Original Paper

Human Heredity

Hum Hered 2004;58:93–107
DOI: 10.1159/000083030

Received: June 11, 2004
Accepted after revision: September 2, 2004

Properties of Structured Association Approaches to Detecting Population Stratification

Shaun Purcell[a, b] Pak Sham[b, c]

[a]Whitehead Institute, Nine Cambridge Center, Cambridge, Mass., USA; [b]Social, Genetic and Developmental Psychiatry Research Centre, Institute of Psychiatry, King's College London, London, UK; [c]Department of Psychiatry and Genome Research Centre, Faculty of Medicine, University of Hong Kong, Hong Kong, PRC

Key Words
Population stratification · Lateral class analysis · Structured association · Genomic control · Linkage disequilibrium

Abstract
Objective: To examine the properties of the structured association approach for the detection and correction of population stratification. ***Method:*** A method is developed, within a latent class analysis framework, similar to the methods proposed by Satten et al. (2001) and Pritchard et al. (2000). A series of simulations illustrate the relative impact of number and type of loci, sample size and population structure. ***Results:*** The ability to detect stratification and assign individuals to population strata is determined for a number of different scenarios. ***Conclusion:*** The results underline the importance of careful marker selection.

Background

Population stratification refers to a recent mixture of subpopulations which may differ in allele frequencies at many loci across the genome. A stratified sample is therefore one in which discrete subpopulations that do not interbreed as a single randomly-mating unit are pooled together. The early population genetic work on population stratification was primarily concerned with its impact on genotypic frequencies and the evolutionary process [1], although subsequently its potential impact in disease-gene association studies was highlighted [2]. If cases and controls are not matched for ethnic background, population stratification effects can lead to spurious association. Although the primary focus was on population stratification generating type I, or false positive errors, stratification can also reduce power (that is, to increase type II errors) if the stratification effect 'masks' the trait locus effect.

A stratified sample will display certain characteristic 'signatures', both at single loci and also across unlinked loci. At a single locus, stratification induces a non-independence between maternal and paternal alleles, i.e. Hardy-Weinberg disequilibrium (HWD). Across unlinked loci, stratification can induce a similar non-independence of alleles, i.e. linkage disequilibrium (LD). Two approaches to detecting these signatures, in order to correct for stratification, have been suggested, now labelled 'genomic control', e.g. [3], and 'structured association', e.g. [4, 5]. Both approaches require multilocus genotype data from across the genome for each individual in the sample. The essence of the genomic control approach is that popu-

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2004 S. Karger AG, Basel
0001–5652/04/0582–0093$21.00/0

Accessible online at:
www.karger.com/hhe

Shaun Purcell
Psychiatric and Neurodevelopmental Genetics Unit
Massachusetts General Hospital, 149 13th Street, Floor 10
Charlestown, MA 02129 (USA)
Tel. +1 617 726 7642, Fax +1 617 726 0830, E-Mail spurcell@pngu.mgh.harvard.edu

lation stratification leads to a systematic 'over-dispersion' of χ^2 statistics in the disease-gene association test, an effect that can be estimated and adjusted for. Structured association attempts to assign individuals to subpopulations and to test for association conditional on subpopulation membership. Pritchard et al. [4] developed the STRUCTURE program based on a Bayesian framework, and Satten et al. [6] adopted a latent class analysis (LCA) [7] approach within a maximum-likelihood (ML) framework, using the E-M algorithm. Although Bayesian and ML approaches differ in the statistical apparatus employed, both share similar underlying models.

Structured association offers certain advantages over the genomic control approach. First, any structure in a sample is of intrinsic interest – rather than simply computing a single inflation factor, it is informative to classify individuals into meaningful groups. Structured association can also handle allelic heterogeneity between subpopulations – subpopulation membership can be entered as an interaction term as well as a covariate in any subsequent association test. Finally, unlike genomic control methods, which merely provide an average correction factor, structured association does not assume that the genetic distance between two groups is constant across the genome.

The present work also adopts a structured association, LCA-based approach similar to Satten et al. [6], albeit with several extensions. Furthermore, we present some simple simulation results that illustrate the conditions under which this method might be expected to perform effectively. Broadly speaking, these results conform with a recent, more comprehensive study of the informativeness of genetic markers for inference of ancestry [8]. Recent work has shown that even modest levels of stratification that might go undetected by standard applications of genomic control or structured association can bias the results of large association studies [9]. Similarly, another recent study assessed stratification empirically, by analyzing data from 11 case-control and case-cohort association studies [10], finding that a larger number of markers than previously thought is necessary to detect even moderate levels of stratification. These results make even more pressing the need to assess the properties of structured association methods and develop guidelines for their use in large-scale association studies: although the number of markers used is one important variable, the results below illustrate some of the other determinants of success for this approach.

Methods

A population is assumed to consist of K hidden sub-populations. The basic model assumes that each individual belongs to one and only one sub-population, that mating occurs randomly within each sub-population and that these sub-populations may vary in allele frequencies at loci all across the genome. The aim is to breakdown a population that, as a whole, potentially displays Hardy-Weinberg and linkage disequilibrium across unlinked loci into a number of sub-populations, such that within each sub-population there is Hardy-Weinberg and linkage equilibrium. In practice, the markers do not necessarily need to be completely unlinked: they must be sufficiently distant to be in linkage equilibrium within subpopulation (about 1 cM in homogeneous populations).

The aim of latent class analysis is to probabilistatically assign individuals to class C of K possible classes on the basis of their responses to multiple variables. In the present context, each class C corresponds to a potential population stratum; individuals' responses correspond to sets of genotypes measured on unlinked loci, G. The LCA model involves three inter-related sets of probabilities $P(C|G)$, $P(G|C)$ and $P(C)$. For a specific K, the main values to be estimated are the posterior class probabilities $P(C|G)$: the probability that an individual belongs to a subpopulation conditional on genotypic configuration. The E-M algorithm [11] is used to iteratively calculate $P(C|G)$ by estimating $P(G|C)$ and $P(C)$. $P(G|C)$ represents class-specific allele frequencies – the probability that an individual picked from a certain class has a certain allele at a particular locus. $P(C)$ are the prior probabilities of class membership: the probability that an individual picked at random belongs to class C irrespective of G. For $K > 1$, $P(C)$ represents the mixing proportions of the various classes. Critically, $P(C|G)$ are calculated under the assumptions of Hardy-Weinberg and linkage equilibrium holding *within* each class.

The posterior probability of individual i belonging to class j is $P(C = j|G_i)$. The relative frequency in class j of allele k at locus l is $P(G_l = k|C = j)$. (Note that when G is indexed by an i subscript, it refers to an individual's multilocus genotype; when G is indexed by l, it refers to a single locus in the entire population.) The E-M algorithm proceeds in two steps; the expectation, or E-step, involves calculating the values of $P(C)$ and $P(G|C)$ implied by $P(C|G)$; the maximisation, or M-step, involves recalculating $P(C|G)$ given the new estimates of $P(C)$ and $P(G|C)$. These two steps then iterate until convergence. Details of the method are presented in Appendix 1.

As well as estimating $P(C|G)$ for $K = 1, 2, ...$ one wants to ask: does a more complex model (i.e. higher K) provide a significantly better description of the data? In particular, is there evidence of *any* stratification (i.e. $K > 1$)? As different solutions involve different numbers of unique parameters and are not nested, the Akaike Information Criterion (AIC) [12], defined as minus twice the log-likelihood plus twice the number of model parameters, is used to evaluate different models. There are $K - 1$ non-redundant parameters in $P(C)$ and $\Sigma_l(K(n_l - 1))$ in $P(\mathrm{G}|C)$ if locus l has n_l alleles. The lowest AIC solution is taken to be the most parsimonious and best-fitting explanation of the data. In the absence of any a priori considerations regarding population substructure, only the $P(C|G)$ from the K-solution with the lowest AIC should be used as covariates in any subsequent association analysis. Although the AIC is commonly used in the context of LCA, it does display a tendency to over-estimate K. However, in the present context (deriving a solution in order to correct for population stratification in subsequent association analysis) this is not of particularly great concern – it will lead towards more conservative tests, but this is probably preferable to under-estimating K in any case.

Admixture Models

So far we have assumed a simple population genetic model: K distinct subpopulations of varying size that differ in allele frequencies at unlinked markers; also that Hardy-Weinberg and linkage equilibrium exist within each subpopulation. A more general and realistic model allows for admixture between subpopulations. That is, we may wish to characterise as *admixed* individuals who have descended from two or more other subpopulations also seen in the sample, rather than assuming that a further distinct class exists. Such a model is potentially more powerful and more revealing of hidden population structure.

