European Journal of Human Genetics (2002) 00, 000–000
© 2002 Nature Publishing Group All rights reserved 1018-4813/02 $25.00
www.nature.com/ejhg

ARTICLE

Optimal selection strategies for QTL mapping using pooled DNA samples

Ansar Jawaid[1], Joel S Bader[2], Shaun Purcell[3], Stacey S Cherny[3,4] and Pak Sham[1,3]

[1]*Department of Psychological Medicine, Institute of Psychiatry, King's College London. London SE5 8AF, UK;* [2]*CuraGen Corporation, 555 Long Wharf Drive, New Haven, CT 06511, USA;* [3]*Social, Genetic and Developmental Psychiatry Research Centre, Institute of Psychiatry, King's College London, London SE5 8AF, UK;* [4]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK*

Uncorrected PROOF

The cost of large-scale association studies may be reduced substantially by analysis of pooled DNA from multiple individuals. Here we examine the optimal symmetric and asymmetric designs for pooling experiments for quantitative traits under a range of assumptions about the underlying genetic model and the sources of experimental errors in allele frequency estimation. The results indicate that, in the absence of experimental errors and for common alleles with additive effects, a symmetric pooling scheme comparing the top 27% with the bottom 27% of the trait distribution is optimal, extracting 80% the total information available. A symmetric design is not optimal for rare or recessive alleles, which require asymmetric (or other) pooling strategies. Allele frequency measurement errors reduce the optimal pooling fraction as well as the overall efficiency of the pooling design. In contrast, random variation in the amount of DNA contributed by individuals to a pool reduces only the overall efficiency of the pooling design. Our results emphasize the importance of minimising experimental errors and suggest a pooling fraction of around 20%.
European Journal of Human Genetics (2002) **00,** 000–000. DOI: 10.1038/sj/ejhg/5200771

Keywords: QTL association analyses; DNA pooling; SNPs; experimental errors; sample selection

Introduction

The limited power of linkage to detect and localize genes of minor or modest effect has led to the widely accepted view that association is the primary tool of gene mapping in humans.[1] Unlike linkage, which extends over long genetic distances, allelic associations to a disease are usually restricted to the susceptibility locus itself and very tightly-linked polymorphisms.[2] Consequently, the screening of even a megabase of DNA may require 50–100 markers. Although single nucleotide polymorphisms (SNPs) occur at sufficient density in the genome,[3] the need to genotype hundreds of individuals for thousands of markers remains prohibitively expensive. One way of considerably reducing cost is to use DNA pooling, whereby DNA samples from multiple individuals are pooled before genotyping. This technique is ideal for screening a large number of markers for associations, although positive results will require confirmation using individual genotype data.[4–10]

For a categorically defined disease, DNA pooling is necessarily restricted to a simple case-control design, in which allele frequencies are compared across a pool of DNA from cases and a pool of DNA from controls. The appropriate method of analysis and the power of this simple design have been examined.[11] A greater variety of pooling designs is possible for quantitative traits. Bader *et al*.[12] considered symmetric designs under a classical biometrical genetic model and showed that the optimal pooling strategy is to define pools by the top 27% and the bottom 27% of the trait distribution. However, they did not consider asymmetric designs or more importantly the impact of different sources of experimental errors.

***Correspondence: Ansar Jawaid, Department of Psychological Medicine, Institute of Psychiatry, King's College London, London, SE5 8AF UK.**
Tel: +44 (0) 20 7848 0966; Fax: +44 (0) 20 7701 9044;
E-mail: a.jawaid@iop.kcl.ac.uk.
Received 28 August 2001; revised 7 December 2001; accepted 12 December 2001

Technical aspects of DNA pooling dictate that errors in allele frequency estimation will arise. For example, DNA quantification, choice of electrophoresis method, 'plus-A' stutter, and sensitivity (the minimum reliable detectable difference between pools) are all factors that contribute to discrepancies in allele frequency estimation. This error can be reduced to $<5\%$,[13,14] and in the absence of experimental bias maybe to as little as 1%.

We have examined the sensitivity of optimal pooling designs for quantitative traits to variations in genetic model parameters and to experimental noise. After confirming the result of Bader *et al*.[12] that a symmetric pooling design with a pooling fraction of 27% in each tail is optimal for a common additive gene, we show the potentially serious loss of power of this design for rare or recessive alleles. We also consider two sources of experimental noise and show that a high level of experimental accuracy is essential for the success of the pooling strategy, and that the impact of experimental noise on optimal design is to lower to pooling fraction. Finally, we provide practical guidelines for optimal sample selection in DNA pooling studies.

