \tag{2}$$

for large R and small F. Estimates based on classical polymorphisms suggest that plausible values for F are of the order of 0.01-0.05 for European populations or other closely related populations, and as much as 0.1–0.3 for the most divergent pairs of populations (summarized in Table 2.3.1A, Cavalli-Sforza *et al.*, 1994). It seems then, that for case-control studies where the population structure is relatively subtle–and hence easy to miss–we might expect values for F on the order of 0.01.

The value of $\sum(f-g)^2$ ranges from 0 to 2, and depends on the specifics of the sampling (c.f. Pritchard and Rosenberg, 1999). Suppose that the prior probability of sampling from each subpopulation is π_k, and the frequency of disease in subpopulation k is ϕ_k. Then the cases (controls) are sampled from subpopulation k at a rate proportional to $\pi_k\phi_k$ and $\pi_k(1-\phi_k)$ respectively. For example, suppose that $K = 2$, $\pi_1 = \pi_2 = 1/2$, and that the rate of disease is two-fold higher in subpopulation 1 ($\phi_1/\phi_2 = 2$). Then $\mathrm{E}(f_1) = 2/3$, $\mathrm{E}(f_2) = 1/3$, and $\mathrm{E}(g_1) = \mathrm{E}(g_2) \approx 0.5$ (if ϕ is small). If f_1 and g_1 take their expected values, $\sum_k (f_k - g_k)^2 = 1/18$.

For this hypothetical example, if we take $F = 0.01$, λ is quite small unless the sample sizes are reasonably large by current standards. But for a sample of 1000 cases and 1000 controls, $\lambda \approx 1.53$, which implies a substantial departure from the null model. Future searches for complex disease genes of small effect will almost certainly need to have samples of thousands of individuals to achieve moderate power, and the impact of population structure will become more serious as sample sizes increase.

Correcting for the effects of population structure in case-control studies

Suppose that we suspect a particular chromosomal region may contain a disease mutation, and indeed, using a case-control sample, we find strong

associations with markers in this region. We want to know whether this is strong evidence for a nearby disease mutation, or whether the result could be due to population structure. Similarly, if we perform a genome scan for association, we want to be able to assess the significance levels of the strongest associations.

Demographic factors such as population structure are expected to have a similar affect on all loci across the genome. For this reason, if there are problems with a particular case-control sample, we can expect that there will be a high rate of significant associations at many markers across the genome (Pritchard and Rosenberg, 1999; Devlin and Roeder, 1999). This suggests the strategy of using genotype data from a series of unlinked markers to assess the impact of population structure on the sample.[2]

Two kinds of approaches have been suggested to capture this type of information. In the presence of population structure, the usual chi-squared test statistic, X^2, may no longer have a χ^2 distribution. One line of attack is to keep this original test statistic but to use the data at unlinked markers to ensure that its observed value is compared against the correct distribution. In the first approach in this direction, Pritchard and Rosenberg (1999) used the unlinked markers to check whether the level of structuring was small enough that the usual statistic still had a χ^2 distribution. Devlin and Roeder (1999) and Reich and Goldstein (2001) took this one step further by using the unlinked markers to estimate the null distribution of the usual test statistic. One natural approach (Reich and Goldstein, 2001) is to estimate this null distribution from the empirical distribution of the chi-squared statistics calculated at each of the unlinked markers. However, as those authors noted, this is not very efficient. A parametric approach was obtained

[2]For convenience, we refer to these as "unlinked" markers. More precisely, we require that they should be far enough apart that an individual's genotypes at each locus are conditionally independent, given his/her ancestry. When studying discrete subpopulations, we merely need the markers to be sufficiently far apart that they are in linkage equilibrium within subpopulations; in such situations, we can often treat markers separated by as little as 1 cM or so as independent (Pritchard and Przeworski, 2001). In admixed populations, where the admixture creates extra correlations between loci, the spacing may need to be greater.

by Devlin and Roeder (1999), who argued theoretically, and by Reich and Goldstein (2001) who used simulations, to show that under a range of assumptions the effect of structure is simply to rescale the null distribution of X^2 by the multiplicative factor λ defined in equation 1. When this is the case, the unlinked marker data can be used to provide an estimate $\hat{\lambda}$ of λ, and the value of $X^2/\hat{\lambda}$ can be compared against the usual χ^2 distribution[3]. Devlin and Roeder (1999) coined the expression "genomic control" (GC) for their method. We will use the term to refer to this general approach. These methods are reviewed in more detail by Devlin *et al.* (2001).

In contrast, the second approach uses the multilocus genotype data to infer the details of the population structure. The tests of association take account of the inferred structure (Pritchard *et al.* (2000a,b); Satten *et al.* (2001); see also related work by Ripatti *et al.* (2001) and Sillanpää *et al.* (2001)). We describe this class of methods as "structured association" (SA) methods.