Admixture is modelled in terms of a finite number of *derived classes* (C_D) that represent an admixture of one or more *ancestral classes* (C_A). Considering discrete sets of admixture proportions by constraining possible proportions to a $1/r$ resolution, where r is specified by the investigator, we can enumerate all possible derived classes for a given number of ancestral classes. For example, if $r = 2$ and there are 3 ancestral classes, six derived classes are implied. The matrix

$$\Theta = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.50 & 0.50 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.50 & 0.00 & 0.50 \\ 0.00 & 0.50 & 0.50 \\ 0.00 & 0.00 & 1.00 \end{bmatrix}$$

represents the mixing proportions of the three ancestral classes (columns) in the six derived classes (rows). Three of these derived classes are *pure* in the sense that they are derived from only one ancestral class, the other three derived classes are *admixed*. Reading across rows, the elements of Θ represent the proportion of an individual's genome that is derived from each ancestral class.

Counting individuals, rather than alleles, is still straightforward: individuals are counted directly into derived classes. If $I()$ represents a count of individuals, the prior derived class probabilities are therefore simply estimated as $P(C_D = d) = I(C_D = d)/N$ where $I(C_D = d) = \Sigma_i P(C_D = d|G_i)$. However, rather than directly counting alleles into derived classes, the two layers of classes must now be considered. Of primary interest are the parameters for the derived classes, which correspond to the simple classes considered previously: posterior probabilities are only calculated for the derived classes, $P(C_D|G)$. The presence of the ancestral classes effectively places constraints on how the allele-counting step proceeds, however. Details of the method are presented in Appendix 2.

Other Extensions

Fixing $P(C|G)$

It may sometimes be desirable to allocate individual i to latent class j, by fixing $P(C = j|G_i)$ to 1 and $P(C \neq j|G_i)$ to 0, rather than estimating these values. This procedure allows the likelihood to be calculated for any classification of individuals based on external criteria (e.g. self-reported ancestry). Additionally, this procedure can be used to 'anchor' the solution: for example, if the sample contains a few 'prototypical' individuals (i.e. those with unambiguous ethnic group information) then these individuals can be fixed to specific classes. This is particularly useful when more complex admixture models are specified.

Haploid Organisms and X Chromosome Data

Although the method above applies to diploid genotypic data, a straight-forward modification enables the analysis of haploid organisms, or of X chromosome data in males. In particular, if $A()$ represent allele counts, only one allele at each locus is now counted in the E-step $A(G_l = k|C = j) = \Sigma_i P(C = j|G_i)D_1$ and so the class-specific allele frequencies are now $P(G_l = k|C = j) = A(G_l = k|C = j)/I(C = j)$ whilst in the M-step, the calculation of $P(C|G)$ becomes $P(C = j|G_i) = P(C = j)\ \Pi_l\ P(G_l = k_{i1}|C = j)/(\Sigma_{j'} P(C = j')\ \Pi_l\ P(G_l = k_{i1}|C = j'))$. (See Appendix 1 for more details on the basic method.)

The Hardy-Weinberg Equilibrium Assumption

Population stratification is not the only cause of Hardy-Weinberg disequilibrium. One other potentially common cause is selective genotyping error. Consider the scenario in which heterozygous individuals are more likely to have a missing genotype. This loss of heterozygosity is not likely to lead to spurious association – but might it lead to 'spurious stratification'. That is, the current method might take HWD due to missing heterozygous genotypes as evidence of stratification and therefore favour a spurious $K > 1$ solution.

This possibility was investigated by simulation: 10 replicate homogeneous datasets of 400 individuals and 40 SNPs (equal allele frequencies) were simulated. Heterozygotes were designated missing with probability 0, 25, 50 and 75%. Therefore, in the last (unrealistically extreme) condition (75% missing) substantial deviations from HWE were observed. The data were analysed for $K = 1$ and $K = 2$ solutions in the standard manner. The approach was also modified, to relax the within-class HWE assumption, by treating genotypes as the unit of response rather than alleles (i.e. equivalent to assuming all individuals to be haploid and that each genotype is a unique allele).

As table 1 illustrates, a large percentage of the heterozygotes must be missing in order to favour a two-class solution (i.e. positive values of $AIC(K = 1) - AIC(K = 2)$) – it is very unlikely that this level of genotyping failure would occur in practice for all markers. Also, the

Table 1. The impact of selective genotyping failure: relaxing the within-class HWE assumption

Missing A_1A_2	$AIC(K = 1) - AIC(K = 2)$	
	HWE assumed	HWE relaxed
0%	–72.63	–68.01
25%	–52.25	–60.78
50%	13.41	–69.79
75%	119.76	–72.55

The figures represent the difference in AIC for a $K = 1$ and a $K = 2$ solution: Positive values therefore indicate that stratification has been detected (i.e. a $K = 2$ solution is favored over a $K = 1$ solution).

specification of equal allele frequencies represents a 'worst-case scenario' (i.e. it gives the highest possible frequency of heterozygotes). Furthermore, when the option to relax the within-class HWE assumption was implemented, the AIC difference remained invariant to the marker HWD. Similar results were obtained when different genotyping artefacts were simulated to induce the HWD: for example, some proportion of heterozygotes being called as homozygotes.

Genetic Outlier Detection

A related goal to detecting subpopulations within a sample is the detection of population outliers using genetic background information. That is, the sample may be relatively homogeneous except for one or two individuals. These individuals would not constitute a class by themselves – but it might be of interest to identify such individuals before embarking on any other analyses. A proposed method is first to calculate the sample log-likelihood ln L_0 for $K = 1$. Then, for each individual i, the sample log-likelihood ln L_i is calculated for $K = 2$ but with individual i fixed to class 2 (i.e. fix $P(C = 2 \mid G_i) = 1$) and all other individuals fixed to class 1 (i.e. fix $P(C = 1 \mid G_j) = 1$ for $j \neq i$). The difference $\ln L_i - \ln L_0$ is a measure of genetic distance and can be inspected to identify genetically outlying individuals. A similar approach has been proposed by Fisher et al. [13].

Diagnostic Statistics

Several diagnostic statistics can be used to aid the model-fitting process. An inter-class genetic distance matrix using Nei's measure of genetic distance [14] is calculated from the class-specific allele frequencies. Nei's genetic distance between two classes, j_1 and j_2, for N loci is calculated

$$d_{\text{Nei}} = -\ln \frac{\Sigma_{l=1}^{N} \Sigma_k \, [P(G_l = k \mid C = j_1) P(G_l = k \mid C = j_2)]/N}{\sqrt{\Sigma_{l=1}^{N} \Sigma_k \, [P(G_l = k \mid C = j_1)^2] \, \Sigma_{l=1}^{N} \Sigma_k \, [P(G_l = k \mid C = j_2)^2]/N}}$$

It is especially convenient to apply a multidimensional scaling technique to the distance matrix, in order to obtain a visual representation of the class structure. The class-specific allele frequencies are also given in the output of the computer program L-POP, which allow the calculation of other useful summary statistics such as the F_{ST} index, representing a general measure of genetic differentiation in the sample.

Additionally, an 'entropy' measure is calculated for each individual, to indicate how well that individual has been classified in the final solution. Entropy for individual i is calculated by summing over all j classes 1 to K: $-\Sigma_{j=1}^{K} P(C = j \mid G_i) \ln P(C = j \mid G_i)$ where $P(C = j \mid G_i) > 0$. The measure ranges between 0 and 1, where a lower value represents a better classification.

Inter-class Nei genetic distances are also calculated for each locus separately. These statistics can be useful for identifying which loci are contributing to solutions with $K > 1$. Typically, one would expect all loci to contribute approximately equally. In cases where only a couple of loci stand out as contributing much more than the others, it is worth investigating the positions of these loci – it might be indicative of the loci being tightly linked. In this case, at least one of the markers should be removed from the dataset. Class-specific locus-specific genetic distances are also calculated (i.e. comparing class j against all other classes for that locus).