Uncorrected PROOF

Method and results

Genetic model

We assume a diallelic quantitative trait locus (QTL) with alleles A_1 and A_2, occurring at frequencies p and q, respectively. We denote the mean trait effects of the genotypes A_1A_1, A_1A_2, and A_2A_2 by a, d, and $-a$, and their frequencies by $P(G)=p^2$, $2pq$, and q^2, respectively. The mean effect in the population is therefore $m=a(p-q)+2dpq$. The dominance ratio (d/a) is denoted as c, while the proportion of trait variance accounted for by the QTL is represented by ${\sigma_Q}^2$. Under Hardy-Weinberg equilibrium, ${\sigma_Q}^2=2pq[a\text{-}d(p\text{-}q)]2+[2pqd]2={\sigma_A}^2+{\sigma_D}^2$. The distribution of trait scores (X) for each genotype, G, is assumed to be normal with mean μ_G equal to $a-m$, $d-m$, and $-a-m$ for genotypes A_1A_1, A_1A_2, and A_2A_2, respectively, and variance ${\sigma_R}^2=1-{\sigma_Q}^2$ within each genotype. The trait distribution in the population is thus a mixture of three normal distributions with overall mean 0 and variance 1.

Test statistic

We assume that trait values are available for all individuals in a random sample of the population. To test the null hypothesis of no linkage disequilibrium between marker locus and QTL we compare the allele frequencies of A_1 in the lower and upper pools. The test statistic for a two-pool design is

$$Z^2 = \frac{(\hat{p}_U - \hat{P}_L)^2}{\sigma^2},$$

where $\hat{p}_L$ and $\hat{p}_U$ are the estimated frequencies of allele A_1 in the lower and upper pools respectively. The variance of $\hat{p}_U - \hat{p}_L$ is $\sigma^2 = V_S + V_U + V_M$, where V_S represents sampling variation, V_U represents variation in the quantity of DNA contributed by the individuals, and V_M represents variation in the measurement of allele frequency. The sampling variance is

$$V_S = \hat{p}(1-\hat{p})\left(\frac{1}{2n_L}+\frac{1}{2n_U}\right), \text{ where}$$
$$\hat{p} = \frac{n_L\hat{p}_L + n_U\hat{p}_U}{n_L + n_U}$$

and n_L and n_U are the numbers of individuals in the lower and upper pools, respectively.

The accuracy of allele frequency estimation using pooled DNA depends on each individual making an equal contribution of DNA to the pool. However, the process of obtaining equal concentrations of DNA for the individual samples, and then pipetting out equal volumes of the solutions to make up the pool, is subject to experimental error. The variance due to unequal DNA contributions is shown in Appendix A to be

$$V_U \approx \hat{p}(1-\hat{p})\tau^2\left(\frac{1}{2n_L}+\frac{1}{2n_U}\right)$$

where τ is the coefficient of variation (i.e. standard deviation over mean) of the number of DNA molecules of locus A contributed by each individual.

The frequency of an allele in a pool of DNA is a quantitative measure that is also subject to measurement error. We assume that the effect of measurement error is to increase the variance of the allele frequency estimate of each pool by a constant quantity, denoted ε^2. Thus the contribution of measurement error to the variance in allele frequency difference between the two independent pools is therefore simply $V_M=2\varepsilon^2$.

Under the null hypothesis, Z^2 has a χ^2 distribution with one degree of freedom. To compare the efficiencies of different designs, we evaluate the non-centrality parameter (NCP) of Z^2 in the presence of a QTL.

From a random sample of N individuals from the population, individuals with trait values below a threshold T_L are selected for the low pool, while those with trait values above another threshold T_U are selected for the high pool. The expected numbers of individuals in the upper and lower pools are given, respectively, by

$$E(n_U) = N\sum_G\left[1-\Phi\left(\frac{T_u-\mu_G}{\sigma_r}\right)\right]P(G)$$
$$E(n_L) = N\sum_G\left[\Phi\left(\frac{T_L-\mu_G}{\sigma_R}\right)\right]P(G)$$

where Φ is the standard normal distribution function. The expected allele frequencies in the two pools are then

$$E(p_U) = \frac{N\left[\left[1-\Phi\left(\frac{T_U-\mu_{A_1A_1}}{\sigma_R}\right)\right]P(A_1A_1)+\frac{1}{2}\left[1-\Phi\left(\frac{T_U-\mu_{A_1A_2}}{\sigma_R}\right)\right]P(A_1A_2)\right]}{E(n_U)}$$

and

$$E(p_L)=\frac{N\left[\Phi\left(\frac{T_L-\mu_{A_1A_1}}{\sigma_R}\right)P(A_1A_1)+\frac{1}{2}\Phi\left(\frac{T_L-\mu_{A_1A_2}}{\sigma_R}\right)P(A_1A_2)\right]}{E(n_L)}$$

The *NCP* of the test statistic is then

$$NCP=\frac{(E(p_U)-E(p_L))^2}{V_S+V_U+V_M}.$$

Optimal asymmetric and symmetric designs

For any set of model parameters, this *NCP* can be maximised over the thresholds T_L and T_U, and therefore the pool sizes n_L and n_U. Because there are only two variables, the optimisation can be therefore achieved simply by a grid search. Thus, if the true genetic model is known, the optimum selection strategy is to select individuals for the upper and lower pools using the thresholds calculated under the asymmetric pooling scheme for the particular model. In addition, we also maximised this *NCP* subject to the constraint that n_L=n_U. These symmetric designs are particularly relevant when there is no knowledge regarding allele frequency or dominance, so that there is no reason to treat the two tails differently.

