Nucleic Acids Research, 2006, Vol. 34, No. 7 ***e55***
doi:10.1093/nar/gkl136

Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates

Stuart Macgregor*, Peter M. Visscher and Grant Montgomery

Genetic Epidemiology, Queensland Institute of Medical Research, Brisbane, Australia

Received January 13, 2006; Revised February 14, 2006; Accepted March 14, 2006

ABSTRACT

Array based DNA pooling techniques facilitate genome-wide scale genotyping of large samples. We describe a structured analysis method for pooled data using internal replication information in large scale genotyping sets. The method takes advantage of information from single nucleotide polymorphisms (SNPs) typed in parallel on a high density array to construct a test statistic with desirable statistical properties. We utilize a general linear model to appropriately account for the structured multiple measurements available with array data. The method does not require the use of additional arrays for the estimation of unequal hybridization rates and hence scales readily to accommodate arrays with several hundred thousand SNPs. Tests for differences between cases and controls can be conducted with very few arrays. We demonstrate the method on 384 endometriosis cases and controls, typed using Affymetrix Genechip© HindIII 50 K arrays. For a subset of this data there were accurate measures of hybridization rates available. Assuming equal hybridization rates is shown to have a negligible effect upon the results. With a total of only six arrays, the method extracted one-third of the information (in terms of equivalent sample size) available with individual genotyping (requiring 768 arrays). With 20 arrays (10 for cases, 10 for controls), over half of the information could be extracted from this sample.

INTRODUCTION

Genome-wide genetic association analysis is set to become one of the primary tools for the identification of loci contributing to susceptibility to complex common human disease. However, the cost remains prohibitively expensive for many projects. Genome scans of suitable size (hundreds of cases/controls, hundreds of thousands of markers) typically cost well over US$1 million. Instead of genotyping the large numbers of markers [typically single nucleotide polymorphisms or (SNPs)] in individual samples on DNA microarrays, a number of authors have proposed pooling the DNA from large numbers of individuals (1–3). The pooled DNA is hybridized to arrays, such as the Affymetrix Genechip© array (4) and the allele frequencies estimated in each pool. In practice, the primary interest is in tests of the difference in allele frequency between the case pool and the control pool. Whilst pooling offers a substantial reduction in genotyping cost, naive tests derived from DNA pool allele frequency estimates have undesirable statistical properties (5). A more appropriate test can be derived by recognizing that DNA pools yield estimated allele counts rather than observed counts. Essentially, the additional variance generated by pooling specific errors must be appropriately taken into account.

We propose a method for analysis of large scale pooling data which utilizes the information available across multiple SNPs to estimate the errors inherent in pooling. By utilizing the information from multiple SNPs we are able to estimate the variance associated with pooling. This allows us to construct a statistical test for association with desirable properties. Moreover, since array data will typically have a regular structure (in terms of multiple measurements per SNP on the array), simple tests (such as *t*-tests) which ignore this structure will be unsatisfactory. We propose the use of general linear model based tests which take into account the structure of the array data. Since the error variance associated with pooling is estimated across SNPs, the need for replication of pools is minimized, thereby decreasing cost. The method does not require prior information on the value of k (a measure of the extent of unequal amplification/hybridization of alleles) and hence avoids the need for expensive individual genotyping of heterozygotes for every SNP of interest. Therefore our method

*To whom correspondence should be addressed at Queensland Institute of Medical Research, Post Office, Royal Brisbane Hospital, 300 Herston Road, Brisbane 4029, Australia. Tel: +61 7 3845 3563; Fax: +61 7 3362 0101; Email: stuart.macgregor@qimr.edu.au

easily scales up to arrays with hundreds of thousands to millions of SNPs. The new method is applied to data on a set of 384 cases and controls from a study on endometriosis (6–8) typed with the Affymetrix Genechip© HindIII array (4). For a subset of this data there were accurate measures of k available. We show that assuming $k = 1$ has a negligible effect upon the results.

MATERIALS AND METHODS

Statistical methods

Pooling tests of association. In genetic association analysis the primary interest is to estimate the difference in the proportion of A alleles between case and control pools. The simplest test for this difference at a SNP involves calculating the average proportion in cases and controls and computing the test statistic.

$$T_{\text{simple}} = \frac{(\tilde{p}_a - \tilde{p}_u)^2}{\text{var}(\hat{p}_a - \hat{p}_u)} \simeq \frac{(\tilde{p}_a - \tilde{p}_u)^2}{\text{var}(\tilde{p}_a - \tilde{p}_u)} \quad \mathbf{1}$$

The population frequency in cases is denoted p_a, the pooling sample estimate of the allele frequency is denoted $\tilde{p}_a$ and the sample estimate if the sample was individually genotyped without error is denoted $\hat{p}_a$. p_u, $\tilde{p}_u$ and $\hat{p}_u$ are defined similarly for controls. Since the values of $\hat{p}_a$ and $\hat{p}_u$ are not available the sample estimates are used as an approximation in the denominator of equation 1. In the absence of errors in the estimation of $\tilde{p}_a$ and $\tilde{p}_u$, var$(\tilde{p}_a - \tilde{p}_u)$ is given by the usual formula for the binomial sampling variance, $V = p_a(1 - p_a)/2n_a + p_u(1 - p_u)/2n_u$ (or in practice $\tilde{V}$ where the V is given a $\sim$ to reflect the fact it is based on sample estimates). The number of cases and controls is n_a and n_u, respectively. T_{simple} will then have a χ^2_1 distribution (under the null hypothesis of no difference). However, in the presence of errors in the estimation of $\tilde{p}_a$ and $\tilde{p}_u$, the term var$(\tilde{p}_a - \tilde{p}_u)$ will be greater than $\tilde{V}$ and the distribution of T_{simple} will no longer be χ^2_1 (5).

Denoting the variance of the pool specific error in allele frequency estimation as var$(e_{\text{pool}-1})$, it is shown in appendix 1 that a corrected test statistic is

$$T_1 = T_{\text{simple}} \times \frac{\tilde{V}}{\tilde{V} + 2\text{var}(e_{\text{pool}-1})} \quad \mathbf{2}$$

The problem is hence to estimate var$(e_{\text{pool}-1})$. This could be estimated from replicate pools for the SNPs in question but to obviate the need for further genotyping we propose using the full set of available SNPs to estimate var$(e_{\text{pool}-1})$. Before doing that, we first describe an efficient means of estimating the difference between cases and controls with array data.