It may also be of interest to compare different solutions against each other, or against an external classification scheme. For each solution, the data can be partitioned by assigning each individual to a single class based on highest posterior probability; for each pair of solutions a two-way contingency table can be constructed. The adjusted RAND index [15] is a measure of agreement specifically designed to compare partitioning schemes of data from clustering methods; importantly, this measure is able to compare solutions with different numbers of classes. The adjusted RAND index varies between 0 and 1 (where 0 represents no agreement and 1 represents complete agreement) and is calculated

$$\text{RAND} = \frac{\Sigma_{i,j} \binom{n_{ij}}{2} - \left[\Sigma_i \binom{n_{i.}}{2} \Sigma_j \binom{n_{.j}}{2} \right] \Big/ \binom{n}{2}}{\frac{1}{2} \left[\Sigma_i \binom{n_{i.}}{2} + \Sigma_j \binom{n_{.j}}{2} \right] - \left[\Sigma_i \binom{n_{i.}}{2} \Sigma_j \binom{n_{.j}}{2} \right] \Big/ \binom{n}{2}}$$

where n_{ij} is the observed count for individuals classified into class i for the first solution and class j for the second solution; the marginal counts for the first and second solutions are represented as $n_{i.}$ and $n_{.j}$ respectively.

Correction for Stratification in Association Analysis

Whereas the approach of Satten et al. [6] combines the test of association for binary disease traits with the detection of stratification, the current approach separates these two aspects of the problem. The most simple strategy is to use posterior probabilities $P(C = 1 \mid G)$ to $P(C = K - 1 \mid G)$ from the best-fit solution as covariates in whatever test of association is required. Alternatively, individuals can be assigned to discrete classes on the basis of their highest $P(C \mid G)$ (although this can induce a bias if the highest posterior probabilities are not very near 1).

We have developed an approach that evaluates the likelihood of observing an individual's genotype conditional on trait score, $P(G \mid X)$. This 'conditioning-on-trait values' approach has been previously adopted in the context of complex segregation analysis [16] and variance components linkage [17]. To allow for stratification effects, association is modelled conditional on belonging to class j of K discrete classes. For each individual, the probabilities of belonging to each class will be the posterior probabilities produced by a method such as the one described above, using genetic background information. Alternatively, these 'probabilities' could be binary variables coded 0/1 based on some other classification scheme, such as self-reported ethnicity. The posterior probabilities are denoted $P(C \mid G)$. The class-conditional likelihood will be based on $P(G \mid X, C)$. The overall likelihood will be the weighted sum $\Sigma_j P(G \mid X, C_j) P(C_j \mid G)$ therefore.

The model is parameterised in terms of class-specific additive genetic values (a_j), dominance deviations (d_j)

and allele frequencies (p_j). Mean-centred class-specific genotypic means are calculated

$$\mu_{11|j} = a_j - (a_j(p_j - q_j) + 2p_jq_jd_j)$$

$$\mu_{12|j} = d_j - (a_j(p_j - q_j) + 2p_jq_jd_j)$$

$$\mu_{22|j} = -a_j - (a_j(p_j - q_j) + 2p_jq_jd_j)$$

and class-specific genotype frequencies $P(G_{11}|C)$, $P(G_{12}|C)$ and $P(G_{22}|C)$ are calculated p_j^2, $2p_jq_j$ and q_j^2. The trait must be standardised prior to analysis using the population mean and variance, which must either be estimated from an unselected sample or obtained from other sources. The residual trait variance is

$$\sigma_R^2 = 1 - \sum_j P(C_j)(P(G_{11}|C_j)\mu_{11|j}^2 + P(G_{12}|C_j)\mu_{12|j}^2 + P(G_{22}|C_j)\mu_{22|j}^2)$$

where $P(C_j)$ is the prior probability of belonging to class j, calculated by summing posterior probabilities over all N individuals in the sample, $\Sigma_i\, P(C_j|G)/N$. Applying Bayes Theorem to $P(G|X,C)$, the likelihood of observing genotype G_i is the mixture of likelihoods summed over all possible classes weighted by the posterior class probabilities

$$L(G_i|X_i) = \sum_j \frac{P(X_i|G_i, C_j)P(G_i|C_j)}{\Sigma_F\, P(X_i|G_F, C_j)P(G_F|C_j)} P(C_j|G_i)$$

where the sum F is over all genotypes. For individual i, the probability of observing the trait score conditional on genotype and class is given by the normal or logistic density function, depending on whether the trait is continuous or binary. By either fixing or equating parameters, likelihood ratio test statistics can then be constructed between null and alternate models as minus twice the difference in log-likelihood. Simulation results show that this formulation is equivalent to a more standard regression-based approach in most circumstances and is more powerful when the continuous trait is non-normal, or if the sample has been selected on the basis of extreme trait values [18]. This basic approach can be easily extended to multiallelic or haplotype analysis, most simply by comparing each specific allele or haplotype versus all others (and correcting for multiple testing by use of permutation procedures, if so desired).

Table 2. A comparison of L-POP and STRUCTURE: Determining *K* in L-POP

Stratification	Loci	*K* = 1	*K* = 2	*K* = 3
N	20	0.00	5.00	8.58
	50	0.00	5.54	18.22
	100	0.00	11.25	31.50
Y	20	338.10	11.59	0.00
	50	1,306.59	2.73	0.00
	100	3,001.63	0.00	32.64

The values represent the AIC values for a *K* = 1, *K* = 2 and *K* = 3 class solution minus the minimum across each row of these three values (i.e. the best-fit solution will have a value of 0). The *K* = 1 solution was correctly selected in all cases when there was no stratification. In contrast, when there was stratification (such that in reality *K* = 2) we see that a *K* = 1 solution is not the best fit (although there is a tendency to overestimate *K* when a small number of markers are used). See the text for a description of the simulated stratification.

Comparison with STRUCTURE

As noted above, the present model is similar to the basic underlying approach used in the STRUCTURE program. We performed some simple simulations to investigate the equivalence of the two methods (using Version 2.1 of STRUCTURE). STRUCTURE offers more than one kind of model, although it is not always clear which model would be most appropriate for a given sample – as the STRUCTURE manual notes, ‘ ... some experimentation on the part of the user’ is required. We therefore used four models in STRUCTURE: first varying whether or not an admixture model was used. In the no admixture model, individuals are assumed to discretely belong to either one population or another. We also applied both correlated and independent allele frequencies models. In general, the authors of STRUCTURE recommend using an admixture model, because of it’s flexibility; also, they suggest the correlated allele frequency model is a good ‘default’ model. Relatively long burn-in (50,000) and run lengths (500,000) were used for STRUCTURE analysis; also, each analysis was repeated three times to assess convergence. In each case, we fit models for $K = 1$, $K = 2$ and $K = 3$, using both STRUCTURE and the method described above, implemented in the computer program L-POP.

In all cases, 500 individuals were simulated, for either 20, 50 or 100 SNPs. Either no population stratification was generated, or a simple two-strata substructure was introduced, such that Wright’s F_{ST} value was 0.04. As these simulations were intended as simple illustrations of the methods rather than comprehensive evaluations, only a single dataset was generated in each condition.

The results from L-POP are shown in table 2, which gives the AIC for $K = 1$, $K = 2$ and $K = 3$ solutions (normal-

Table 3. A comparison of L-POP and STRUCTURE: determining *K* in STRUCTURE

ADMX	CORR	*M*	No stratification (*K* = 1)			Stratification (*K* = 2)		
			K = 1	*K* = 2	*K* = 3	*K* = 1	*K* = 2	*K* = 3
Y	Y	20	1.000	0.000	0.000	0.000	0.063	0.937
			1.000	0.000	0.000	0.000	0.646	0.354
			1.000	0.000	0.000	0.000	0.917	0.083
		50	1.000	0.000	0.000	0.000	0.997	0.003
			1.000	0.000	0.000	0.000	0.750	0.250
			1.000	0.000	0.000	0.000	0.083	0.917
		100	1.000	0.000	0.000	0.000	0.401	0.599
			1.000	0.000	0.000	0.000	0.953	0.047
			1.000	0.000	0.000	0.000	0.214	0.786
	N	20	1.000	0.000	0.000	0.000	0.426	0.574
			1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	0.000	1.000
		50	1.000	0.000	0.000	0.000	0.154	0.846
			1.000	0.000	0.000	0.000	0.881	0.119
			0.999	0.000	0.001	0.000	0.964	0.036
		100	0.000	0.012	0.988	0.000	0.622	0.378
			0.000	0.000	1.000	0.000	0.475	0.525
			0.000	0.995	0.005	0.000	0.475	0.525
N	Y	20	1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	1.000	0.000
		50	1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	1.000	0.000
		100	1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	1.000	0.000
	N	20	1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	1.000	0.000
			1.000	0.000	0.000	0.000	1.000	0.000
		50	0.000	0.000	1.000	0.000	0.000	1.000
			0.000	0.000	1.000	0.000	0.000	1.000
			0.000	0.000	1.000	0.000	0.000	1.000
		100	0.000	0.000	1.000	0.000	0.000	1.000
			0.000	0.000	1.000	0.000	0.000	1.000
			0.000	0.000	1.000	0.000	0.000	1.000

Values represent the posterior probabilities of each particular solution, e.g. a value of 1.000 for a single solution indicates that this is the most likely solution. The simulations were repeated using the same datasets under the four STRUCTURE models, as described in the text. The analysis was repeated three times in each condition, to assess convergence – the three rows under each condition represent these repeat runs. See the text for a description of the simulated stratification.

ized by subtracting the minimum of these three values from each, so that the best-fitting solution will score 0.00). The column labelled *M* indicates the number of markers used. When there is no stratification present, L-POP correctly identifies the *K* = 1 solution in all cases. When stratification is present, L-POP selects either a *K* = 2 or a *K* = 3 solution – in all three cases it rejects a *K* = 1 solution which is perhaps the most important feature from the point of view of correcting for stratification in tests of association (i.e. it is better to over-estimate than to under-estimate the number of classes). Pleasingly, the condition with the most markers (M = 100) gave the correct K = 2 solution.