We calculated the *NCP* for the optimal asymmetric and symmetric designs for a test of $\sigma_A>0$ under 8 different sets of model parameters, encompassing different levels of QTL heritability, allele frequency and dominance, assuming the absence of experimental errors (Table 1). As expected, with equal allele frequency (P=0.5) and no dominance (c=0), the optimal design is symmetric.[12,15] The optimal design is asymmetric when allele frequencies are not equal or when there is dominance, and the degree of asymmetry and the ratio of the *NCP* for asymmetric and symmetric designs both depend strongly on the magnitude of a/σ_R (R^2=0.98 for regression of the natural log of the *NCP* ratio on a/σ_R, P-value =4×10^{-19}). When $s/\sigma_R<0.5$, the symmetric and asymmetric designs provide equal information. As a/σ_Rincreases, the QTL has a major gene effect and a multimodal phenotypic distribution arises. The asymmetric design essentially selects the individuals who become separated from the main phenotypic distribution.

Although asymmetric pooling is potentially more informative than symmetric pooling, our usual lack of knowledge on allele frequency and dominance means that we would normally adopt a symmetric design. A symmetric design would also be appropriate for a more general test of $\sigma_Q{}^2\neq0$.

Figure 1 shows the expected percentage of total information (obtained by individual genotyping) retained by symmetric designs with different pooling fractions, for the eight models. Here, the *NCP* for individual genotyping is simply the QTL heritability, $\sigma_Q{}^2$.[2,12,16] The optimal pooling fraction is about 27% for all models with the exception of rare

Table 1 Optimum thresholds, pool sizes, and frequency of allele A_1 in pools, for individual models over range of heritabilities, under Symmetric and Aymmetric Pooling schemes, sample size N=1000.

				Optimal Symmetric Pooling					*Optimal Asymmetric Pooling*				
Model	$h^2_{(\%)}$	*p*	*c*	n_L	$n_{U\ (\%)}$	$P_{L\ (\%)}$	P_U	*NCP*	$n_{L\ (\%)}$	$n_{U\ (\%)}$	$P_{L\ (\%)}$	P_U	*NCP*
1	10	0.1	−1	2	2	9.1	44.1	11.3	93	1	9.1	83.0	98.2
2			0	32	32	3.7	18.3	69.3	45	10	4.6	26.4	99.3
3			1	32	32	3.7	18.2	69.1	45	12	4.5	24.4	92.4
4		0.25	−1	15	15	20.3	39.6	26.5	59	7	21.1	48.5	48.6
5			0	28	28	14.2	37.6	79.7	36	19	15.5	40.3	85.5
6			1	28	28	14.3	36.3	71.7	33	25	15.2	37.1	73.3
7		0.5	0	28	28	36.4	63.6	82.6	28	28	36.3	63.6	82.7
8			1	28	28	37.7	59.8	54.8	21	38	35.5	58.7	58.2
1	5	0.1	−1	3	3	9.1	28.0	6.6	79	1	9.2	54.5	32.9
2			0	28	28	5.1	16.3	37.2	41	14	5.8	19.4	45.0
3			1	28	28	5.1	16.1	36.3	40	15	5.8	18.6	42.2
4		0.25	−1	22	22	21.1	32.3	14.3	49	10	21.8	36.6	20.0
5			0	28	28	17.2	33.7	40.2	34	21	17.8	35.0	41.6
6			1	28	28	17.4	32.8	35.3	31	25	17.8	33.4	35.7
7		0.5	0	27	27	40.3	59.7	40.9	27	28	40.4	59.7	40.9
8			1	27	27	41.4	57.3	27.2	22	35	40.3	57.0	28.1
1	1	0.1	−1	15	15	9.2	12.0	1.3	50	7	9.4	13.2	2.0
2			0	27	27	7.6	12.7	8.0	33	21	7.8	13.2	8.3
3			1	27	27	7.6	12.7	7.6	33	21	7.8	13.0	7.8
4		0.25	−1	27	27	22.9	27.6	3.1	36	19	23.1	28.3	3.4
5			0	27	27	21.3	28.8	8.1	30	25	21.5	29.0	8.2
6			1	27	27	21.5	28.5	7.0	29	26	21.7	28.6	7.0
7		0.5	0	27	27	45.7	54.3	8.1	27	27	45.8	54.2	8.1
8			1	27	27	46.3	53.4	5.4	25	30	46.2	53.2	5.4

n_L=sample size of individuals in lower pool (%), n_U=sample size of individuals in upper pool (%). p=frequency of allele A_1, P_L=frequency of allele in A_1 in lower pool, P_U=frequency of allele A_1 in upper pool.