Pritchard *et al.* (2000a) assume a model in which the sample contains a mixture of individuals drawn from K subpopulations, where K might be unknown. In order to allow for the possibility of admixture, individuals are permitted to have some fraction of their ancestry in each of the K subpopulations. Each subpopulation is characterized by a set of allele frequencies at each marker locus; these are typically unknown in advance. Since the sample contains a mixture of individuals from different subpopulations, there can be departures from Hardy-Weinberg and linkage equilibrium in the overall sample. However, it is assumed that *within* subpopulations there is Hardy-Weinberg equilibrium, and linkage equilibrium between all markers. Inference is performed within a Bayesian framework; an MCMC (Markov chain Monte Carlo) algorithm is used to sample from the joint posterior distribution of the subpopulation allele frequencies, and the ancestries of individuals. Loosely speaking, the algorithm seeks to identify subpopulations

[3]Notice that in a particular study the value of λ will depend, amongst other things, on the realized values of the f_k and g_k, defined above. That is, replicate samples of the same numbers of cases and controls from the same population, would, in general, lead to different values of λ. In correcting the distribution of the test statistic one should use the value of λ for the study at hand; this is the value of λ estimated by the marker data.

of individuals who are genetically similar. Use of the Bayesian framework makes it straightforward to incorporate any prior information, such as partial data on ethnicity for some individuals. Model choice for K is performed by running the Markov chain separately at different values of K; an approximate method is used to estimate posterior probabilities for each value of K.[4]

Pritchard *et al.* (2000b) then describe a method of testing for association that conditions on the inferred ancestries of individuals. The standard null hypothesis in association studies is that allele frequencies at the candidate locus do not depend on phenotype. In the presence of population structure, this is replaced with a null hypothesis that there is no such dependence *within subpopulations*. The alternative model proposed is quite general: namely that the allele frequencies depend on both phenotype and subpopulation (see Pritchard *et al.* (2000b) for formal definitions of the models). This alternative allows for the possibility that the alleles at the test locus may have different effects in different subpopulations (for example, the test locus might be in strong linkage disequilibrium with a disease mutation in some subpopulations, but not in others). A test statistic is constructed by computing the likelihood ratio under the two hypotheses, conditional on the inferred ancestries of individuals. P-values are obtained by simulation. In the case where there are many candidate loci, as in a genome scan, all the loci are used to infer population structure, and then each locus is tested in turn for association with the phenotype. These methods are implemented in a pair of programs, *structure* and *STRAT*, available from `www.stats.ox.ac.uk/mathgen/software.html`.

More recently, Satten *et al.* (2001) have developed a closely related method. Instead of inferring population structure within a Bayesian framework, Satten *et al.* (2001) use maximum likelihood, implemented via the EM algorithm. More importantly, their model does not allow for admixture between subpopulations. The number of populations, K, is chosen to maximise a penalized likelihood, where the penalty (based on the Akaike

[4]Alternatives such as reversible jump MCMC (Green, 1995) are difficult to implement in this context due to the very large number of parameters.

Information Criterion, AIC) increases with the number of parameters in the model. In the settings where there is enough information for the subpopulation membership to be estimated well, both the Bayesian and maximum likelihood methods would be expected to perform well. We note however, that in the large sample limit, the use of AIC may lead to systematic overestimation of the number of subpopulations K (e.g., Hurvich and Tsai, 1989). Unlike the other papers described above, which focus on significance testing, the primary aim of Satten *et al.* (2001) is to estimate the size of effect at the candidate locus. This is done under the assumption that the effect can be modeled with the same parameter in all subpopulations. Note that all of their simulations in the original paper were performed under the assumed model. In the Discussion, we return to a comparison of the methods of Pritchard *et al.* (2000b) and Satten *et al.* (2001).

The structured association approach was recently used by Thornsberry *et al.* (2001) who were studying *Dwarf8*, a candidate gene for the control of flowering time in maize. The authors collected flowering-time data and sequence data from 92 inbred lines. It was clear from the outset that population structure could be a major confounder. The sample was geographically diverse, including both tropical and north American lines representing several major subgroups of maize. Moreover, it is quite likely that the trait of interest, flowering time, is under divergent selection in different parts of the geographic range. Indeed, it was found that there was a high rate of significant associations at random microsatellite loci, indicating serious confounding with population structure (c.f. Pritchard and Rosenberg, 1999).

Thornsberry *et al.* applied the structured association approach of Pritchard *et al.* (2000b) to tackle this problem, using a modified test statistic to allow for quantitative variation. Analysed in this way, the rejection rate at the microsatellite markers was close to the nominal value, though still slightly high (8% at the 5% level). Some of this excess may be due to "real" associations with quantitative trait loci; it may also be that the sample of 92 inbred lines is not large enough to fully capture the details of maize population structure. The authors reported two sets of p-values: (1) the raw p-values obtained from the structured association test, and (2) p-values

based on the empirical distribution of structured association p-values at the microsatellite loci. They argued that the true values probably lie somewhere in between. The study reports that significant associations at the *Dwarf8* locus were replicated across five field studies.

Comparison of methods

Theory. To understand the differences between the various approaches, it may help to think of association mapping in the presence of population structure as a missing data problem. To simplify the discussion we ignore the existence of admixed individuals, but we note that an important advantage of the approach of Pritchard *et al.* (2000b) is that it does allow for the possibility of admixture.