The results from STRUCTURE are shown in table 3. The first two columns indicate which of the four STRUCTURE models was applied; for each value of M three rows are shown, representing the three repeated runs of STRUCTURE on the same dataset. The values in the final six columns indicate the posterior probability of the K = 1, K = 2 and K = 3 solutions. These are calculated from the estimated log-likelihood of the data given K which STRUCTURE estimates, $P(X|K)$, re-expressed to give $P(K|X)$ where X are the data. The first three of these columns represent the case when no stratification was generated: the models that allow correlated allele frequencies perform better in this case, correctly identifying the K = 1 solution in all cases. The models that do not allow for correlated allele frequencies, in this particular setting, give a less clear set of results (for example, consistently selecting the K = 3 solution when M = 50 and M = 100 in the no admixture model). When stratification was generated (the final three columns) then, similar to L-POP, a K = 1 solution was never selected, although sometimes a K = 3 model is selected. In some cases, despite the relatively long run times, quite different results are obtained from different runs on the same dataset, indicating problems converging. The no admixture/correlated allele frequencies model correctly selects the K = 2 model every time. However, this is not the suggested default model and one would not expect this pattern of results to hold for all datasets (e.g. when more subtle stratification and/or admixture does exist).

It is also of interest to compare how L-POP and STRUCTURE assign individuals to classes, for a specific $K > 1$ solution. For example, in the case when two strata actually are present in the data, for a K = 2 solution, how do the posterior class probabilities assigned to each individual compare between methods? Here we find almost complete agreement between the different STRUCTURE models and between L-POP and STRUCTURE. To assess this, we consider the correlations between the posterior probabilities; also, we consider agreement in terms of individuals' 'best-fit class' (i.e. determined by the maximum posterior probability). For M = 20, the different STRUCTURE models misassigned either 61 or 62 of the 500 individuals (note: it was the same 61/62 individuals who were misassigned across different methods); L-POP misassigned the same 62 individuals. The minimum correlation between the STRUCTURE methods and repeat runs was 0.996 – i.e. whatever model was applied, for a specific K, individuals were similarly classified. The minimum correlation between the STRUCTURE and L-POP posterior probabilities was 0.988. For M = 50, all STRUCTURE methods and repeat runs misassigned the same 10 individuals as L-POP; all correlations between STRUCTURE and L-POP posterior probabilities were 1.000. For M = 100 we observed perfect assignment and correlations of 1.000.

In summary, both L-POP and STRUCTURE were broadly able to distinguish between samples with no stratification versus samples with a simple two-strata pattern of stratification. The ability to select the correct $K > 1$ solution seems less well developed, although it is fair to say that this is a general problem with all classification methods in any context. For a specific K solution, however, there was near complete agreement in how L-POP and STRUCTURE assigned individuals to classes, at least in these simple cases. For more subtle, complex patterns of population substructure, we might not expect such a high degree of convergence between L-POP and STRUCTURE, or between the different models in STRUCTURE: such an analysis is beyond the scope of the present work, however.

Simulation Study

By simulating data under a number of different conditions, we aimed to explore the effect of sample size, number of marker loci and genetic distance between subpopulations on the ability to detect stratification. A number of other properties of the markers used were also varied (e.g. number of alleles and the distribution of between-subpopulation allele frequency differences). In all, thirteen different conditions were examined. In each condition, five datasets were generated, with 10, 20, 50, 100 and 200 marker loci respectively. In all cases, two models were applied to the data: K = 1 and K = 2. All results are shown in table 4. Only L-POP was used for these analyses: based on the results obtained above, we would expect many of the properties observed below to be qualitatively similar for STRUCTURE.

The 'Original' condition simulated 2,000 individuals from two subpopulations (P_1 and P_2, with 1,000 individuals from each). All marker loci were diallelic, with an average allele frequency of 0.5 and an average between-subpopulation allele frequency difference (δ) of 0.2. For all markers, one allele had a frequency of 0.4 in P_1 and 0.6 in P_2. The results for all simulations reported in this section are given in table 4. A two-class solution is correctly

Table 4. Detection of population stratification: simulation results

M	Δ_{AIC}	$P(C)$	*Corr*	PP_1	PP_2	Δ_{AIC}	$P(C)$	*Corr*	PP_1	PP_2
	‘Original’: $\delta = 0.2$; $N = 1{,}000 + 1{,}000$					‘Small’: $\delta = 0.2$; $N = 100 + 100$				
10	409.42	0.557	0.817	0.314	0.799	29.07	0.485	0.785	0.250	0.719
20	1,157.10	0.491	0.893	0.145	0.837	110.22	0.483	0.860	0.151	0.814
50	5,484.41	0.501	0.979	0.032	0.971	447.65	0.488	0.960	0.034	0.941
100	13,359.38	0.502	0.996	0.006	0.998	1,253.27	0.501	1.000	0.001	0.999
200	28,984.38	0.500	1.000	0.000	0.999	2,855.87	0.500	1.000	0.000	1.000
	‘Delta’: $\delta = 0.1$; $N = 1{,}000 + 1{,}000$					‘Delta-Small’: $\delta = 0.1$; $N = 100 + 100$				
10	37.07	0.636	0.648	0.538	0.735	–0.97	0.918	0.500	0.907	0.929
20	72.66	0.480	0.720	0.341	0.620	2.75	0.830	0.585	0.756	0.904
50	594.97	0.527	0.847	0.192	0.752	16.46	0.487	0.800	0.204	0.769
100	1,898.15	0.506	0.911	0.115	0.872	83.40	0.484	0.875	0.106	0.862
200	5,414.64	0.500	0.979	0.030	0.969	357.87	0.517	0.950	0.065	0.969
	‘Unequal’: $\delta = 0.2$; $N = 250 + 1{,}750$					‘Absolute’:$\delta = 0.2$; $N = 1{,}000 + 1{,}000$				
10	134.81	0.112	0.432*	0.554	0.935	578.91	0.477	0.833	0.209	0.745
20	473.36	0.134	0.704*	0.308	0.944	1,785.51	0.498	0.915	0.119	0.876
50	2,250.96	0.129	0.964*	0.012	0.945	6,814.84	0.502	0.986	0.023	0.980
100	5,589.78	0.125	0.996*	0.000	0.994	16,594.40	0.500	1.000	0.000	0.999
200	12,168.17	0.125	1.000*	0.000	0.999	34,974.14	0.500	1.000	0.000	1.000
	‘Split1’: $\delta = 0.4$, 0.0 (mean $\delta = 0.2$); $N = 1{,}000 + 1{,}000$					‘Split2’: $\delta = 0.8$, 0.0 (mean $\delta = 0.2$); $N = 1{,}000 + 1{,}000$				
10	1,341.61	0.515	0.903	0.151	0.878	4,004.12	0.495	0.986	0.016	0.973
20	4,039.44	0.504	0.965	0.050	0.958	12,107.36	0.499	0.998	0.001	0.998
50	13,334.15	0.499	0.998	0.002	0.997	33,266.78	0.500	1.000	0.000	1.000
100	29,769.10	0.500	1.000	0.000	0.999	71,147.01	0.500	1.000	0.000	1.000
200	63,487.39	0.500	1.000	0.000	1.000	144,108.03	0.500	1.000	0.000	1.000
	‘Multi’: $F_{ST} \approx 0.04$; $N = 1{,}000 + 1{,}000$					‘Multi-Absolute’: $F_{ST} \approx 0.04$; $N = 1{,}000 + 1{,}000$				
10	26.00	0.542	0.652	0.443	0.641	19,303.76	0.500	1.000	0.000	1.000
20	234.01	0.490	0.763	0.287	0.692	41,693.66	0.500	1.000	0.000	1.000
50	1,053.92	0.501	0.871	0.167	0.835	108,478.78	0.500	1.000	0.000	1.000
100	2,956.16	0.501	0.953	0.068	0.934	220,230.66	0.500	1.000	0.000	1.000
200	8,098.05	0.500	0.992	0.012	0.988	443,559.82	0.500	1.000	0.000	1.000
	‘Multi-Split’: $N = 1{,}000 + 1{,}000$									
10	2,033.58	0.501	0.945	0.066	0.936					
20	2,091.13	0.499	0.948	0.064	0.934					
50	2,012.31	0.496	0.950	0.052	0.939					
100	1,874.20	0.509	0.946	0.071	0.947					
200	1,586.13	0.490	0.948	0.045	0.934					
	‘Null’: $\delta = 0.0$; $N = 1{,}000 + 1{,}000$					‘Null-Small’: $\delta = 0.0$; $N = 100 + 100$				
10	2.51	0.989		0.991	0.987	0.77	0.175		0.174	0.176
20	–16.49	0.513		0.506	0.519	5.70	0.874		0.853	0.895
50	4.20	0.045		0.043	0.046	0.13	0.818		0.794	0.842
100	8.88	0.034		0.038	0.031	–31.05	0.551		0.553	0.550
200	–21.66	0.142		0.144	0.139	–48.94	0.532		0.537	0.526