A general linear model for array data. When arrays are used for pooling there are typically multiple probe measurements available. With Affymetrix Genechip© arrays there is a measure on each strand of the DNA (strand replication), several measures (up to 7 with the 50 K chips) at different probe positions on the chip (probe replication) and, typically, multiple arrays per sample (array replication). Arrays from other manufacturers can be accommodated in our method by simple modification of the model to reflect the different structure of replicated measurements. Although t-tests can be applied to these multiple measurements, a more efficient way of dealing with this data is to explicitly model the data structure. To do this we propose fitting a general linear mixed model (GLMM). An introduction to mixed models (models with both fixed and random effects) is given in Armitage (9). In the linear model the response variable is the estimates of proportion of A alleles in cases for a given SNP; this is calculated using $p = A/(A + B)$ where A and B are measures of the fluorescent intensities for alleles A and B, respectively. Note that no correction is made for unequal hybridization of the alleles, see also Data application and Discussion sections below. Since there are multiple probe measurements there are multiple measures of p. Let p_{ijlm} denote the mth probe measure of p on strand l of replicate j in sample i. With C samples (e.g. case, control), R array replicates, S strand measures and D probe measures, and the vector p will contain up to $C \times R \times S \times D$ values per SNP. Here we consider two possible linear models; a nested (or hierarchical) model and a non-nested model. The nested model for a measure p_{ijkl} is

$$p_{ijlm} = c_i + r_j + s_{jl} + d_{ijlm}.$$

This model nests the strand measures within replicates. The predictor variables on the right hand side of the linear model are a factor for case/control status c_i, a factor for array replicate r_j, a factor for strand s_{jl} and a factor for probe position d_{ijlm}.

An alternative, non-nested model is

$$p_{ijlm} = c_i + r_j + s_l + d_m + \epsilon,$$

where ϵ is an error term and the other terms are as before. Note that by not modeling the nesting there is now scope for the estimation of a probe term and a separate error term.

For case-control data the factor c_i has two levels; case and control. We arbitrarily set 'control' to be the baseline level with 'case' a deviation from this baseline. We can hence refer to this factor as simply c, the deviation of cases from controls in the GLMM. Case/control status is treated as fixed in the linear model whilst strand, probe and array are treated as random. Estimation for the GLMM is by restricted maximum likelihood (10).

The nested model allows estimation of the contribution of the variance components to the pooling allele frequency estimate p (11). We focus our attention on just the case samples for estimation of the variance components. We chose to look only at cases here for simplicity; the results in the control sample would be expected to be similar. Hence in the nested linear model above we drop the factor for case/control status. The variance of the allele frequency estimate is decomposed as follows

$$\tilde{V}_a + \frac{\hat{\sigma}^2_r}{n_r} + \frac{\hat{\sigma}^2_s}{n_r \times n_s} + \frac{\hat{\sigma}^2_e}{n_r \times n_s \times n_e} \quad \mathbf{3}$$

where $\tilde{V}_a = \tilde{p}_a(1 - \tilde{p}_a)/2n_a$ is the binomial sampling variance in cases only, $\hat{\sigma}^2_r$ and n_r are estimated variance component and repeat count for replicate (similarly for sense and probe). To obtain estimates of the variance contributions to the difference in allele frequency estimates between cases and controls, we double the estimates for the contribution to p.

The main interest for association analysis is in the estimate $\hat{c}$ for the case/control factor. $\hat{c}$ is analogous to the estimate of $\bar{p}_a - \bar{p}_u$ from a *t*-test (where $\bar{p}_a$ and $\bar{p}_u$ denote the mean values of p in the case and control pools, respectively). If we were to drop the random effects other than the error term from the model we would recover a basic *t*-test which ignored the structure of the array data. If in practice, some of the probes fail and there are (e.g.) only probe measures available on one DNA strand for a particular SNP, the random effect for strand is dropped from the model. We show in appendix 2 that the GLMM case/control estimate c (or in the simpler *t*-test case, $\bar{p}_a - \bar{p}_u$) can be used to estimate $\mathrm{var}(e_{\mathrm{pool}-1})$. This allows construction of a test statistic with good statistical properties.

Data application

Case control sample. DNA pools were constructed from 384 endometriosis cases and 384 ethnically matched controls (8). DNA concentrations were measured using PicoGreen (Molecular Probes) for the quantitation of double-stranded DNA in solution on a Fluoroskan Ascent CF plate reader (Labsystems, Chicago). Concentrations of DNA samples were carefully adjusted by serial dilutions to a final concentration of 25 ng/μl (M SD = 25.19 ± 0.55). Individual DNA samples were tested in at least two PCR to ensure samples containing high quality DNA.

Array data. SNPs were genotyped on DNA pool samples using Affymetrix Genechip© HindIII arrays. Arrays were treated according to standard protocols (Affymetrix, San Diego). The arrays yield multiple measures of fluorescent intensity with each giving up to seven probe measures on both the sense and anti-sense strand of the DNA. In practice, the 10 best probes (from a possible $2 \times 7 = 14$ per array, counting both strands) are selected by Affymetrix for inclusion in the data supplied to end users [(12), Supplementary Data]. In some cases there were up to seven probe measures on one strand of the DNA (in this case the other strand would have a maximum of three probe measures) and this necessitated the use of seven levels in the factor for probe position (i.e. for term d in the linear model). Single pools of 384 case and 384 control samples were constructed and aliquots of each pool were hybridized to three replicate arrays. The maximum possible number of intensity measures was $5 \times 2 \times 3 = 30$ per sample (case or control).

The intensity measures consist of perfect-match/mis-match pairs for each allele. Corrected perfect-match values are calculated by subtracting the average mis-match value (across the two alleles) from the perfect-match value for each SNP. The corrected perfect-match values for each allele is used to calculate the proportion, p, of A alleles in the pool for each SNP. The intensity measures for alleles A and B are known to vary due to differential hybridization between SNPs. This is analogous to the situation where the differential amplification occurs with previous genotyping technologies; this is typically addressed by estimating k, the A:B ratio in heterozygotes (13,14). Although the unequal hybridization adversely affects allele frequency estimates in single pools the primary interest here is in the difference in frequency between case and control pools. Previous work has shown that there is a negligible effect of the changing values of k on differences in frequency in the majority of cases [(5,15,16), see also Discussion]. What is affected however, is the type I error of a naive test statistic based on the difference between pool frequencies in cases and controls. We deal with this problem in the statistical method described above. To calculate the proportion of A alleles in the pool we use $p = A/A + B$ (i.e. we assume $k = 1$). By not requiring an estimate of k from individually typed heterozygotes, our method has the potential to substantially reduce the cost of pooling experiments based on large numbers of SNPs.

A quality control step was implemented in the analysis to ensure both perfect-match intensities always exceeded the average of the mis-match intensity over the two alleles. The maximum number of p values for any SNP was 30. A small proportion of SNPs (1.3%) had less than 8 p measures available and these SNPs were removed from the analysis. Preliminary analysis showed that results based on fewer than eight intensity measures were particularly unreliable. A total of 56 494 SNPs, each with between 8 and 30 p measures, were taken forward into the full analysis.