In a case-control study, the data we have available to us include the phenotypic information and the genotype at the candidate locus for each individual. In an ideal world, we would also like to know how many subpopulations there are, and which sampled individuals belong to which subpopulations. If we had this missing information, we would construct a test for association by comparing allele (or genotype) frequencies between cases and controls *within* each subpopulation and then combine the results of these comparisons *across* subpopulations. (The "best" way to do this in the complete data setting will in general depend on assumptions about how susceptibility alleles, and their effect, vary across subpopulations.)

In the real world we do not have the information about subpopulation membership, but, in the settings considered here, we have the additional genotype data for each individual at unlinked loci across the genome. The statistical challenge is how best to use this information to construct tests which are (i) valid, and (ii) have good power against plausible alternatives. There is a fundamental dichotomy in the range of methods which have been developed.

As described above, the first approach (genomic control) makes use of the usual chi-squared statistic (or trend test, Devlin and Roeder, 1999). Informally, the effect of population stratification is often to inflate the null

distribution of the test statistic. The genomic control approach aims to use the distribution of values of the test statistic across unlinked markers to estimate the extent of this inflation. Corrected p-values at any given locus can then be obtained using an adjusted distribution that accounts for any inflation observed.

The second general approach is fundamentally different. Rather than stick with the usual statistic and correct its distribution, the approach uses the unlinked marker information to *estimate* the missing data. That is, the additional genotype information is used to estimate both the number of subpopulations and the subpopulation to which each sampled individual belongs. Having estimated the missing data, a test for association can then be constructed as it would be in the complete data setting. Within this broad framework, which we call "structured association" (SA), there are choices about how to estimate the missing data, and further choices about the construction of the test of association based on the complete data (e.g., Pritchard *et al.*, 2000b; Satten *et al.*, 2001).

We can now see the tradeoffs between the genomic control and structured association approaches. In a sense, the SA approach is more ambitious. It aims to get a lot more mileage out of the genotype data at unlinked markers, in estimating the subpopulation membership of each individual, rather than just the inflation factor λ. General statistical considerations suggest that when this missing information is estimated well, its use can result in better tests (more powerful against a wider range of alternatives) than comparing the standard test statistic to a corrected distribution. On the other hand, if the missing subpopulation information is not estimated well, subsequent tests may be invalid.

How well the missing information can be estimated will depend on (i) the estimation method used, (ii) the complexity of the underlying population structure, and the extent to which is captured by the model, and (iii) the amount of data (both number of individuals and number of independent loci) available for its estimation. As a general rule, if there is enough information for good estimation of the missing subpopulation data, the additional power and flexibility offered by SA approaches, such as STRAT, should make them

preferable to GC methods. As we move toward association studies based on hundreds or thousands of markers, this will increasingly become the case.

Simulations. We have performed a simulation study to compare the performance of the genomic control method of Bacanu *et al.* (2000) (a frequentist test based on Devlin and Roeder, 1999) and STRAT, the structured association method of Pritchard *et al.* (2000b). For the purpose of this study, we focused on the situation in which L marker loci are used to infer population structure, and then additional candidate loci are tested for association. An alternative model is that a single set of loci are used both to test for association, and to learn about structure (as, for example, in a genome screen for association). We would expect similar results for both cases: the proportion of markers that are in genuine association with the phenotype in a genome scan will be very small, and these should have little impact on the inferred structure.

We present results from the following model. Samples of 200 cases and 200 controls were drawn from a population consisting of three subpopulations. The prior probability of sampling individuals from each subpopulation was 1/3; the prevalence of the disease was two-fold higher in subpopulation 3. Thus, the expected numbers of cases and controls from each subpopulation were 50, 50, 100, and 66.7, 66.7, 66.7, respectively (the controls were assumed to be a random sample, rather than ascertained for *not* having the disease). The actual numbers drawn from each subpopulation were fixed at close to their expected values, namely, 50, 50, 100, and 66, 67, 67.

The subpopulation allele frequencies at biallelic marker loci were modeled using the Balding-Nichols (1995) model, as in Devlin and Roeder (1999): the allele frequency at locus l in population i was simulated from a Beta distribution with parameters $\{(F_i p_l/(1-F_i), F_i(1-p_l)/(1-F_i)\}$, where $i \in \{1,2,3\}$. The ancestral allele frequency p_l was drawn from a Uniform distribution in (0.1, 0.9). Wright's F for the three subpopulations was set to 0.01, 0.02, and 0.04 respectively.

We present examples where 50, 200, and 1000 biallellic loci were used either to estimate λ (GC), or to infer the details of the population structure

(STRAT). The various parameters were chosen to illustrate the performance of the methods both in situations where STRAT performs well, and in situations where it may struggle to infer the structure adequately. (The accuracy of inferences about population structure improves with sample size, number of loci, and the degree of divergence between populations (Pritchard *et al.*, 2000a).) These parameters correspond to a modest amount of structure: the inflation factor $\lambda \approx 1.24$ for the sample size given above, and the chi-square test of association is significant about 14% of the time at the 10% level, and significant 2% of the time at the 1% level.

Results. The program *structure* was used to estimate population structure for each of twenty data sets of 50, 200, and 1000 markers, simulated under the null hypothesis of no association[5]. With 50 loci, *structure* was unable to detect the presence of three distinct subpopulations. For four of the data sets, the posterior mode for K was 1; for the remaining sixteen $\hat{K} = 2$. In the latter cases, the population with $F = 0.04$ was separated from the other two. $\hat{K}$ was 3 for all twenty examples with 200 and 1000 loci. Figure 1 shows three typical examples of the accuracy of *structure* at inferring the ancestry of individuals. The performance is relatively poor with 50 loci, but with 200 loci most individuals are correctly assigned, and with 1000 loci, the assignments are essentially perfect.