The 13 conditions are described in the text. Each table shows results for different number of markers (M): the difference in AIC (Δ_{AIC}) between a $K = 1$ and a $K = 2$ solution (such that positive values indicate a $K = 2$ solution); the prior class probability ($P(C)$); the proportion of the sample correctly assigned under a $K = 2$ solution (*Corr*); the average posterior probability of belonging to a specific class for individuals who truly belong to the first stratum (PP_1) and individuals who truly belong to the second stratum (PP_2). See the text for more details.

* In the Unequal condition, Corr represents the proportion of the minority subpopulation correctly assigned (the values would be artificially high if the usual definition of correct assignment were used).

favoured in all five conditions (the first column M is the number of marker loci). The columns labeled $\Delta_{AIC} = AIC(K = 1) - AIC(K = 2)$, so a positive value is evidence for a two-class solution over a one-class solution. Even with only 10 markers, this difference is large (409.42). The final four columns refer to parameter estimates under the $K = 2$ solution. The $P(C)$ column gives the prior class probability for class '1' – in all cases this value is near 0.5 (i.e. as the two classes were simulated at equal frequencies), but the estimate increases in precision with increasing number of markers. The *Corr* column gives the proportion of individuals correctly classified on a highest posterior probability basis.

For the 'Original' condition, the classification rate rises from round 80 to 100% as the number of markers increases. That is, although a two-class solution is favoured with only 10 markers, the accuracy of the classification is not perfect. However, for such a small number of markers, arguably 80% accuracy is acceptable. The columns labelled PP_1 and PP_2 give the average posterior probability for belonging to class '1' for individuals from subpopulations '1' and '2' respectively. Perfect classification would correspond to one of these values being 0 and the other being 1. No classification would correspond to both values equalling the prior probability for class '1'. (Note that the values have been ordered such that the smaller value always corresponds to P_1 – in practice, whether or not estimated class '1' corresponds to P_1 or P_2 is random and arbitrary.) As can be seen, with increasing number of markers, the separation between the two classes increases – by 100 markers, the classification is almost perfect.

The 'Small' condition was similar to the 'Original' condition, except only 100 individuals from each subpopulation were generated. Although a two-class solution is favoured in all cases, the difference in AIC has dropped considerably. However, the accuracy of classification has remained approximately equal to the 'Original' condition. (Note that with the smaller sample size, the precision of the classification estimates themselves will be lower).

The 'Delta' condition reduces the genetic distance between the two groups, making them less distinct and therefore harder to separate. In this condition, the δ value is 0.1 instead of 0.2 (i.e. all markers are simulated using an allele frequency of 0.45 in the first subpopulation and 0.55 in the second). This leads to a reduction in the AIC difference, although a two-class solution is still consistently favoured. However, the classification ability of the model also drops under this condition. For example, with only 10 markers, the prior probability of class '1' is 0.636 (i.e. it should be 0.5); the posterior probabilities are both above 0.5 for P_1 and P_2 (i.e. one should be near 0, the other near 1). Under these conditions, around 200 markers are required before classification becomes near-perfect.

The next condition combines the 'Small' and 'Delta' conditions. Here, the evidence for the two-class solution is greatly attenuated, especially with a smaller number of markers. With only 10 markers, the model favours a one-class solution, and shows no evidence of classifying individuals correctly (it performs at chance).

Summarising the last four conditions, it is clear that small sample size has an extra deleterious effect when conditions are poor to begin with. That is, the small sample size represents 10% of the large sample size. When $\delta = 0.2$, the small-sample Δ_{AIC} is also approximately 10% of the large-sample Δ_{AIC}. For example, for 20 and 200 markers, it is 9.51 and 9.85%, respectively (i.e. 110.22/1,157.10 and 2,855.87/28,984.38, respectively). However, when the genetic distance between groups is smaller (i.e. $\delta = 0.1$), then the evidence for stratification is proportionally less in the small sample compared to the large sample: the small-sample Δ_{AIC} is only 3.77% and 6.59% of the large-sample value, for 20 and 200 markers, respectively.

However, sample size and average δ value are not the only variables which impact on the model's ability to detect stratification and classify individuals. In the 'Unequal' condition, the subpopulations were simulated with unequal mixing proportions such that one class formed a minority, the other a majority, rather than the 50:50 balance previously used. In this condition, although the overall sample size was held constant (2,000), 250 individuals were simulated from the first class, 1,750 individuals were simulated from the second. Compared to the 'Original' condition, there has been some reduction in the Δ_{AIC} values, and the classification ability has been affected also. The values in the *Corr* column represent the proportion of the minority subpopulation correctly assigned (the values would be artificially high, if the usual definition of correct assignment were used).

The next 'Absolute' condition investigated the effect of average allele frequency: keeping δ fixed at 0.2, the allele frequencies were simulated at 0.2 and 0.4 for the two subpopulations rather than 0.4 and 0.6. There does not appear to be any great effect of absolute allele frequency, compared to the 'Original' condition, at least under these conditions. However, this does not address the issue of whether or not rare alleles are, in practice, more or less likely to show differences between different ethnic groups.

Table 5. Allele frequencies used for 'Multi' and 'Multi-Absolute' scenarios

Allele	'Multi'		'Multi-Absolute'	
	P_1 freq.	P_2 freq.	P_1 freq.	P_2 freq.
1	0.36000	0.14000	0.2305	0.0000
2	0.19625	0.30375	0.3695	0.4000
3	0.30375	0.19625	0.4000	0.3695
4	0.14000	0.36000	0.0000	0.2305

Rare alleles are subject to greater fluctuation in frequency due to genetic drift than common alleles, and so may be expected to show greater between-population differences. In fact, recent studies looking at SNP frequencies in different races have concluded that less frequent SNPs are more likely to be specific to one or two races [19, 20].

Of course, the average δ value across a set of markers does not capture all the information about allele frequency differences between two groups. In this 'Split1' and the subsequent 'Split2' conditions, the impact of the distribution of frequency differences was examined, whilst keeping the average δ value constant. In the 'Split1' condition, half the markers were simulated to show no difference between groups (i.e. $\delta = 0$, both groups simulated using 0.5 allele frequency) and half the markers were simulated using as exaggerated allele frequency difference ($\delta = 0.4$, groups simulated using 0.7 and 0.3 allele frequencies). In this way, the average between group distance was still $\delta = 0.2$. We find that the pattern of allele frequency differences gives greater power to detect stratification and better classification rates also. With only 50 markers (25 of which show no between-subpopulation differences) near-perfect classification can be achieved.

The 'Split2' condition represents a more extreme version of the 'Split1' condition. Rather than splitting the markers into two equal-sized groups, three-quarters of them were set to show no differences with only the remaining quarter showing an increased δ of 0.8 (i.e. allele frequencies 0.1 and 0.9). (For the 10-marker condition, 2 markers had $\delta = 0.8$, one marker $\delta = 0.4$ and seven markers $\delta = 0$; similarly for the 50-marker condition). In this way, the average δ value is still 0.2 in all conditions. This more extreme split results in even better ability to detect and characterise subpopulation structure, despite the fact that the majority of loci do not show any allele frequency differences between groups at all. This means that a few well-selected markers with large between-group variation might be all that are needed to accurately distinguish between the major ethnic groups.