RESULTS

The estimates of $\sqrt{\mathrm{var}(e_{\mathrm{pool}})}$ for the GLMM and *t*-test estimation methods are given in Table 1. Estimates are given on the standard deviation scale. On the variance scale the $\mathrm{var}(e_{\mathrm{pool}-1})$ is $\simeq 0.00058$ irrespective of the estimation method. Note that although the estimates of $\mathrm{var}(e_{\mathrm{pool}-1})$ in Table 1 are similar for the different estimation methods, the test statistics calculated on the basis of this value of $\mathrm{var}(e_{\mathrm{pool}-1})$ will vary because the estimate of $\hat{c}$ is different for the different estimation methods. The calculation of $\mathrm{var}(e_{\mathrm{pool}-2})$ takes into account the precision of the estimate of $\hat{c}$ for each SNP (which varies by estimation method) and so the estimates of $\mathrm{var}(e_{\mathrm{pool}-2})$ vary depending on the estimation method used. $\mathrm{Var}(e_{\mathrm{pool}-2})$ is always smaller than $\mathrm{var}(e_{\mathrm{pool}-1})$, since $\mathrm{var}(e_{\mathrm{pool}-1})$ includes the error involved in estimating the pool frequency difference (i.e. $\hat{c}$) from the available probes. The estimate of the total variance associated with pooling is either $\mathrm{var}(e_{\mathrm{pool}-1})$ (this averages over the varying precision of estimates over SNP and hence applies to all SNPs) or $\mathrm{var}(e_{\mathrm{pool}-2}) + \mathrm{var}(\hat{c}_{\mathrm{X}})$ (this is specific to a particular SNP, in this case to SNP X).

Use of the nested model allows estimation of the contribution of the different sources to the variance of the pooling allele frequency estimate. We focus here on the results for the case sample but the control sample results are similar (data not shown). With no missing data for three replicates of the

Table 1. Estimation of $\sqrt{\mathrm{var}(e_{\mathrm{pool}})}$ by estimation method

Estimation method	$\sqrt{\mathrm{var}(e_{\mathrm{pool}-1})}$	$\sqrt{\mathrm{var}(e_{\mathrm{pool}-2})}$
t-test	0.0241	0.0060
Nested GLMM	0.0241	0.0112
Non-nested GLMM	0.0239	0.0152

See appendix 1 for the definition of $\mathrm{var}(e_{\mathrm{pool}-1})$ and appendix 2 for the definition of $\mathrm{var}(e_{\mathrm{pool}-2})$. By construction $\mathrm{var}(e_{\mathrm{pool}-2})$ must be smaller than $\mathrm{var}(e_{\mathrm{pool}-1})$. $\mathrm{var}(e_{\mathrm{pool}-1})$ gives an approximate estimate of the overall variance associated with pooling; this estimate averages over all SNPs without taking into account the differing precision (i.e. number of functioning probes) available for each SNP.

HindIII array we would have $n_r = 3$, $n_s = 2$ and $n_e = 7$. With missing data variance component estimates are computed by calculating a weighted mean where the weights depend on the number of probe measures available for that SNP. The weighted mean variance component estimates for replicate, strand and probe (across all 56 494 SNPs) are given in Table 2. Also given in Table 2 is the approximate contribution of such components to the variance of the pooling allele frequency estimate. For convenience these are given both as variances and as standard deviations. The estimate of the contribution of each source of variance to the difference in allele frequency between cases and controls is twice the values given in Table 2. For example, the contribution to the difference in allele frequency from variation at the 'probe' level is $2 \times 0.00044 = 0.00088$. The contribution from variance at the 'strand' level is of similar magnitude, with the 'replicate' variance contributing relatively little variance. In practice, all three variances would be reduced by simply increasing the number of replicate arrays applied to each pool. In addition to reducing variance by increasing array replication, if it is possible to obtain arrays with larger number of probe measures per strand then this would lead to useful decreases in variance. Note that since the maximum likelihood estimation procedure cannot yield negative estimates of the variance components, the variance component estimates in Table 2 will be upwardly biased. For many SNPs the estimates for one or more of the variance component estimates was on the boundary of the parameter space (i.e. at 0). This means the upward bias may be substantial.

The results from the available tests for differences between cases and controls are given in Table 3. Although we know the *t*-test based statistics are not optimal, the results of these are given for comparison. The GLMM based statistics are computed for the nested and the non-nested case. We also computed the GLMM based statistics with the factors for strand and array replicate regarded first as fixed and then as random. Only the random effect results are given here; the results with strand and array as fixed effects are very similar. Since we are not interested in the specific values of these factors they are most reasonably modeled as random draws from a population of factor values (i.e. as random effects). We are interested in directly testing the effect of case/control status so this factor is treated as a fixed effect.

Since we expect the vast majority of the 56 494 SNPs not to be associated with disease we would expect the proportion of SNPs that reach significance at the $\alpha = 1\%$ level to be very close to 1%. Table 3 shows that the uncorrected test statistic has a grossly inflated type I error. The type I error of the corrected *t*-test based statistics is a substantial improvement compared with the uncorrected statistic but the type I error remains significantly ($P < 0.0001$) higher than the level expected by chance (under the null hypothesis of no association). The type I error of the nested GLMM (T_2 nested) is slightly above ($P = 0.01$) the level expected under the assumption of no true positives. The type I error of the non-nested GLMM (T_2 non-nested) is at the level expected under the assumption of no (or only a few) true positives.

In Table 4 the proportion of the SNPs exceeding the $\alpha = 0.1\%$ level are shown. In this case the proportion of SNPs exceeding the nominal level is slightly higher than expected for the GLMM (assuming no true positives). The 95% confidence interval (CI) on the number of SNPs exceeding the 0.1% level assuming no true associations is (43,71). The estimates of the number of SNPs reaching this level of significance (69 with the non-nested GLMM, 81 with the nested GLMM) are around the upper end of this CI, suggesting there may be a few SNPs that are truly associated in this sample.

There was considerable (but not complete) overlap between the SNPs identified by the different statistics. Of the 69 SNPs significant at the 0.1% level with the non-nested GLMM, 54 also appeared in the list of those exceeding the 0.1% level for the nested GLMM (i.e. 54 of the total set of 81 SNPs shown in Table 4). For T_1 (*t*-test based) the overlap was 52 (i.e. 52 of 162 from Table 4 overlapped). For T_2 (*t*-test based) the overlap was 46 (i.e. 46 of 91 from Table 4 overlapped). In the last two cases (*t*-test based statistics) the type I error was inflated. This means that the actual overlap for the last two cases is likely to be slightly less than the values given here.

Graphs of the observed test statistic against the test statistic expected under a χ^2_1 distribution are shown in Figure 1. This figure clearly shows the inappropriate type I error for the uncorrected and *t*-test based test statistics.