In order to estimate the inflation factor λ, Bacanu *et al.* (2000) suggest using the median value of the trend statistic divided by 0.456. We simulated 10^4 samples of 50, 200, and 1000 loci under the model described above, to study the behavior of $\hat{\lambda}$. The sampling distributions of $\hat{\lambda}$ for these examples are plotted in Figure 2. Notice from the figure that even with many loci, there is considerable uncertainty in the estimates of λ. Large values of $\hat{\lambda}$ will lead to reduced power, while low values of $\hat{\lambda}$ lead to inappropriately high type 1 error rates. Bacanu *et al.* (2000) report that, in their simulations, values of $\hat{\lambda}$ below 1 were adjusted upward to $\hat{\lambda} = 1$; as noted by those authors, the latter adjustment causes the estimates to be biased upwards.

[5]We applied *structure* using the admixture model with correlated allele frequencies (Pritchard *et al.*, 2000a).

After applying this adjustment, the mean estimates of $\hat{\lambda}$ were 1.33, 1.25 and 1.24 respectively.

In order to study rejection rates under the null hypothesis, we used the initial data sets of 50, 200, and 1000 loci to infer structure, and then simulated additional "candidate" loci under the same model. Table 1 reports rejection rates for STRAT and GC. When 200 or 1000 loci were used to infer structure, the rejection rates for STRAT were very close to the nominal p-values. (The results report average rejection rates based on the twenty data sets used to infer population structure. Longer runs for each of a subset of these data sets indicated little variation in rejection rates across data sets.)

When 50 loci were used to infer structure, STRAT produced unreliable results. For the four data sets in which $\hat{K}$ was 1, the rejection rates were unacceptably high (the test at the candidate markers is then similar to the standard Chi-square test). For the other sixteen data sets, *structure* had some success in identifying the most divergent subpopulation (see Figure 1), and the resulting rejection rates were actually quite close to the target p-values. As in previous simulations (Pritchard *et al.*, 2000b), it seems that STRAT is often surprisingly robust to some mis-estimation of the population structure, provided that there is enough information in the marker data to capture the major features of the structure.

The average rejection rates for GC were close to the nominal values in all three cases. The reported rates represent average rejection rates over repeated experiments (based on 10^4 estimates of λ). As described above, in any particular experiment, the estimated value of λ may be considerably smaller or larger than the true value.[6] Notice also that the accurate rejection rates were achieved when forcing $\hat{\lambda}$ to be ≥ 1, which makes the test more conservative on average. When λ is large, the effect of the truncation is less severe, and simulations under a similar model can produce rejection rates slightly higher than the nominal values for small L (results not shown; see also Bacanu *et al.* (2000)).

[6]Reich and Goldstein (2001) prefer to use an upper bound estimate of λ; this will incur some loss in power.

Power of STRAT and GC. We have also performed simulations comparing the power of STRAT and the Bacanu *et al.* (2000) version of GC under various models. Typical examples are presented in Table 2. When the same effect is present in all subpopulations, the GC test statistic (Bacanu *et al.*, 2000) usually has slightly better power than the test statistic implemented in STRAT. However, the test statistic used by STRAT allows tests against a wider range of alternatives: in particular, it allows for different effects in different populations. In the cases where the effect at the candidate locus varies across subpopulations, STRAT performs rather better—sometimes much better—than GC.

Table 2 also presents power results for the TDT under the same models (see also Bacanu *et al.*, 2000; Pritchard *et al.*, 2000b). These tests are not precisely comparable since the TDT involves 50% more genotyping at the candidate locus. Genomic control and STRAT require genotypes from many unlinked marker loci, but these may come at no extra expense if all the loci are potential candidates. If we compare the power of R parent-offspring trios to a case-control study of R cases and R controls, then TDT achieves slightly better power than either GC or STRAT (when the effect is the same in all subpopulations); but the power of TDT studies with $2R/3$ trios is generally slightly lower than either GC or STRAT (Bacanu *et al.*, 2000; Pritchard *et al.*, 2000b).

Discussion

In this article we have discussed the recent work on using unlinked marker loci to construct valid case-control tests for association. One of our primary goals has been to discuss some of the statistical issues underlying the various approaches.

The major difference between the genomic control and structured association approaches is that the latter attempts to estimate the population structure. One shortcoming of this approach is that the method is likely to perform poorly (in terms of Type 1 error rates) if the inferred structure is an inadequate representation of the truth. Luckily, simulations suggest

that STRAT is often surprisingly robust (Pritchard *et al.*, 2000b). Nonetheless, when there are few marker loci available, or the inference problem is particularly challenging, genomic control may sometimes perform better, as in the example with 50 loci presented above. One departure from model assumptions that may cause difficulties is inbreeding; however the examples given by Devlin and Roeder (1999) are rather extreme, and it is unclear how serious the issue of inbreeding is in practice.