All previous simulations have been for a diallelic locus: the method is equally applicable to mutli-allelic markers however. A δ value of 0.2 corresponds to $F_{ST} = 0.04$ when the average allele frequency is 0.5. That is, the average expected heterozygosity within each subpopulation is $(1 - 0.6^2 - 0.4^2) + (1 - 0.4^2 - 0.6^2)/2 = 0.48$ and the expected heterozygosity across all populations based on the average allele frequencies is $1 - 0.5^2 - 0.5^2 = 0.50$ and so $F_{ST} = (0.50 - 0.48)/0.50 = 0.04$. In this 'Multi' condition, the performance of using multi-allelic markers with comparable F_{ST} values was examined. For two populations, the allele frequency values shown in table 5 were used to simulate the markers for the two subpopulations, to give a F_{ST} value of approximately 0.04. Performance is worse under these conditions, especially for smaller numbers of markers, presumably due to the average between-population allele frequency differences being smaller.

In the 'Multi-Absolute' condition, a different set of allele frequencies were employed but with a similar F_{ST} value (0.04). The critical factor in this condition is that some of the subpopulation-specific allele frequencies were set to 0 (also shown in table 5). This condition shows a markedly different set of results: there is a massive increase in the ability to select a two-class solution and classification is essentially perfect with only 10 markers. In the 'Multi' condition the average difference in allele frequency between subpopulations was 0.16375; in this 'Multi-Absolute' condition the average difference is even smaller, only 0.1305. It would appear that the presence of allele frequencies of 0 allows the model to easily distinguish between classes.

In this 'Multi-Split' condition, the three different types of multi-allelic marker used above are combined. Markers with allele frequencies corresponding to the 'Multi-Absolute' condition are labeled 'Type I' markers. Markers with allele frequencies corresponding to the 'Multi' condition are labelled 'Type II'. Finally, markers with four equifrequent alleles (i.e. 0.25 in both subpopulations) are labelled 'Type III' markers.

In all of the five marker number conditions, only 2 markers are of type I, 2 are of type II and the remaining $M - 4$ are of type III. That is, unlike all previous scenarios, where we would expect increasing information with increasing M, only the uninformative marker count rises with M in this condition. In all five cases, there are only four out of M markers that show any difference between subpopulations: when $M = 200$ there are 196 markers which should not contribute anything except noise.

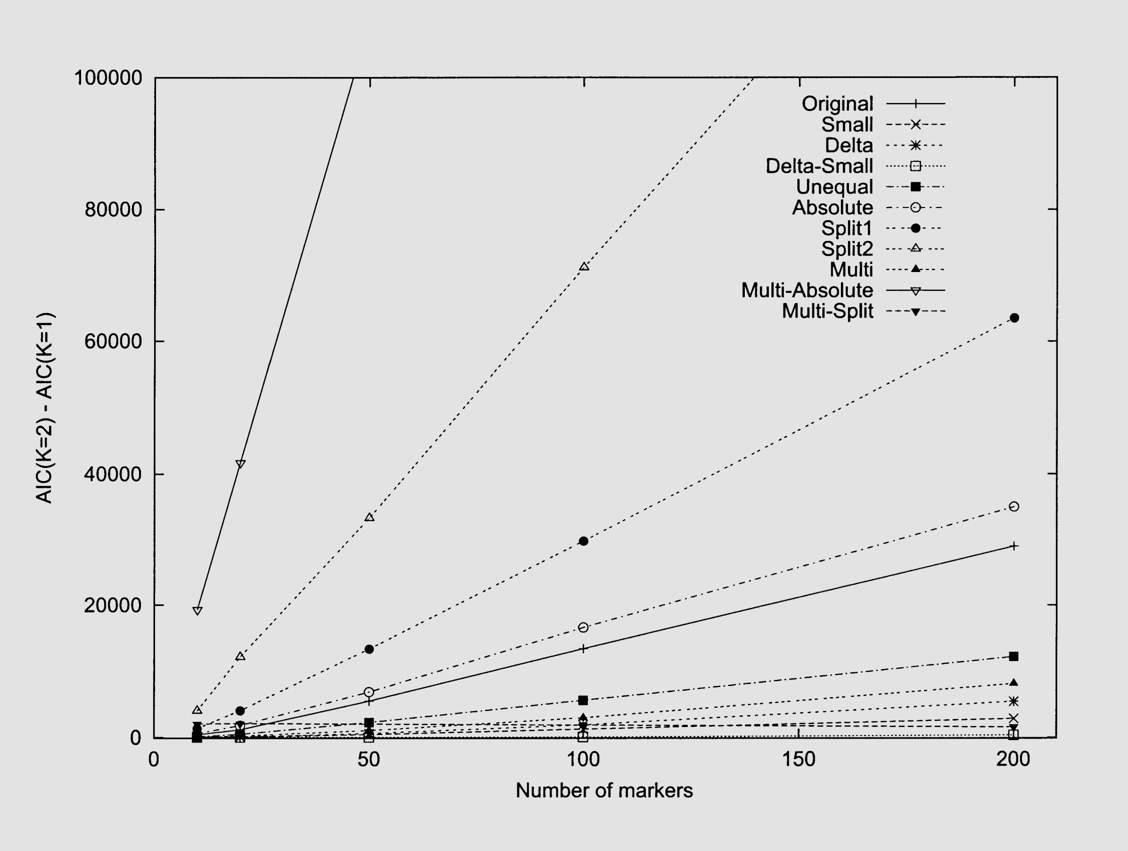

Fig. 1. Simulations result: Δ_{AIC} for different models.

In all cases a two-class solution is selected, with large AIC differences, although this decreases with increasing *M*. The classification ability of the model remains roughly constant over the different *M*, with a *Correct* value of around 0.95 and posterior probabilities around 0.05 and 0.95. This performance is roughly equivalent to the 'Original' condition with 50 markers – despite the fact that only four markers will be contributing to the solution.

Utilising the diagnostic output features of L-POP (see Implementation section below), the inter-class locus-specific genetic distances are tabulated for $M = 10$, showing clearly the relative contribution to the solution (table 6).

The final two conditions examine performance under the null – that is, when there are no true allele frequency differences between subpopulations at any of the markers. Although the two groups are simulated separately, there are no genetic differences between them, so one would expect a single-class solution in all cases. The results show that this is not necessarily the case, however. In fact, in

Table 6. Inter-class locus-specific genetic distances for multiallelic scenarios when $M = 10$

Locus	Type	Inter-class locus variation
1	I	0.6193
2	I	0.6177
3	II	0.0721
4	II	0.0965
5	III	0.0121
6	III	0.0139
7	III	0.0215
8	III	0.0301
9	III	0.0241
10	III	0.0324

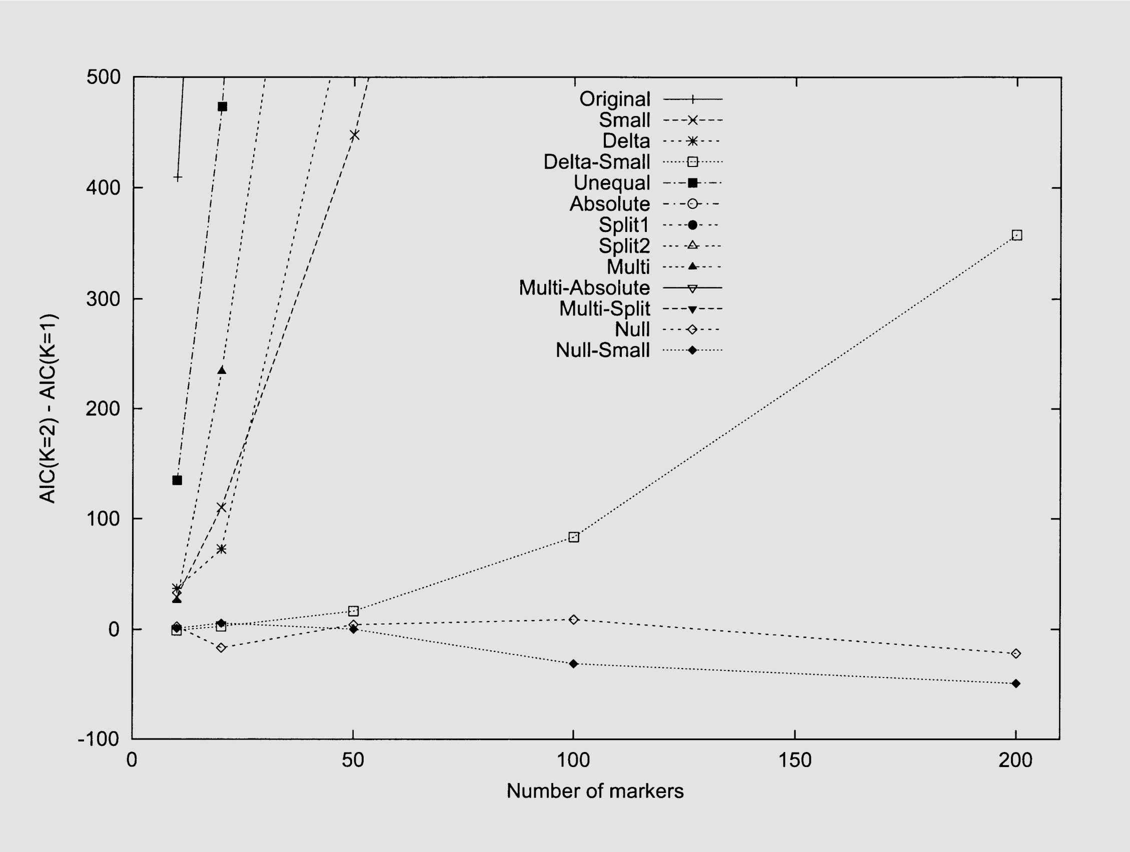

Fig. 2. Simulation results: Δ_{AIC} for different models, reduced scale.