Table 2. Variance component estimates from nested model (case sample only)

	Replicate	Strand	Probe
Variance component (as variance)	0.00026	0.00268	0.01054
Variance component (as standard deviation)	0.01612	0.05176	0.10266
Contribution to variance in allele frequency estimate (as variance)	0.00009	0.00044	0.00042
Contribution to variance in allele frequency estimate (as standard deviation)	0.00931	0.02097	0.02049

The first two rows give the variance component estimates for each source (as variance and as standard deviation). These values are then inserted into Equation 3 to give the contribution to the variance in allele frequency estimate (variance/standard deviation given in last two rows).

Table 3. Test statistic comparison at 1% level

Test statistic	# SNPs exceeding 1% level	Proportion exceeding 1% level
T_{simple} (*t*-test based, uncorrected)	9370	0.16580
T_1 (*t*-test based)	734	0.01299
T_2 (*t*-test based)	666	0.01179
T_2 (nested GLMM)	620	0.01097
T_2 (non-nested GLMM)	554	0.00981

The total number of SNPs is 56 494. The number expected to exceed the 1% level under the null hypothesis of no true associations is 565.

Table 4. Test statistic comparison at 0.1% level

Test statistic	# SNPs exceeding 0.1% level	Proportion exceeding 0.1% level
T_1 (*t*-test based)	162	0.00287
T_2 (*t*-test based)	91	0.00161
T_2 (nested GLMM)	81	0.00143
T_2 (non-nested GLMM)	69	0.00122

The total number of SNPs is 56 494. The number expected to exceed the 0.1% level under the null hypothesis of no true associations is 56.

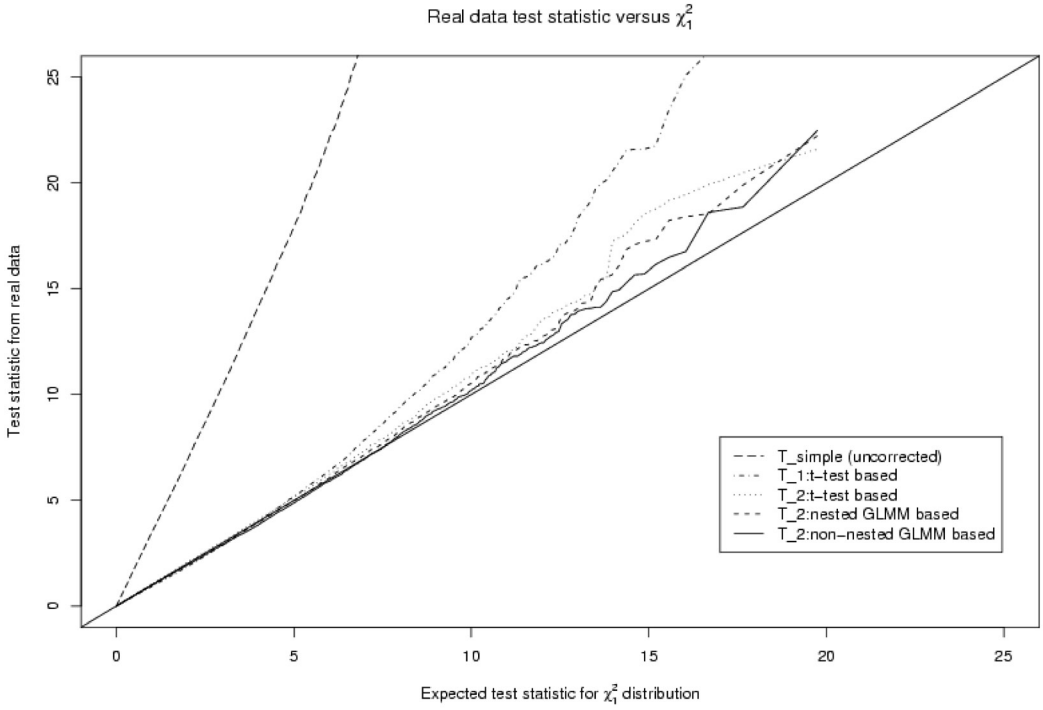

Figure 1. Comparison of test statistic performance. Results for 56 494 SNPs. Each statistic is plotted against the expected distribution under the hypothesis that there are no (or very few) true positives. A test statistic with distribution exactly equal to the expected asymptotic distribution will lie along the plotted $y = x$ line. The plotted lines exhibit considerable stochastic variation at high values (>18) because there are few data points in this range.

DISCUSSION

We have described a structured analysis method using internal replication information in large scale genotyping sets. The proposed method takes advantage of information from SNPs typed in parallel (typically on an array) to construct a test statistic with appropriate type I error. The method does not require the use of additional arrays for the estimation of unequal hybridization rates. As a result, the method can be applied with very few arrays. In our sample of 384 endometriosis cases and controls, we were able to obtain good results with a total of only six arrays.

In the optimal case, the difference between cases and controls was estimated using a GLMM. This approach differs from that taken by other authors in that the nested structure of the data are taken into account. In contrast, the Affymetrix GDAS software utilizes the median intensity score (across sense and strand measures) to evaluate the allele frequency in a pool (12). One of the main advantages of the median, compared with say the mean, is the robustness to outliers. To test whether outliers were having a large effect on the results that we obtained we recalculated the GLMM based test statistics on a dataset where the largest and smallest probe measures were removed from the data; this resulted in a dataset, which contained estimates of allele frequency differences based on between 6 and 28 probe measures. The results were very similar to those obtained on the full dataset (data not shown), indicating that outliers were not having a substantial effect on our results.

There are many disadvantages in using the median. Firstly, an analysis based on the median will not take into account the structure of the probe measurements. Secondly, the median discards information on the precise magnitude of the actual observations and typically has greatly increased sampling variance compared with the mean (9). Finally, there is no computationally rapid method of evaluating the standard error of the median. This final disadvantage is particularly unfortunate in the context of the work presented in this paper because the method we describe works best when the corrected test statistic takes into account the variable precision present when estimating the allele frequency differences from variable numbers of probes.

In the linear model we have assumed that the response variable (p) was unbounded when in reality it is bounded by 0 and 1. In practice, most of the p values of interest are in the range 0.1–0.9 (i.e. minor allele frequency or MAF > 0.1) and the bounding will not affect the results greatly. For loci with smaller MAF the model would be less appropriate and a model which explicitly dealt with frequency data (e.g. a generalized linear model with a logit link function) would give better results. In practice, for loci with small MAF, the power to detect the disease loci would be greatly decreased (compared with loci with larger MAF). This lack of power is likely to impact on results more than the inadequacy of the model for low MAF loci.