Uncorrected PROOF

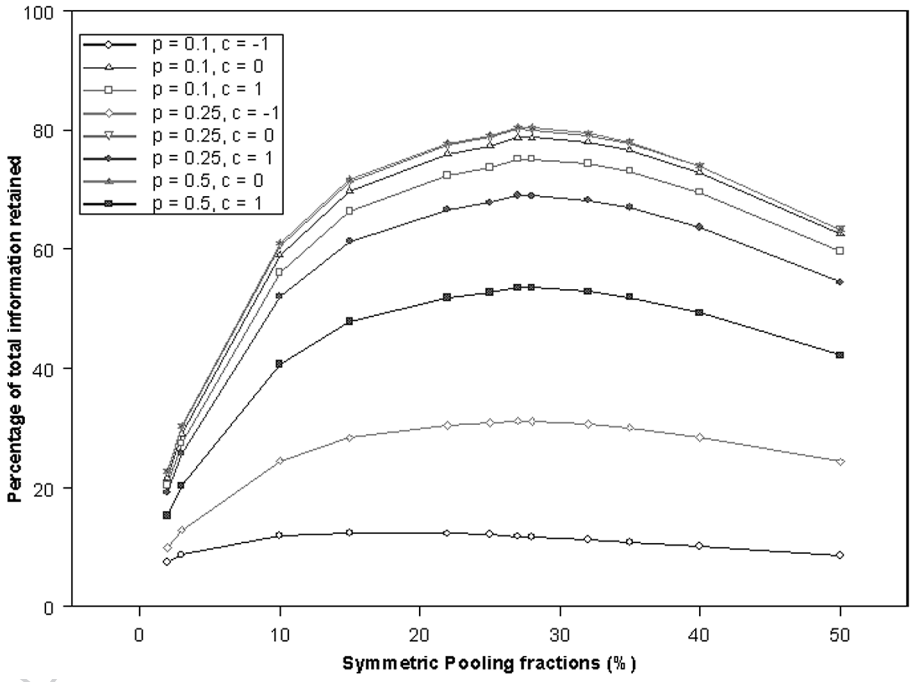

Figure 1 Symmetric pooling scheme compared with individual genotyping for h^2=0.01. Proportion of information is the ratio of the test statistic under symmetric pooling scheme to the test statistic under individual genotyping.

or recessive alleles; the information retained approaches 80% for common additive alleles but is particularly poor for rare recessive alleles.

Experimental error

Next, we investigated the impact of experimental error, focusing only on symmetric pooling designs, employing an equal allele frequency, additive model with 1% QTL heritability. The impact of increasing level of measurement error is to reduce the information retained and to decrease the optimal pooling fraction (Figure 2). Random variation in the amount of DNA contributed by individual subjects reduces the information retained, but does not affect the optimal pooling fraction (Figure 3). In absolute terms, even relatively small values of ε (>0.01) or τ (>0.2) can have a large impact on the information retained by pooling.

For small and additive QTL effects, the *NCP* in the presence of experimental noise is shown in Appendix B to be

$$NCP = 2N\left(\frac{\sigma_A^2}{1+\tau^2}\right)\left(\frac{(\phi(\Phi^{-1}(f)))^2}{f+f^2\kappa^2}\right),$$

where the term κ^2 represents the ratio of the measurement error to other sources of error,

$$\kappa^2 = \frac{\varepsilon^2}{(pq/2N)(1+\tau^2)}.$$

In the absence of experimental noise this reduces to the formula derived by Bader *et al.*[12], which implies an optimal pooling fraction of 27%. It can be seen from these formulas that the optimal fraction remains at 27% regardless of τ when ε is zero, but having $\varepsilon>0$ reduces the optimal pooling fraction. The reason for this difference is that increasing the pool size does not reduce measurement error. Analytical estimates for the optimal pooling fraction, derived in Appendix C, are

$$f = \begin{cases} \Phi(-0.61-0.26\kappa^2), & \kappa<1; \\ \Phi(-(0.82-0.25\ln\kappa^2), & \kappa>1. \end{cases}$$

These analytical estimates are shown in Figure 4 to be quite accurate when compared to the numerical results.

Discussion

We have illustrated that in the absence of experimental error a symmetric pooling sampling scheme, whereby the top and bottom 27% are separately pooled and genotyped, results in a pooling association study that is optimally powerful across a wide range of possible genetic models underlying the trait. The information retained relative to individual genotyping approaches 80% for common additive alleles but is particularly poor for rare recessive alleles.