three out of the five simulations, there was evidence for a two-class solution, although this was quite slight, compared to the previous AIC differences obtained. As expected, the $P(C)$ values (calculated under a two-class solution) are quite meaningless – what is significant is that the P_1 and P_2 posterior probability values are both similar to this value. (Of course, in practice, one would not be aware of the P_1 vs. P_2 distinction.) When the model favours a two-class solution, it appears that one of the classes is very small (around 2–4% of the sample). This suggests that the AIC may have a tendency to over-estimate the true value of K. Further work will be required to investigate the conditions under which the null model is retained, and also the possible use of metrics other than AIC to evaluate model-fit. The final 'Null-Small' condition simulates under the null but with the smaller ($N = 200$) sample size. Results appear to be similar to the 'Null' condition above.

Summary Results of Simulations

Figure 1 plots the AIC difference for the 13 different conditions as a function of M. In all cases, the AIC difference appears to increase with increasing number of markers in a roughly linear manner. The 'Split' conditions resulted in increased ability to detect the two-class solution; using multi-allelic markers with allele frequencies as in the 'Multi-Absolute' condition had the greatest impact (note: the line goes off the scale). As expected, decreasing the number of markers, genetic distance between groups and sample size all result in reduced AIC differences.

Figure 2 plots the same information, but changes the scale of the Y-axis as appropriate for the conditions with little or no AIC difference. Although there is a trend for the AIC difference to become negative under the 'Null' conditions (and therefore represent a $K = 1$ solution) this is not as striking as the performance under the alternative. This plot also shows the poor performance of the 'Delta-Small' condition with fewer than 100 markers.

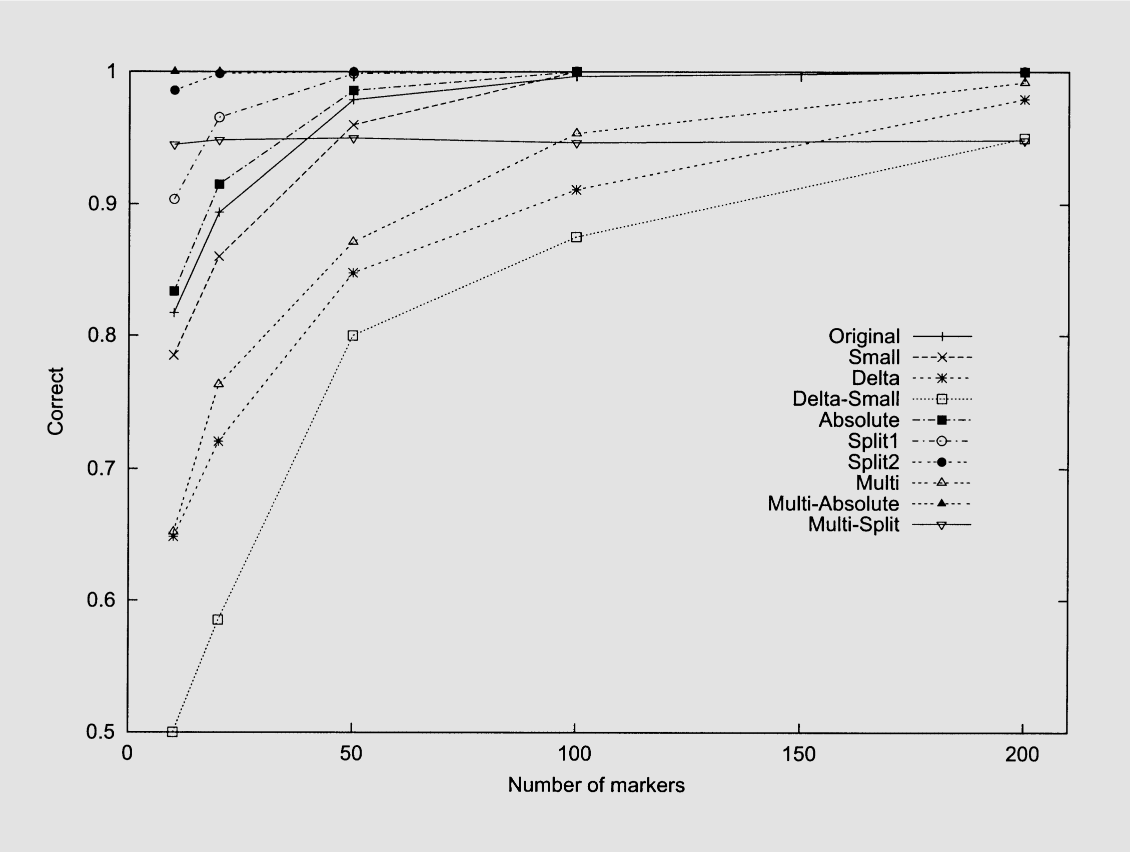

Fig. 3. Simulation results: Proportion of correct classification *(Corr)* for different models.

Finally, figure 3 plots the classification accuracy rate for the different conditions by increasing M. Note that the line for the 'Multi-Split' condition is flat, as expected, as the number of informative markers does not increase with M. In most conditions, performance is acceptable with around 50 markers (above 95% accuracy) and near-perfect with around 200 markers.

Discussion

The new genetic background methods (genomic control and structured association) still require a fully comprehensive evaluation of power issues. Bacanu et al. [21] found that in the absence of stratification genomic control approaches are more powerful, especially with common diseases. However, in the presence of stratification the results are more complex. Overall, it seems that these methods can work with as few as 20 loci – this figure seems roughly supported by the present simulations.

Unlike family-based methods, which logically control for stratification, genetic background methods only probabilistically control for stratification. That is, whether or not the stratification is correctly detected in the first place is subject to certain power constraints. Cardon et al. [22] point out that the overall type I error rate can still be inflated (doubled) at low levels of stratification (i.e. when the power of the genomic control method is significantly less than 100%). These results suggest that genetic background methods do not provide absolute protection against stratification.

However, one potential advantage of the structured association approach is that cluster-membership can become a variable in analysis to do more than just control for stratification effects. For example, it is possible to look for cluster-based interaction effects that might represent $G \times E$ or allelic heterogeneity (i.e. E is indexed by cluster).

A very important issue is the selection of an optimal marker set – several studies have begun to look at divergence in allele frequency for many markers between the major ethnic groups [19, 20]. Using these markers would be preferable for two reasons: first, they provide the greatest discriminatory ability due to the greater divergence in allele frequencies; second, as the allele frequencies are well-estimated for major ethnic groups, it would be possible to create pseudo-classes that have class-specific allele frequencies fixed to these values. In this way, it might become apparent, for instance, that a large proportion of a sample is an admixture between Caucasian and African-American ancestry despite the fact that there are no pure African-American individuals in the sample. That is, currently, to detect a class as admixed, the ancestral classes must exist in the sample in the pure form also. Online resources such as the ALFRED database project should assist this effort [23]. The simulation results support the notion that a handful of well-chosen markers may provide much more discriminatory power than hundreds of randomly-selected markers.

Implementation

The methods described above have been implemented in the computer program L-POP[1]. The program can handle missing genotypic data, autosomal and X chromosome markers and haploid organisms. Posterior probabilities can be estimated for each individual; alternatively, individuals can be fixed to belong to a particular class. Options to specify admixed solutions, relax certain assumptions and calculate the diagnostic measures mentioned above are incorporated.