We employed two different linear models for the structure of the array data. The nested model has the advantage of allowing estimation of the different components of variance that contribute to the variance in the allele frequency estimate [see Table 2 and (11)]. The non-nested model does not explicitly model the nested structure of the data but has the advantage of allowing estimation of an overall error term in addition to terms for replicate, strand and probe (there are not

enough data points to estimate a separate error term for probe when a nested model is used). Having a separate term for probe is desirable because this allows the structure of the probe measurements to be modeled. For example, level 1 of the probe factor (which corresponds to a probe quartet with a central interrogation position on Affymetrix HindIII arrays [(12), Supplementary Data] is assumed to be the same across different replicates and strands. Similarly for factor levels 2, 3, ... (which correspond to up and downstream interrogation positions on Affymetrix HindIII arrays).

Using the nested linear model allowed us to evaluate the different sources of variation for the variance in estimate of allele frequency (Table 2). In practice the interest is often in the difference in allele frequency between cases and controls. The contribution of each of the sources to the variance of the difference in allele frequency in cases and controls is twice the values given in Table 2. That is, we can rewrite equation 3 to now represent the variance associated with the difference in allele frequency between cases and controls;

$$\tilde{V} + 2\left(\frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_s^2}{n_r \times n_s} + \frac{\hat{\sigma}_e^2}{n_r \times n_s \times n_e}\right) \quad \mathbf{4}$$

where $\tilde{V}$ is the binomial sampling term for the difference in allele frequency (i.e. not for simply the allele frequency). Evaluating $\hat{\sigma}_r^2/n_r + \hat{\sigma}_s^2/(n_r \times n_s) + \hat{\sigma}_e^2/(n_r \times n_s \times n_e)$ gives a value of ~0.00089. In the absence of errors in estimation $\hat{\sigma}_r^2/n_r + \hat{\sigma}_s^2/(n_r \times n_s) + \hat{\sigma}_e^2/(n_r \times n_s \times n_e)$ should equal $\mathrm{var}(e_{\mathrm{pool}-1})$ (to see this compare equations 4 and A1). In practice $\mathrm{var}(e_{\mathrm{pool}-1})$ yields a value of 0.00058. This indicates that the estimate of $\hat{\sigma}_r^2/n_r + \hat{\sigma}_s^2/(n_r \times n_s) + \hat{\sigma}_e^2/(n_r \times n_s \times n_e)$ has been inflated by the maximum likelihood estimation procedure. Because of the way it is calculated (from the mean of $\hat{c}^2$ values over all SNPs), the estimate of $\mathrm{var}(e_{\mathrm{pool}-1})$ is unlikely to be subject to the same biases and is likely to be a more reliable estimate.

In our experiment we generated array replicates by using replicate arrays on the same pool. An alternative approach would be to use replicate arrays on replicate pools. In the former case there would be 'technical' variance as a result of differences between arrays. In the latter case there would be both 'technical' variance and 'pooling construction' variance. When we applied the nested model to the case only sample we obtained estimates of the variance from different sources (see Table 2); the estimate for the replicate variance would only reflect the 'technical' variance and not the 'pooling construction' variance. To gain an estimate of both the 'technical' and the 'pooling construction' variance in a single sample one would need to perform an experiment with replicate arrays on replicate pools [e.g. as done by Brohede *et al*. (2)]. However, if one is only interested in the difference between cases and controls, we would expect the method described in appendix 2 to do a reasonable job of taking into account both the 'technical' variance and the 'pooling construction' variance. To see this consider the way in which $\mathrm{var}(e_{\mathrm{pool}})$ is estimated (appendix 2). The estimate of $\mathrm{var}(e_{\mathrm{pool}})$ is calculated on the basis of case-control differences and will hence include both sources of variance. This assumes that the vast majority of SNPs are not associated with disease and that for virtually all SNPs the 'cases' and 'controls' are just an independently constituted pool/sample. Since this is likely to be a reasonable assumption, the estimate we obtained for $\mathrm{var}(e_{\mathrm{pool}})$ (≃0.00058) is unlikely to be substantially inflated.

Since we expect the vast majority of SNPs not to be associated with disease in this sample, the corrected test statistics should follow a χ_1^2 distribution for most of the observed range. This occurs for the GLMM test statistics, with the only deviation occurring at the upper end of the test statistic range. Although it is difficult to reliably discriminate between true and false positives, the GLMM test statistics are consistent with there being a few real positives. To further evaluate this we added 10 'real' positives (where 10 randomly chosen points from the top decile had 10 added to their test statistic) to a set of 56 000 data points from a χ_1^2 distribution. Plotting these in a similar way to the test statistics shown in Figure 1 reveals a picture very similar to that seen for the GLMM results (data not shown).

The method described here does not use information on the difference in peak heights in a heterozygote individual (k). Although k can be estimated given data on individually genotyped heterozygotes, such a procedure adds to the cost of a pooling experiment, particularly given that our method allows users to reliably conduct a whole pooling experiment with only a few arrays per sample. In the near future association studies will be based on several hundred thousand SNPs and reliable estimation of k for every SNP will be difficult to co-ordinate. The effect of k has been considered by a number of authors (5,8,16,17). Although assuming k is 1 when it is not known to lead to biases, such biases are greatly diminished when pooling is used for case-control studies because the same error in the specification of k is made in both pools. The effect of k on power was considered in depth by Moskvina *et al*. (16). Moskvina *et al*. (16) derive a statistic, $k_{\max}$, which gives the value of k for which statistical power is maximized. The $k_{\max}$ values for a range of case (columns) and control (rows) frequencies are given in Table 5. Note that the $k_{\max}$ values are not entered in the diagonal cells of Table 5 as these correspond to the case of no difference in frequency between cases and controls. Clearly, when $k_{\max}$ is close to 1 (cells with boldface in Table 5) then assuming $k = 1$ will not lead to appreciable losses in power. The only occasions in which $k_{\max}$ deviates substantially from 1 are those when the minor allele frequency of either cases or controls is around 0.1 (cells with italic font in Table 5). Unfortunately, these cases are also the cases in which

Table 5. $k_{\max}$ values at varying frequencies in cases and controls

Frequency case/control	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1		*0.17*	0.22	0.27	0.33	0.41	**0.51**	**0.67**	**1**
0.2	*0.17*		0.33	0.41	**0.5**	**0.61**	**0.76**	**1**	**1.5**
0.3	0.22	0.33		**0.53**	**0.65**	**0.8**	**1**	**1.31**	**1.96**
0.4	0.27	0.41	**0.53**		**0.82**	**1**	**1.25**	**1.63**	**2.45**
0.5	0.33	**0.5**	**0.65**	**0.82**		**1.22**	**1.53**	**2**	3
0.6	0.41	**0.61**	**0.8**	**1**	**1.22**		**1.87**	2.45	3.67
0.7	**0.51**	**0.76**	**1**	**1.25**	**1.53**	**1.87**		3.06	4.58
0.8	**0.67**	**1**	**1.31**	**1.63**	**2**	2.45	3.06		*6*
0.9	**1**	**1.5**	**1.96**	2.45	3	3.67	4.58	*6*	

$k_{\max}$ values between 0.5 and 2 are in boldface, $k_{\max}$ values in the range (0.2–0.5) and (2,5) are in normal font and $k_{\max}$ values less than 0.2 or greater than 5 are in italic font.

estimates of k are most difficult to obtain (large numbers of individuals must be screened to find a suitable number of heterozygotes).