Our results on the impact of experimental error emphasize the importance of accuracy in both the constitution of the pools and the measurement of allele

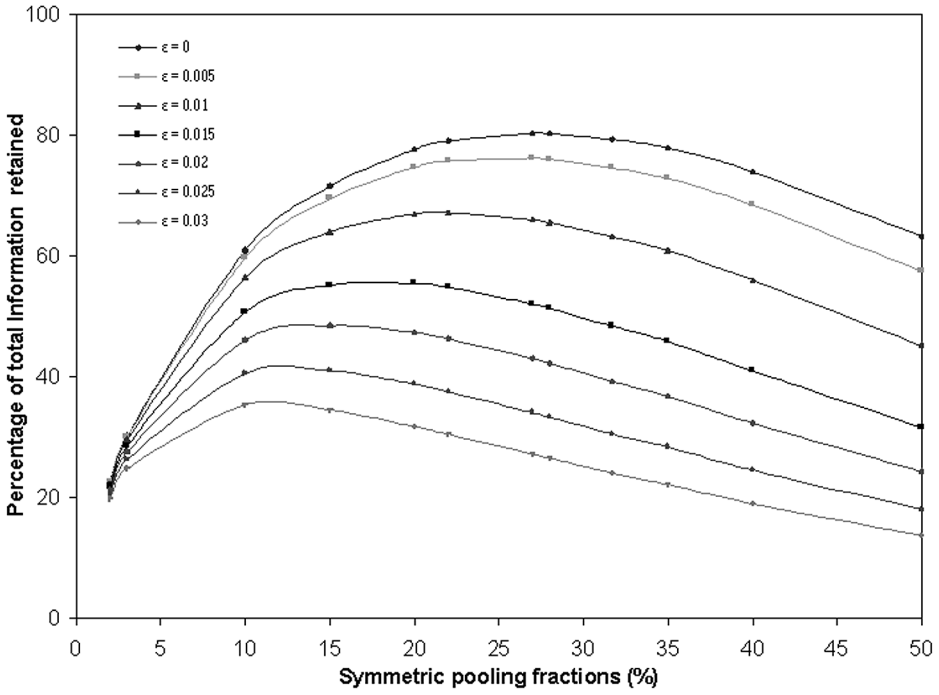

Figure 2 Effect of error from measurement on symmetric pooling scheme assuming various values for standard error of allele frequency measurement error (ε), P=0.5, h^2=0.01, c=0, N=1000.

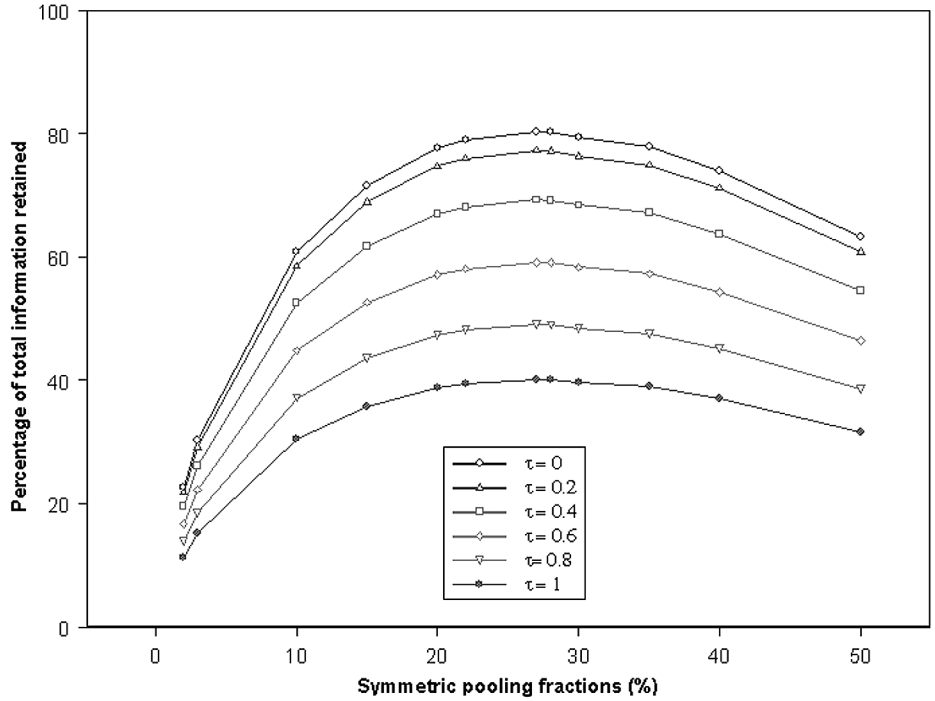

Figure 3 Effect of error due to unequal DNA contribution from individuals on symmetric pooling scheme, assuming various values for the coefficient of variation (τ) of the number of DNA molecules of locus A, P=0.5, h^2=.01, c=0.