It is important then, when using structured association methods, to be able to detect situations in which the inferred structure is not an adequate representation of reality. Some information about this should be contained in the MCMC output–for instance looking at the posterior distributions of K (the number of subpopulations) and Q (the estimated ancestry of individuals). However, it is not immediately clear exactly how to use this sort of information, and this issue may be worthy of further study. An alternative approach was proposed by Pritchard *et al.* (2000b), who suggested that the empirical distribution of p-values across the marker loci could be used to check the performance of STRAT in any particular study. In a study of many loci we expect that very few are in genuine association with the phenotype. Thus, after using a structured association method to test for association at many loci, the resulting distribution of p-values should be approximately uniform. If it is not, this indicates that the method may be performing poorly.

Unlike the structured association approach, genomic control assumes a particular parametric distribution of the value of the test statistic. It is not known yet how appropriate this model is in practice, and it will be important to validate the model empirically. In particular, if the actual distribution has thicker tails than predicted, this could lead to high Type-1 error rates. For instance, selection at a candidate locus (or at nearby loci) could drive the alleles to different frequencies in different subpopulations, making large values of the test statistic much more likely. For this reason, genomic control may not be well-suited to studies in the MHC region, for instance, where it is known that selection is important. Variation in effective population size across the genome (for instance due to background selection, or differences

between the X chromosome and autosomes) may also lead to variation in F, by changing the rate of genetic drift. Currently, a shortcoming of the genomic control approach is that it is restricted to biallelic markers, while SA can handle more complicated data, including microsatellites or haplotypes.

Another important difference between the various methods lies in the way in which they use the subpopulation information. Pritchard *et al.* (2000b) construct a test statistic which compares allele frequencies between cases and controls within subpopulations and combines the information across subpopulations. The genomic control approach uses a test statistic that averages the information across subpopulations (which are unknown in that case). Similarly, Satten *et al.* (2001) fit a particular (logistic) model of the way genotype affects susceptibility, assuming, in effect, that this is the same in all subpopulations. The statistical trade-offs between the two approaches are clear. If the effect of the candidate locus really is consistent across subpopulations, then these tests will perform better than the more general approach of Pritchard *et al.* (2000b) (which requires the estimation of more parameters). But this comes at a price. If the effects vary across subpopulations, the use of the less general approach will tend to result in poor estimates or tests with low power.

It seems biologically plausible that effects could vary substantially across subpopulations. First, we will typically be testing for association at a marker that is (hopefully) in linkage disequilibrium with nearby disease mutations, rather than testing for association at the actual site of a disease mutation. Even if the same disease mutation is most common in each subpopulation, the patterns of LD at linked markers may differ. Moreover, we may not expect that, at a given locus, it is the same disease mutations that are common in each subpopulation. In this case the direction of the effects at a candidate locus in each subpopulation will be essentially independent. Environmental conditions or epistatic effects could also differ among subpopulations, giving rise to differences in gene effects. For these reasons, it seems preferable at this time to consider reasonably general models that are powerful against such variation across subpopulations.

There is considerable scope for further investigation into statistical tests

within the SA framework. One strength of the SA approach is that it can readily accomodate a variety of test statistics, and it will be of interest to study the performance of such tests against a range of alternative models. For example, one might think of alternative models in which the candidate locus effects are correlated in closely related subpopulations, but less so in more divergent subpopulations. It would also be natural to extend SA methods to other settings, including quantitative traits (see Thornsberry *et al.*, 2001) and haplotype methods.

We now turn our attention to a comparison of two SA approaches. In their paper, Satten *et al.* (2001) raise four concerns about the method of Pritchard *et al.* (2000b) which we discuss here, as follows.

1. They champion estimation of genotype effects rather than simply testing for association. Each has its place, but we do not see this as a fundamental distinction between the methods. Having estimated subpopulation information, this could be used to estimate estimate effects, either parametrically, or non-parametrically, within the framework of Pritchard *et al.* (2000b), in contrast to the claim of Satten *et al.* (2001).

2. Satten *et al.* (2001) claim that their method properly accounts for variability in selecting K, the number of subpopulations. In fact, because they use an inconsistent estimator of K, it is not clear that their bootstrap approach will behave well. As the authors note, their simulation study was too small to assess coverage properties adequately, even for data simulated under the model they used for estimation. (Their simulations were also for subpopulations much more distinct than might be realistic in most applications.) The test of Pritchard *et al.* (2000b) conditions on the inferred K. We would suggest use of structured association in situations in which this can be estimated well, in which case there will be little practical difference between conditional and unconditional tests.

3. Satten *et al.* (2001) claim that our method "requires a Gibbs sampler that changes the number of parameters in the model". This is simply false. As noted above, and in the original papers, we avoid such a procedure precisely because of the difficulties involved.

4. Satten *et al.* (2001) use genotype data at the candidate locus in addition

to the unlinked markers to infer structure, and say that this information is not used by Pritchard *et al.* (2000b). While the simulation examples provided in this paper do not use the candidate loci, this was done primarily to simplify the simulations. In fact, the publicly available programs *structure* and *STRAT* are designed to use all the loci (see also Model C of Pritchard *et al.* (2000b)). Satten *et al.* (2001) are correct, however, that we do not use the phenotype information when inferring structure; this means that the model used to infer structure is not completely correct at a (true) candidate locus. This effect is probably negligible, however, given that hundreds of SNPs may be needed to apply these methods successfully in practical problems.