For a small subset of the data (74 SNPs) we had access to high precision estimates of k from a sample of ~3000 individuals described in Craig *et al*. [(18), Supplementary Data]. These allowed us to assess the impact of assuming $k = 1$ in our main analysis. The estimates of k were used to recalculate the allele frequency differences from the raw intensity scores and test statistics (and their associated P-values) were calculated using the method described in the Materials and Methods section (based on the nested GLMM but the non-nested GLMM gives very similar results). The P-values were converted to $-\log10(p)$ for comparison between the $k = 1$ denoted $p_{k=1}$) and $k \neq 1$ (denoted $p_{k\neq1}$) cases. The correlation between the two was very high (>0.98 for correlation between $-\log10$ transformed values, >0.99 for correlation between untransformed values). Figure 2 shows the regression of $-\log10(p_{k\neq1})$ on $-\log10(p_{k=1})$. The equation of the regression line (drawn on Figure 2) was $0.000 + 0.989 \times (-\log10(p_{k=1}))$ with standard errors of 0.013 and 0.021 for the estimates of intercept and slope, respectively. The vast majority of the points in Figure 2 fall very close to the line $y = x$ (for clarity the line $y = x$ is not included in Figure 2 as it is almost indistinguishable from the estimated regression line). These results indicate that even if reliable estimates of k were available, this would have a negligible effect on the results shown here. Including k may also cause problems if the sample of individually genotyped individuals is small since the error in estimating k will be substantial. This error must be taken into account in any method which does not use substantial numbers of individuals to estimate k. Although some efforts have been made to centralize the estimation of k values from suitably large samples (1), estimation of k from large numbers of individuals is the exception rather than the norm to date (13,15,17). Furthermore, it is unclear how comparable estimates of k are across different platforms. The recent HapMap paper reported that there was considerable inconsistency between different SNP typing platforms, with fewer than 20% of HapMap SNPs being successfully typed by all of the tested platforms (19), Supplementary Data]. Methods which do not require estimates of k will be preferable when there is doubt about the transferability of results across platforms. When reliable estimates of k are available e.g. from the resource described in (1)], they can be simply incorporated into the analysis we describe here. A modified version of the scripts to implement our method with k estimates is available on request.

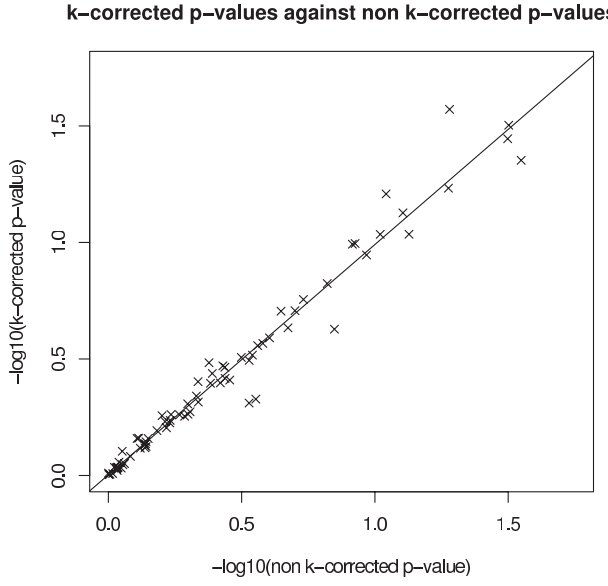

Figure 2. Comparison of log transformed P-values calculated assuming $k = 1$ with log transformed P-values where k was estimated from ~3000 individuals. Data from 74 SNPs are shown. The regression line, $y = 0.000 + 0.989\times$ is drawn on the plot. The line $y = x$ is not shown but is virtually indistinguishable from the regression line.

For SNPs with minor allele frequency >0.1 the binomial sampling standard deviation [i.e. $\sqrt{V}$, where $V = p_a(1 - p_a)/2n_a + p_u(1 - p_u)/2n_u$ for a 384 individual sample is in the range (0.015–0.025)]. The standard deviation of the estimate of allele frequency difference (c) between cases and controls had interquartile range (25th–75th percentile) of (0.018–0.028) for the data examined here (using non-nested GLMM estimates). In the quality control stage of the analysis we discarded SNPs for which there were not at least eight probe measures per pool. If we set a much more stringent criteria for inclusion (>20 probes, resulting in half of all SNPs being discarded), the interquartile range was not substantially smaller (0.018–0.026). Alternatively, if only two arrays per sample were available (8–20 probes per sample), this interquartile range increased to (0.022–0.035). Increasing the number of arrays would decrease the error in allele frequency difference estimation. However, the fixed size of the binomial sampling variance means that the return from increasing the number of arrays will be diminished with more than a handful of arrays per sample. For the variance correction method we describe above, only using one array (restricted to the case where we had 6–10 probes per array) per sample was insufficient to recover a (GLMM corrected) test statistic with appropriate type I error (i.e. the estimate of $\text{var}(e_{\text{pool}})$ was not sufficiently accurate). With two arrays the type I error with the GLMM corrected test statistic was acceptable (although a larger number of arrays may be required for good power, see below).

The error involved in estimating the allele frequency difference in pools will lead to a loss of power. Using the estimate of $\text{var}(e_{\text{pool}})$ based on $\text{var}(e_{\text{pool}-1})$ ($\simeq$0.00058) we can calculate the approximate effective sample size if the same sample were individually genotyped. Equation 2 gives an expression for the effective relative sample size (ERSS) as $\tilde{V}/(\tilde{V} + 2\text{var}(e_{\text{pool}-1}))$ (5). We assume here that the allele frequency $\simeq$1/2 and the number of cases and controls is 384; this gives $\tilde{V} = 0.00065$. Multiplying ERSS by the sample size gives the effective sample size (ESS). For the array data presented here, the ERSS is $\sim(0.00065)/(0.00065+2\times 0.00058)$ or $\simeq$1/3 that of an individually genotyped sample. With 10 arrays the ERSS would be $\simeq$1/2. Increasing the ERSS to 2/3 would require ~40 arrays. Note that distributing the available arrays across sub-pools instead of a single large pool will not affect the ERSS. Increasing the number of sub-pools by a factor f increases $\tilde{V}$ by a factor $\sqrt{f}$ but assuming that the total number of arrays is fixed, the number of arrays per sub-pool will decrease by f, resulting in an increase in $\text{var}(e_{\text{pool}})$ by

a factor $\sqrt{f}$. Since these cancel in $\sqrt{f}\tilde{V}/(\sqrt{f}\tilde{V}+2\sqrt{f}\mathrm{var}(e_{\mathrm{pool}-1}))$ using say 12 arrays on 384 individuals yields the same ERSS as using 6 arrays on each of two sub-pools of 192 individuals.