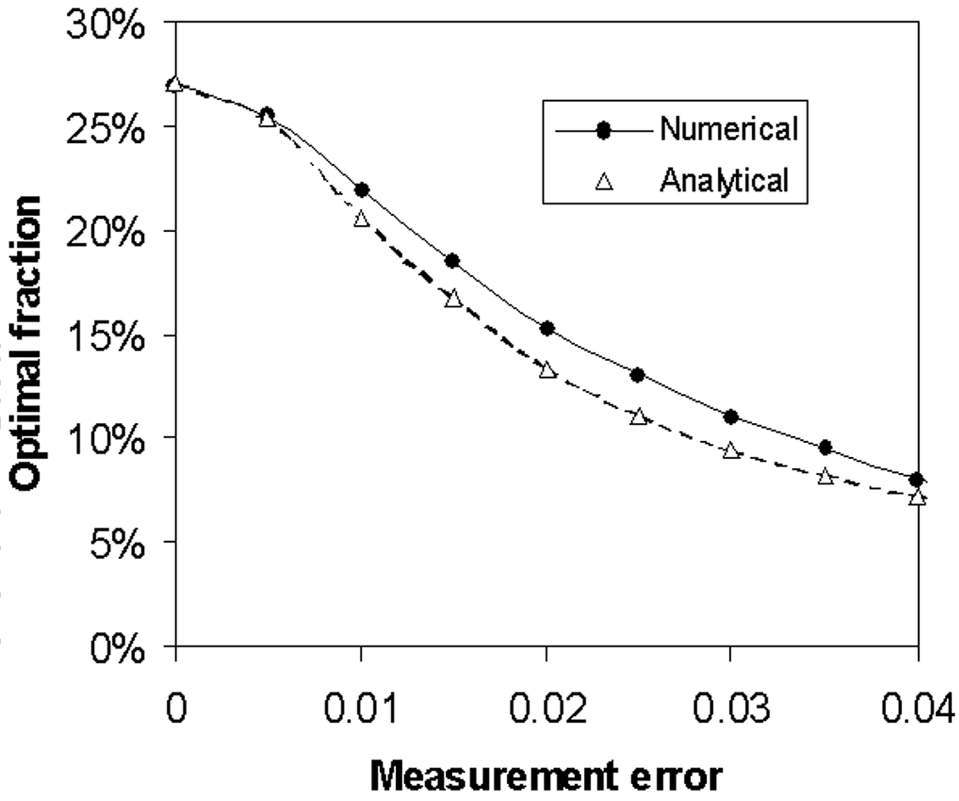

Figure 4 Optimal pooling fraction for a symmetric scheme assuming various values for standard error of allele frequency measurement error (ε), P=0.5, h^2=0.01, c=0, N=1000, from numerical calculations (solid line) and analytical approximations (dashed line).

frequencies. We have shown that random variation in the amount of DNA contributed by individuals to a pool reduces the efficiency of the pooling scheme and that allele frequency measurement errors reduce the optimal pooling fraction as well as the overall efficiency of the scheme. Thus, providing errors from allele frequency estimation can be minimised to within 1% in standard deviation, we recommend symmetric pooling fractions of around 20% as opposed to the 27% that would be optimal in the absence of experimental errors. It may be preferable to replicate the pooling to reduce the DNA concentration variance or, more importantly, to repeat the allele frequency measurement to reduce the effective experimental measurement error.

In our calculations we have assumed that the true values of ε and τ are known, so that they can be specified correctly in the calculation of the test statistic. Clearly, under-specification of ε and τ will lead to liberal P values, whereas over-specification of ε and τ will lead to a conservative test. In practice, values of ε and τ may be estimated from laboratory experiments prior to the actual association study, or inferred from the distribution of values of the test statistics, in a way similar to the use of genomic control for population stratification.[17–20]

While the results we have obtained suppose a single-locus test, biological systems may exhibit multi-locus epistatic effects or gene-environment interactions. The primary effect in the context of the single-locus pooled tests described here is to re-scale the additive variance. For example, suppose that a fraction s of the population is exposed to a sensitising factor that enhances the genotypic effects by a factor λ, from a, d, and $-a$ to λa, λd, and $-\lambda a$ for the genotypes A_1A_1, A_1A_2, and A_2A_2 respectively, and additionally that the sensitising factor is correlated with allele A_1 with correlation constant r. The re-scaled value of σ_A, the correlation between the fraction of allele A_1 in a genotype and the phenotypic shift, is $[1+(\lambda-1)s]\sigma_A + r\sqrt{s(1-s)}(\lambda-1)(p-q)^2 d$. This analysis supposes that the sensitising factor has no independent effect, which is appropriate if phenotypic variables have been conditioned on the applicable covariates. Although interaction terms do not interfere with single-locus pooled tests, estimating the size of the interaction terms would require individual genotyping.

Finally, family-based tests provide additional means to control for environmental effects, and the optimised tests for unrelated populations described here may be extended to family-based studies (JS Bader and P Sham, personal communication).