The software L-ASSOC was developed to implement the ML method for assessing association in structured populations: it can perform a test of association (additive and dominance effects) controlling for potential substructure is specified; ignoring substructure; allowing different magnitudes of QTL effect between class; testing only for homogeneity of allele frequency and/or effect between classes.

Appendix 1: The Basic E-M Algorithm

The count I of individuals in class j of K is obtained by summing over all i individuals

$$I(C=j) = \sum_i P(C=j\,|\,G_i).$$

The allele counts A for each class are calculated in an analogous fashion

$$A(G_l = k\,|\,C=j) = \sum_i P(C=j\,|\,G_i)(D_{i1} + D_{i2})$$

where, for nonmissing allele data, D_{i1} is 1 if individual i's first allele at locus l is k and otherwise 0; D_{i2} is similarly defined for the individual's second allele. For missing data at a locus, values are imputed into D_{i1} and D_{i2} for each possible allele k to represent the probability of that allele occurring in that individual, which equals $P(C = j\,|\,G_i)P(G_l = K\,|\,C = j)$ where $P(G_l = k\,|\,C = j)$ is the estimated class-specific allele frequency from the previous E-M iteration (or the starting values on the first iteration). The prior class probabilities are then

$$P(C=j) = \frac{I(C=j)}{N}$$

whilst class-specific allele frequencies are

$$P(G_l = k\,|\,C=j) = \frac{A(G_l = k\,|\,C=j)}{2I(C=j)}$$

as class j contains $2I(C=j)$ chromosomes.

In estimating $P(C\,|\,G)$, the probability of observing individual i is first calculated

$$P(G_i) = \sum_j P(C=j) \prod_l \tau P(G_l = k_{i1}\,|\,C=j)P(G_l = k_{i2}\,|\,C=j)$$

where k_{i1} and k_{i2} are the two alleles at locus l and $\tau = 1$ if $k_{i1} = k_{i2}$ (i.e.homozygous genotype) or $\tau = 2$ if $k_{i1} \neq k_{i2}$ (i.e. heterozygous genotype). To handle missing data, $P(G_l = \mathit{missing}\,|\,C = j)$ is defined as 1 and so will not contribute to the product term. It is this step that defines the intra-class properties of Hardy-Weinberg and linkage equilibrium: within each subpopulation all alleles are assumed to occur independently within and across loci. Summing over all classes weighted by the prior class probability then gives the overall likelihood of observing that individual, $P(G_i)$. For individual i the posterior probability of belonging to class j is

$$P(C=j\,|\,G_i) = \frac{P(C=j)\,\Pi_l\,\tau P(G_l = k_{i1}\,|\,C=j)P(G_l = k_{i2}\,|\,C=j)}{\Sigma_{j'}\,P(C=j')\,\Pi_l\,\tau P(G_l = k_{i1}\,|\,C=j')P(G_l = k_{i2}\,|\,C=j')}$$

whilst the sample log-likelihood on E-M iteration n is $\lambda_n = \Sigma_i \ln P(G_i)$. The E-M algorithm converges if $|\lambda_n - \lambda_{n-1}|$ falls below some arbitrary tolerance value. Otherwise, returning to the E-Step, $P(G\,|\,C)$ and $P(C)$ are recounted on the basis of the newly-revised estimates of $P(C\,|\,G)$.

[1] L-POP is available for download from http://statgen.iop.kcl.ac.uk/lpop/.

Appendix 2: Admixed Classes

We calculate the expected contribution from ancestral class a of allele k conditional on the data and current model estimates as the unit to be used in the ancestral class allele count. For individual i, considering allele k at locus l, we can calculate the expected contribution from ancestral class a

$$\alpha(G_l = k \mid C_A = a, G_i)$$

$$= \sum_d P(C_D = d \mid G_i) \left[\frac{[\Theta]_{da} P(G_l = k \mid C_A = a)}{\Sigma_{a'} [\Theta]_{da'} P(G_l = k \mid C_A = a')} \right]$$

where the first sum is over d derived classes. (Note that if $\Sigma_{a'} [\Theta]_{da'} P(G_l = k \mid C_A = a')$ equals zero, it can be set to any nonzero number without adverse effect, to avoid computational problems.)

The sample contribution to the allele k count for ancestral class a is therefore

$$A(G_l = k \mid C_A = a) = \sum_i \alpha(G_l = k \mid C_A = a, G_i)(D_{i1} + D_{i2})$$

which is equivalent to the original allele-counting formula in the case of a 'pure' derived class where $[\Theta]_{da}$ takes only 1 or 0 values as $\alpha(G_l = k \mid C_A = a)$ will only ever equal $P(C_D = d \mid G_i)$ or 0.

Missing data have to be handled slightly differently, however. The contribution to the ancestral class count a for each possible allele k is calculated by summing over all derived classes d

$$A(G_l = k \mid C_A = a) = \sum_i \left(\sum_d \alpha(G_l = k \mid C_A = a, G_i)[\Theta]_{da} \right) (D_{i1} + D_{i2})$$

where, as before, for missing alleles $D_{i1} = D_{i2} = P(C_D = d \mid G_i)P(G_l = k \mid C_D = d)$.

The ancestral class individual counts can be calculated by simply summing over all the k allele counts for any one locus l $I(C_A = a) = \Sigma_i \Sigma_k A(G_l = k \mid C_A = a)/2$. Having counted the number of individuals and alleles in each ancestral class, we calculate the allele frequencies in the ancestral classes $P(G_l = k \mid C_A = a) = A(G_l = k \mid C_A = a)/2I(C_A = a)$ and then finally the derived class allele frequencies which are simply weighted sums of the constituent ancestral class allele frequencies $P(G_l = k \mid C_d = d) = \Sigma_a P(G_l = k \mid C_A = a)[\Theta]_{da}$. Having calculated the derived class prior probabilities $P(C_D)$ and allele frequencies $P(G \mid C_D)$, the M-step proceeds, for derived classes only, as in the no-admixture case described above.

References

1 Wright S: The genetical structure of populations. Ann Eugenics 1951;15:323–354.

2 Li CC: Population subdivision with respect to multiple alleles. Ann Hum Genet 1969;33:23–29.

3 Devlin B, Roeder K: Genomic control for association studies. Biometrics 1999;55:997–1004.

4 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959.

5 Pritchard JK, Donnelly P: Case-control studies of association in structured or admixed populations. Theor Popul Biol 2001;60:227–237.

6 Satten GA, Flanders D, Yang Q: Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 2001;68:466–477.

7 Lazarsfeld PF, Henry NW (eds): Latent structure analysis. Boston, Houghton Mifflin, 1968.

8 Rosenberg NA, Li LM, Ward R, Pritchard JK: Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 2003;73: 1402–1422.

9 Marchini JJ, Cardon L, Phillips M, Donnelly P: The effects of human population structure on large genetic association studies. Nat Genet 2004;36:512–517.

10 Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel N, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D: Assessing the impact of population stratification on genetic association studies. Nat Genet 2004; 36:388–393.

11 Dempster AP, Laird NM, Rubin DB: Maximum likelihood for incomplete data via the EM algorithm. J R Stat Soc B 1977;39:1–38.

12 Akaike H: A new look at the statistical model identification. IEEE Trans Automatic Control, AC 1974;19:716–723.

13 Fisher SA, Lewis CM: Methods to identify population outliers using genetic markers. GeneScreen 2001;1:125–129.

14 Nei M: Molecular Evolutionary Genetics. New York, Columbia University Press, 1987.

15 Hubert L, Arabie P: Comparing partitions. J Classif 1985;2:193–218.

16 Ewens WJ, Shute NCE: A resolution of the ascertainment sampling problem. Theor Popul Biol 1986;30:388–412.

17 Sham PC, Zhao JH, Cherny SS, Hewitt JK: Variance components qtl linkage analysis: Conditioning on trait values. Genet Epidemiol 2000;19(suppl 1):S22–S28.

18 Purcell S: Sample Selection and Complex Effects in Quantitative Trait Loci Analysis. PhD thesis, King's College London, London University, 2003.

19 Cargil M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al: Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 1999;22:239–247.

20 Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A: Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 1999;22:239–247.

21 Bacanu SA, Devlin B, Roeder K: The power of genomic control. Am J Hum Genet 2000;66: 1933–1944.

22 Cardon LR, Bell JI: Association study designs for complex diseases. Nat Rev Genet 2001;2: 91–99.

23 Cheung KH, Miller PL, Kidd JR, Kidd KK, Osier MV, Pakstis AJ: ALFRED: A Web-accessible allele frequency database. Pac Symp Biocomput 2000:639–650.