In practice the most effective study design will vary depending upon whether the total number of individuals is fixed. If the total is not fixed, greater efficiency (in terms of ESS) can be achieved by allocating fewer arrays to larger numbers of individuals. For example, if the limiting factor is that 48 arrays are available, the ESS with 384 individuals typed with 24 arrays per sample (case, control) is $384 \times (0.00065/(0.00065 + 2 \times 0.00058 \times \sqrt{\frac{3}{24}} = 235))$. In comparison with $2 \times 384 = 768$ individuals typed on 12 arrays per sample the ESS equals 406. With 4×384 individuals and six arrays per sample the ESS equals 679. Note that, as discussed above, we recommend that at least two arrays are typed per sample to allow accurate estimation of $\mathrm{var}(e_{\mathrm{pool}})$.

Study design in conventional (not array based) pooling studies was examined by Barratt *et al*. (11). They assessed different possible designs with the cost of the different stages of the pooling experiment factored into a cost efficiency calculation. For the scenarios they consider, cost efficiency is maximized with two sub-pools instead of one large pool [Figure 3 in Barratt *et al*. (11)]. Although the relative sample size increases with increasing numbers of sub-pools, the total cost of the experiment also increases. A detailed examination of the cost efficiency (factoring in the cost of the arrays, sample preparation and so on) of an array based study would be an interesting area for further study.

We have focused our attention on pooling in case-control samples. Although the simplest application of pooling is to such samples, pooling has previously been applied to datasets where the trait of interest was quantitative (20,21). In such circumstances there is no simple way of 'canceling out' the effect of k because the statistical test does not simply involve the difference between two groups. The importance of gaining an appropriate estimate of k would hence be increased compared with the case-control situation we address here.

Based on the results we have obtained some recommendations can be made. Firstly, we recommend investigators do not expend resources on obtaining estimates of k from their own data. At least in the context of case-control studies, funds should instead be spent on replicating arrays in pools and on replicating results with pools across independent samples of cases and controls. When reliable estimates of k are available they can of course be included but the results are unlikely to change substantially compared with the situation where k is taken to be 1. We recommend the use of suitably large numbers of replicate arrays when pooling DNA from large numbers of individuals. With 384 individual samples, up to 10 times replication would give reasonable value for money in terms of increase in power. With smaller samples, the need for replication is decreased. For example with 96 individuals per pool, three arrays would be sufficient to obtain reasonable power; the ERSS would be $\simeq 1/2$ with three arrays if $\mathrm{var}(e_{\mathrm{pool}})$ was similar to the value we observed in our endometriosis data.

We have described a method for screening SNPs from arrays on DNA pools in case-control association studies. By estimating the pooling variance from parallel typed SNPs the method minimizes the number of arrays required. This will facilitate large scale association analysis of suitably large samples at a cost well within the reach of most laboratories.

ACKNOWLEDGEMENTS

The authors acknowledge Sue Treloar for her pioneering work in establishing the endometriosis study. The authors thank staff in the Molecular and Genetic Epidemiology Laboratories for expert assistance in collection and preparation of the DNA samples and Renee Mayne for preparation of the DNA pools. The study and sample collections were partly supported by grants 339430, 339446 and 389892 from the National Health and Medical Research Council and by the Cooperative Research Centre for the Discovery of Genes for Common Human Diseases established and supported by the Australian Government's Cooperative Research Centre's Program. Funding to pay the Open Access publication charges for this article was provided by QIMR.

Conflict of interest statement. None declared.

REFERENCES

1. Simpson,C.L., Knight,J., Butcher,L.M., Hansen,V.K., Meaburn,E., Schalkwyk,L.C., Craig,I.W., Powell,J.F., Sham,P.C. and Al-Chalabi,A. (2005) A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. *Nucleic Acids Res.*, **33**, e25.
2. Brohede,J., Dunne,R., McKay,J.D. and Hannan,G.N. (2005) PPC: an algorithm for accurate estimation of SNP allele frequencies in small equimolar pools of DNA using data from high density microarrays. *Nucleic Acids Res.*, **33**, e142.
3. Rautanen,A., Zucchelli,M., Makela,S. and Kere,J. (2005) Gene mapping with pooled samples on three genotyping platforms. *Mol. Cell. Probes*, **19**, 408–416.
4. Di,X.J., Matsuzaki,H., Webster,T.A., Hubbell,E., Liu,G.Y., Dong,S.L., Bartell,D., Huang,J., Chiles,R., Yang,G. *et al*. (2005) Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958–1963.
5. Visscher,P.M. and Le Hellard,S. (2003) Simple method to analyze SNP-based association studies using DNA pools. *Genet. Epidemiol.*, **24**, 291–296.
6. Treloar,S., Hadfield,R., Montgomery,G., Lambert,A., Wicks,J., Barlow,D.H., O'Connor,D.T., Kennedy,S. and I.E.S. Group (2002) The International Endogene Study: a collection of families for genetic research in endometriosis. *Fertil. Steril.*, **78**, 679–685.
7. Treloar,S.A., Wicks,J., Nyholt,D.R., Montgomery,G.W., Bahlo,M., Smith,V., Dawson,G., Mackay,I.J., Weeks,D.E., Bennett,S.T., *et al*. (2005) Genomewide linkage study in 1176 affected sister pair families identifies a significant susceptibility locus for endometriosis on chromosome 10q26. *Am. J. Hum. Genet.*, **77**, 365–376.
8. Zhao,Z.Z., Nyholt,D.R., James,M.R., Mayne,R., Treloar,S.A. and Montgomery,G.W. (2005) A comparison of DNA pools constructed following whole genome amplification for two-stage SNP genotyping designs. *Twin Res. Hum. Genet.*, **8**, 353–361.
9. Armitage,P., Berry,G. and Matthews,J.N.S. (2002) *Statistical Methods in Medical Research*. Blackwell, Oxford, UK.
10. Lynch,M. and Walsh,B (1998) *Genetics and Analysis of Quantitative Traits*. Sineaur Associates, Sunderland, USA.
11. Barratt,B.J., Payne,F., Rance,H.E., Nutland,S., Todd,J.A. and Clayton,D.G. (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.*, **66**, 393–405.
12. Matsuzaki,H., Dong,S.L., Loi,H., Di,X.J., Liu,G.Y., Hubbell,E., Law,J., Berntsen,T., Chadha,M., Hui,H. *et al*. (2004) Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nature Meth.*, **1**, 109–111.
13. Le Hellard,S., Ballereau,S.J., Visscher,P.M., Torrance,H.S., Pinson,J., Morris,S.W., Thomson,M.L., Semple,C.A.M., Muir,W.J., Blackwood,D.H.R., Porteous,D.J. and Evans,K.L. (2002) SNP genotyping on pooled DNAs: comparison of genotyping technologies