Acknowledgments

We would like to thank Jing Hua Zhao for helpful comments. This research was supported in part by a UK MRC research studentship to A Jawaid, UK MRC component grant G9700821, Wellcome Trust grant 055379, and National Institutes of Health grant EY-12562.

Uncorrected PROOF

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Abecasis GR, Noguchi E, Heinzmann A *et al*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001; **68**: 191–197.
3 Collins FS, Guyer MS, Chakarvarti A: Variations on a theme: cataloging human DNA sequence variation. *Science* 1997; **274**: 1580–1581.
4 Barcellos LF, Klitz W, Field LL *et al*: Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet* 1997; **61**: 734–747.
5 Daniels J, Holmans P, Williams N *et al*: A simple method for analysing microsatellite allele image patterns generated from DNA pools and its applications to allelic association studies. *Am J Hum Genet* 1998; **62**: 1189–1197.
6 Fisher PJ, Turic D, Williams NM *et al*: DNA pooling identifies QTLs on chromosome 4 for general cognitive ability in children. *Hum Mol Genet* 1999; **8**: 915–922.
7 Hill L, Craig IW, Asherson P *et al*: DNA pooling and dense marker maps: a systematic search for genes for cognitive ability. *Neuroreport* 1999; **10**: 843–848.
8 Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A: Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res* 1998; **8**: 111–123.
9 Stockton DW, Lewis RA, Abboud EB *et al*: A novel locus for Leber congenital amaurosis on chromosome 14q24. *Hum Genet* 1998; **103**: 328–333.
10 Suzuki K, Bustos T, Spritz RA. Linkage disequilibrium mapping of the gene for Margarita Island ectodermal dysplasia (ED4) to 11q23. *Am J Hum Genet* 1998; **63**: 1102–1107.
11 Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res* 1998; **8**: 1273.
12 Bader JS, Bansal A, Sham P. Efficient SNP-based tests of association for quantitative phonotypes using pooled DNA. *Genescreen* 2001; (in press).
13 Hoogendoorn B, Norton N, Kirov G *et al*. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum Genet* 2000; **107**: 488–493.
14 Breen G, Sham PC, Li T, Shaw D, Collier D, Clair ST: Accuracy and sensitivity of DNA pooling with microsatellite repeats using capillary electrophoresis. *Mol Cell Probes* 1999; **13**: 1–7.
15 Schork NJ, Nath SK, Fallin D, Chakarvati A: Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control Subjects. *Am J Hum Genet* 2000; **67**: 1208–1218.
16 Sham, PC, Cherny SS, Purcell S, Hewitt JK: Power of linkage versus association analyses of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 2000; **66**: 1616–1630.
17 Satten GA, Flanders DW, 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.
18 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 2001; **55**: 788–808.
19 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.
20 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220–228.
21 Mood AM, Graybill FA, Boes DC: *Introduction to the theory of statistics*. McGraw-Hill Book Company, 3rd edn. 1974, pp 181.
22 Falconer: The inheritance of liability to certain diseases estimated from the incidence among relatives. *An Hum Genet* 1965; **51**: 227–233.

Appendix A: Variance due to unequal contribution of DNA samples

Restricting the terminology to this appendix, let X_i represent the total number of alleles contributed by individual i in a pool made up of n individuals, with $X_i \sim N(\mu,\tau^2\mu^2)$. Let Y_i represent the number of A_1 alleles contributed by individual i. For genotype A_1A_1 with population frequency p_2, $Y_i=X_i$; for A_1A_2 with frequency p_1, $Y_i \sim \mathrm{Bin}(X_i,1/2)$; and for A_2A_2 with frequency p_0, $Y_i=0$. The population frequency of allele A_1 is $p=p_1/2+p_2$, and the frequency of allele A_1 in the pool is

$$p^* = \frac{\sum Y_i}{\sum X_i}, \text{ with}$$

$$\mathrm{Var}(p^*) \approx \frac{1}{n} \cdot \frac{E(Y_i)^2}{E(X_i)^2} \cdot \left(\frac{\mathrm{Var}(Y_i)}{E(Y_i)^2} + \frac{\mathrm{Var}(X_i)}{E(X_i)^2} - \frac{2\mathrm{Cov}(Y_i)}{E(Y_i)E(X_i)} \right)$$

being the approximate variance for a quotient of correlated random variables.[21] The required terms are

$$E(Y_i) = p_1 E\left(\frac{X_i}{2}\right) + p_2 E(X_i) = p\mu$$

$$E(Y_i^2) = p_1 E\left(\frac{X_i^2}{4} + \frac{X_i}{4}\right) + p_2 E(X_i^2) = \left(\frac{p_1}{4} + p_2\right)(1+\tau^2)\mu^2 + \frac{p_1}{4}\mu$$

$$\mathrm{Cov}(Y_i X_i) = p_1 E\left(\frac{X_i^2}{2}\right) + p_2 E(X_i^2) - p\mu^2 = p\tau^2\mu^2, \text{yielding}$$