In summary, the recent work on statistical methods for constructing valid case-control studies seems to represent an important step forward in the use of association studies for mapping complex disease genes. These methods will be particularly appropriate for use in the large-scale studies that will be needed to find genes of small effect, because while the impact of population structure becomes more severe as sample sizes and number of loci increase, the inference of population structure becomes easier. We have discussed at length the differences between the genomic control and structured association approaches; both have advantages, and both may suffer problems in different settings. Structured association methods can potentially provide some statistical advantages, provided that the data are sufficiently informative to infer population structure adequately. As mapping studies move towards genotyping hundreds or thousands of markers this inference will become increasingly reliable.

Acknowledgements

This work was supported by a grant to JKP from Burroughs-Wellcome Fund, and by grants from EPSRC (GR/M14197) and BBSRC (43/MMI09788) to PD. We thank Warren Ewens, Kathryn Roeder and Noah Rosenberg for comments on the manuscript.

References

Bacanu, S. A., Devlin, B. and Roeder, K. (2000) The power of genomic control. *Am. J. Hum. Genet.*, **66**, 1933–1945.

Balding, D. J. and Nichols, R. A. (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Boehnke, M. and Langefeld, C. D. (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am. J. Hum. Genet.*, **62**, 950–961.

Cardon, L. R. and Bell, J. I. (2001) Association study designs for complex diseases. *Nature Reviews Genetics*, **2**, 91–99.

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. (1994) *The history and geography of human genes*. Princeton: Princeton University Press.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Devlin, B., Roeder, K. and Wasserman, L. (2001) Genomic control, a new approach to genetic-based association studies. *Theor. Pop. Biol.*, **In Press**.

Ewens, W. J. and Spielman, R. S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.*, **57**, 455–464.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hurvich, C. M. and Tsai, C. L. (1989) Regression and time-series model selection in small samples. *Biometrika*, **76**, 297–307.

Knowler, W. C., Williams, R. C., Pettitt, D. J. and Steinberg, A. G. (1988) Gm3-5,13,14 and Type-2 Diabetes-Mellitus—an association in American-Indians with genetic admixture. *Am. J. Hum. Genet.*, **43**, 520–526.

Lander, E. S. and Schork, N. (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–2048.

Lazzeroni, L. C. and Lange, K. (1998) A conditional inference framework for extending the transmission/disequilibrium test. *Hum. Hered.*, **48**, 67–81.

Pfaff, C. L., Parra, E. J., Bonilla, C., Hiester, K., McKeigue, P. M., Kamboh, M. I., Hutchinson, R. G., Ferrell, R. E., Boerwinkle, E. and Shriver, M. D. (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.*, **68**, 198–207.

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

Pritchard, J. K. and Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Gen.*, **65**, 220–228.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000a) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000b) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170–181.

Reich, D. E. and Goldstein, D. B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, **20**, 4–16.

Ripatti, S., Pitkäniemi, J. and Sillanpää, M. J. (2001) Joint modeling of genetic association and population stratification using latent class models. *Genetic Epidemiology*, **In Press**.

Risch, N. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.

Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Satten, G. A., Flanders, W. D. and Yang, Q. (2001) Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.*, **68**, 466–477.

Sillanpää, M. J., Kilpikari, R., Ripatti, S., Onkamo, P. and Uimari, P. (2001) Bayesian association mapping for quantitative traits in a mixture of two populations. *Genetic Epidemiology*, **In Press**.

Spielman, R. S. and Ewens, W. J. (1996) The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. of Hum. Genet.*, **59**, 983–989.

Spielman, R. S. and Ewens, W. J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.*, **62**, 450–458.

Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **52**, 506–513.

Terwilliger, J. D. and Weiss, K. M. (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotech.*, **6**, 578–594.

Thomas, D. C. and Witte, J. S. (2001) Population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiology, Prevention, and Biomarkers*, **In Press**.

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D. and Buckler, E. S. (2001) *Dwarf8* polymorphisms associate with variation with flowering time. *Nature Genetics*, **28**, 286–289.

Wright, S. (1951) The genetical structure of populations. *Ann. Eugen.*, **15**, 323–354.

L	P-value	STRAT	GC	Chi-square
50	0.1	.107 (.13)	.099	.141
	0.01	.013 (.02)	.010	.021
	0.005	.007 (.01)	.004	.010
200	0.1	.099	.100	.139
	0.01	.010	.010	.021
	0.005	.005	.005	.012
1000	0.1	.099	.100	.142
	0.01	.011	.011	.021
	0.005	.006	.006	.012

Table 1: Rejection rates under the null hypothesis, estimated from 10^4 simulated candidate loci. L shows the number of loci used to estimate population structure for STRAT (Pritchard *et al.*, 2000b), or GC (method of Bacanu *et al.*, 2000). "Chi-square" refers to rejection rates using a standard χ^2 test that ignores population structure (these do not depend on L). The results for STRAT show average rejection rates obtained when using twenty simulated data sets of L loci to infer structure before testing for association at 500 candidate loci each; the results for GC are the average rates after simulating 10^4 sets of L loci to estimate λ. When using 50 loci, *structure* could not detect all three subpopulations. The results for STRAT at $L = 50$ are split up according to whether $\hat{K} = 2$, or $\hat{K} = 1$ (the latter, in parentheses, occurred for four of the twenty data sets).