and a semi automated method for data storage and analysis. *Nucleic Acids Res.*, **30**, 30–74.
14. Sham,P., Bader,J.S., Craig,I., O'Donovan,M. and Owen,M. (2002) DNA pooling: A tool for large-scale association studies. *Nature Rev. Genet.*, **3**, 862–871.
15. Norton,N., Williams,N.M., Williams,H.J., Spurlock,G., Kirov,G., Morris,D.W., Hoogendoorn,B., Owen,M.J. and O'Donovan,M.C. (2002) Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum. Genet.*, **110**, 471–478.
16. Moskvina,V., Norton,N., Williams,N., Holmans,P., Owen,M. and O'Donovan,M. (2005) Streamlined analysis of pooled genotype data in SNP-based association studies. *Genet. Epidemiol.*, **28**, 273–282.
17. Norton,N., Williams,N.M., O'Donovan,M.C. and Owen,M.J. (2004) DNA pooling as a tool for large-scale association studies in complex traits. *Ann. Med.*, **36**, 146–152.
18. Craig,D.W., Huentelman,M.J., Hu-Lince,D., Zismann,V.L., Kruer,M.C., Lee,A.M., Puffenberger,E.G., Pearson,J.M. and Stephan,D.A. (2005) Identification of disease causing loci using an array-based genotyping approach on pooled DNA. *BMC Genomics.*, **6**, 138.
19. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
20. Butcher,L.M., Meaburn,E., Dale,P.S., Sham,P., Schalkwyk,L.C., Craig,I.W. and Plomin,R. (2005) Association analysis of mild mental impairment using DNA pooling to screen 432 brain-expressed single-nucleotide polymorphisms. *Mol. Psychiatry*, **10**, 384–392.
21. Jawaid,A., Bader,J.S., Purcell,S., Cherny,S.S. and Sham,P. (2002) Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur. J. Hum. Genet.*, **10**, 125–132.

APPENDIX 1

With pooled data, the sample allele frequency is subject to pool specific error in addition to the usual binomial sampling error. The sample frequency estimate, $\tilde{p}_a$(for cases, for controls replace a with u), from pooled data can be written

$$\tilde{p}_a = \hat{p}_a + e_{\text{pool}-1} = p_a + e_b + e_{\text{pool}-1}$$

where p_a is the true population frequency, e_b is the binomial sampling error and $e_{\text{pool}-1}$ is the error associated with estimating the frequency from the pool (5). This implies that

$$\text{var}(\tilde{p}_a - \tilde{p}_u) = V + 2\text{var}(e_{\text{pool}-1}) \qquad \textbf{A1}$$

and

$$E(T_{\text{simple}}) = \frac{E(\tilde{p}_a - \tilde{p}_u)^2}{\text{var}(\hat{p}_a - \hat{p}_u)} = \frac{V + 2\text{var}(e_{\text{pool}-1})}{V}$$

and that a corrected test statistic is

$$T_1 = T_{\text{simple}} \times \frac{V}{V + 2\text{var}(e_{\text{pool}-1})}$$

In practice we cannot calculate V and we substitute $\tilde{V}$ for V.

APPENDIX 2

The aim here is to construct a corrected statistic for association testing. First we calculate the value of the GLMM case/control estimate and square it (i.e. we compute $\hat{c}^2$) for each SNP. If a simple t-test was applied to the array data instead of the GLMM we would have $\hat{c}^2 = (\bar{p}_a - \bar{p}_u)^2$. Since the mean of $\hat{c}^2$ over all SNPs provides an estimate of $\text{var}(\tilde{p}_a - \tilde{p}_u)$ we may obtain an estimate of $\text{var}(e_{\text{pool}-1})$ by re-arranging equation A1 from the appendix 1 as follows

$$\text{var}(e_{\text{pool}-1}) = \frac{1}{2}(\text{var}(\tilde{p}_a - \tilde{p}_u) - \tilde{V}) \simeq \frac{1}{2}(\text{mean}(\hat{c}^2) - \tilde{V}) = \frac{1}{2}\text{mean}(\hat{c}^2 - \tilde{V})$$

Note again that $\tilde{V}$ is again in practice substituted for V. This allows the calculation of the test statistic in equation 2. This new test-statistic will be substantially less anti-conservative than T_{simple} but will remain slightly anti-conservative because it fails to take into account the varying precision with which the allele frequencies are estimated. With array data, differing numbers of intensity measures will be used to calculate the GLMM case/control estimate (c) for each SNP so there will typically be variation in the precision of the estimates from SNP to SNP. We re-write equation A1 to take into account this error as follows

$$\text{var}(\tilde{p}_a - \tilde{p}_u) = \tilde{V} + 2\text{var}(e_{\text{pool}-2}) + \text{var}(\hat{c})$$

where $\text{var}(\hat{c})$ is the square of the estimated standard error (e.s.e.) of the GLMM case/control estimate $\hat{c}$. Given the linear model specified above this e.s.e. is obtained from the square root of the relevant diagonal element (i.e. the element corresponding to the fixed effect c) of the inverse of the information matrix in a GLMM. This e.s.e. is available in the output of GLMM packages, such as ASReml (http://www.vsn-intl.com/ASReml/). This in turn provides us with a new estimate for the error associated with estimating the frequency from the pool

$$\text{var}(e_{\text{pool}-2}) = \frac{1}{2}\text{mean}\{\hat{c}^2 - \tilde{V} - \text{var}(\hat{c})\}$$

The new corrected test statistic now requires correction for the overall $\text{var}(e_{\text{pool}})$ term and for the SNP specific $\text{var}(\hat{c})$ term. The test statistic for a given SNP, say X [with estimated variance $\text{var}(\hat{c}_X)$], is hence

$$T_{2-\text{X}} = T_{\text{simple}} \times \frac{\tilde{V}}{\tilde{V} + 2\text{var}(e_{\text{pool}-2}) + \text{var}(\hat{c}_{\text{X}})}$$

Note that $\text{var}(e_{\text{pool}-1})$ will typically be larger than $\text{var}(e_{\text{pool}-2})$. Note also that, since the correction in T_2 corrects for a portion of variation that is specific to that SNP, T_2 should be more accurate than T_1.

The test statistics above can be calculated using R (http://www.r-project.org/) and ASReml. Scripts which implement the above method are available on request from the corresponding author.