$$\mathrm{Var}(p^*) = \frac{1}{n}\left[\left(\frac{p_1}{4} + p_2 - p^2\right)(1+\tau^2) + \frac{p_1}{4\mu}\right]$$

after simplification. Assuming Hardy-Weinberg equilibrium and large μ, this reduces to

$$\mathrm{Var}(p^*) = \frac{p(1-p)}{2n}(1+\tau^2.$$

Appendix B: Optimal symmetric design in the presence of experimental error

Let G be the proportion of A_1 alleles in a genotype, so that $G=0$, 1/2 or 1, and $\mathrm{Var}(G)=pq/2$. According to an additive genetic model, the expected value of the trait X given G is $E(X|G) = -[a + a(p-q)] + 2aG$. Using the implied covariance, $\mathrm{Cov}(X,G)=pqa$, and a linear approximation, the

expected value of G given X is $E(G|X) = p + pqaX$. In the lower pool, $E(X) \approx -\phi(\Phi^{-1}(f))/f$, where ϕ is the standard normal density function, Φ^{-1} is the inverse standard normal distribution function, and f is the lower pooling fraction.[22] The expected values of G is in the lower pool and, by symmetry, the upper pool are therefore

$$E(p_L) = p - \frac{pqa\phi(\Phi^{-1}(f))}{f}$$

$$E(p_U) = p + \frac{pqa\phi(\Phi^{-1}(f))}{f}, \text{with}$$

$$\mathrm{Var}(p_U - p_L) = V_S + V_U + V_M = \frac{pq}{Nf}(1+\tau^2) + 2\varepsilon^2$$

from before. The NCP is therefore

$$NCP = \frac{\left(\frac{2pqa\phi(\Phi^{-1}(f))}{f}\right)^2}{\frac{pq}{Nf}(1+\tau^2) + 2\varepsilon^2} = \frac{4Npqa^2(\phi(\Phi^{-1}(f)))^2}{f(1+\tau^2) + \frac{2Nf^2\varepsilon^2}{pq}}$$

$$= 2N\sigma_A^2\left(\frac{(\phi(\Phi^{-1}(f)))^2}{f(1+\tau^2) + \frac{2Nf^2\varepsilon^2}{pq}}\right).$$

Appendix C: Analytical approximation for the optimal symmetric design in the presence of experimental error

The design is optimised by maximising the value of the *NCP*, which is equivalent to maximising the value of $y^2/(f+f^2\kappa^2)$, where $y = \phi(z)$ and $f = \Phi(z)$ for normal deviate z. Taking the derivative with respect to z and multiplying by non-zero terms yields

$$y + 2zf + 2f\kappa^2(y + zf) = 0$$

as the equation specifying the minimum. When $\kappa = 0$, the solution to this equation occurs at $z_0 = -0.61$, with $f_0 = 0.27$ and $y_0 = 0.33$ (Bader *et al.*[12]). For small κ, we write $z = z_0 + \delta$. To lowest order in δ, the above equation is

$$(z_0 y_0 + 2f_0)\delta - 2z_0 f_0 \kappa^2 = 0, \text{yielding}$$
$$\delta = z_0 f_0 \kappa^2/(1 - z_0^2) \text{ and}$$
$$f = \Phi[z_0 + z_0 f_0 \kappa^2/(1 - z_0^2)] = \Phi[-0.61 - 0.26\kappa^2].$$

When κ is large, we use the asymptotic expansion $f = -(y/z) + (y/z^3)$, and the equation specifying the optimum reduces to $-2y\kappa^2/z^3 = 1$. Taking the natural logarithm of both sides and equating exponents,

$$z^2/2 - 3\ln(-z) + \ln(2\pi) + \ln[(2/\pi)^{1/2}\kappa^2] = 0.$$

Writing $x = - + \delta$ yields $\delta = (1/8) - (1/4)\ln[(2/\pi)^{1/2}\kappa^2]$ to lowest order in δ. The result of this perturbation theory expansion for large κ is

$$f = \Phi[-0.82 - (1/4)\ln(\kappa^2)].$$

An appropriate crossover between the small-κ formula and the large-κ formula is $\kappa = 1$.

Uncorrected PROOF

EJHG		
Manuscript No.	180_01	
Author		
Editor		
Master		
Publisher		

European J. of Human Genetics
Typeset by Elite Typesetting
for Nature Publishing Group

ELITE TYPESETTING TECHNIQUES
www.elitetypesetting.com

QUERIES: to be answered by AUTHOR

AUTHOR: The following queries have arisen during the editing of your manuscript. Please answer the queries by marking the requisite corrections at the appropriate positions in the text.

QUERY NO.	QUERY DETAILS	QUERY ANSWERED
1	Ref 12 Bader et al. 2001 in refs, in press. Please supply any further details and confirm year	
2	Please supply a short running title	
3	Reference 22. Please supply initials for author 'Falconer'	