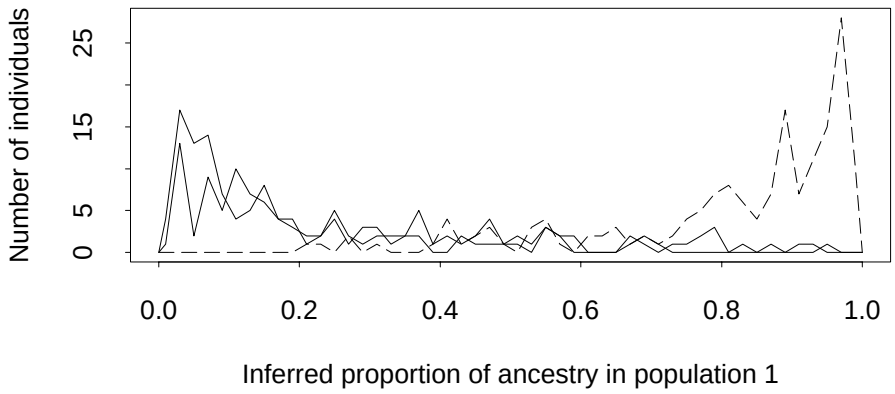
R_1, R_2, R_3	L	P-value	STRAT	GC	TDT
1.5, 1.5, 1.5	200	.01	.29	.35	.45
	1000	.01	.30	.35	
	200	.001	.12	.14	.21
	1000	.001	.12	.14	
1.0, 1.0, 2.0	200	.01	.36	.22	.30
	1000	.01	.36	.22	
	200	.001	.16	.07	.12
	1000	.001	.16	.07	
0.5, 0.5, 1.5	200	.01	.64	.04	.06
	1000	.01	.66	.04	
	200	.001	.40	.008	.01
	1000	.001	.43	.005	

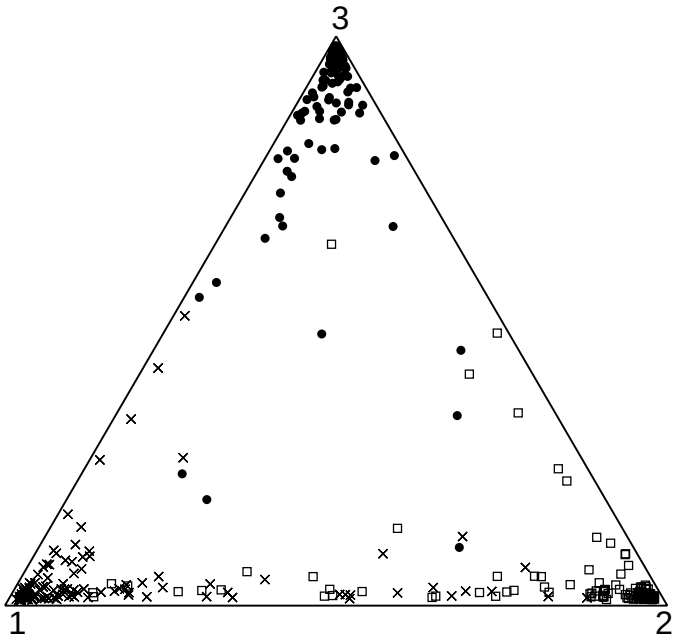
Table 2: Estimated power of three tests of association under the model described in the text, for different p-values. R_1, R_2, and R_3 give the relative risk of alleles at the candidate locus, in subpopulations 1, 2 and 3, assuming a multiplicative model of allele interactions (Pritchard *et al.*, 2000b). L gives the number of loci used by STRAT, and GC, respectively, to infer population structure. Note that the TDT is not precisely comparable since the study design is different: the TDT results are for 200 parent-offspring trios, instead of 200 cases and 200 controls. As for Table 1, the results for GC are based on 10^4 sets of L loci and the results for STRAT are averages from 20 sets of L loci.

Figure captions

Figure 1: Ancestry of sampled individuals estimated from 50 biallelic loci (top plot), 200 loci (middle plot), and 1000 loci (bottom plot) using *structure* (Pritchard *et al.*, 2000a). The data were simulated under a model with three closely-related subpopulations (see text). For the simulated data presented in the top plot, only two subpopulations were detected. Hence, the ancestry of each individual is represented by a number between 0 and 1; the solid lines plot histograms of the estimated values for individuals who were actually from subpopulations 1 or 2; the dashed line is for individuals from subpopulation 3 (which had the most divergent allele frequencies). With 200 and 1000 loci, the correct number of subpopulations (three) was inferred for all the simulated data sets. Then the inferred ancestry for each individual is a vector with three elements summing to 1; these are represented as points on an equilateral triangle (each element of the vector is given by the distance to one edge). Individuals from subpopulations 1, 2 and 3, are represented by crosses, open squares, and filled circles, respectively. All of the points in the extreme corners of the lower two plots (where they are hard to resolve) are correctly classified.

Figure 2: Distributions of $\hat{\lambda}$ obtained from the median estimator (Bacanu *et al.*, 2000) using 50 loci (solid line), 200 loci (dotted line), and 1000 loci (dashed line) under the model with three subpopulations described in the text. Distributions estimated from 10^4 data points. In our simulations of genomic control we adjusted values of $\hat{\lambda} < 1$ upward to 1.0, as in Bacanu *et al.* (2000).


Number of individuals
0
5
15
25
0.0
0.2
0.4
0.6
0.8
1.0
Inferred proportion of ancestry in population 1


3
1
2

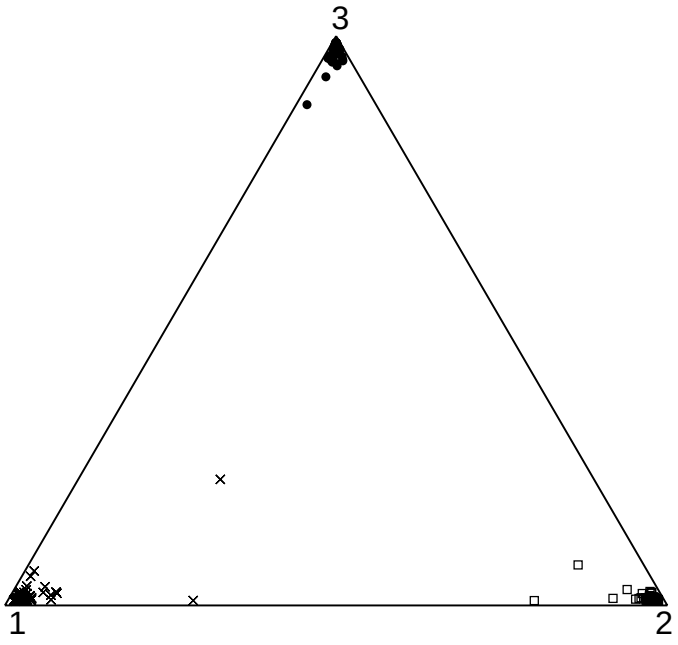

3
1
2

Figure 1

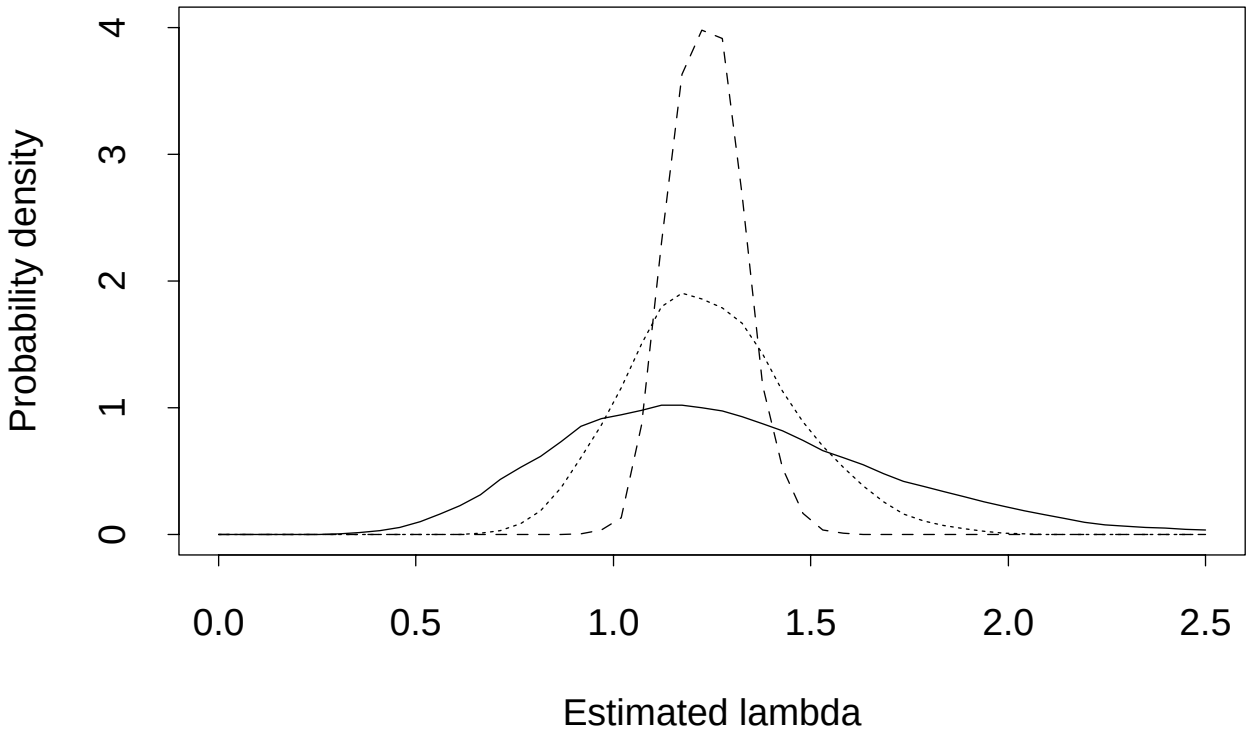

Probability density
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
Estimated lambda

Figure 2