European Journal of Human Genetics (2002) 00, 00–00

www.nature.com/ejhg

ARTICLE

Family-based association tests for quantitative traits using pooled DNA

Joel S Bader*,[1] and Pak Sham[2,3]

[1]CuraGen Corporation, 555 Long Wharf Drive, New Haven, CT 06511 USA; [2]Department of Psychological Medicine, Institute of Psychiatry, King's College London, UK; [3]Social, Genetic and Developmental Psychiatry Research Centre, Institute of Psychiatry, King's College London, UK

Interest in whole-genome QTL mapping has spurred efforts to reduce the cost of studies now based primarily on individual genotyping. Pooled DNA tests are a possible solution, and understanding how measurement error affects test power could assist in study design. Here we pooled tests explicitly optimised for measurement error, including family-based tests robust to population stratification. Our results suggest that pooled DNA whole-genome screens may be feasible with current instruments.
European Journal of Human Genetics (2002) **00,** 00–00. doi:10.1038/sj.ejhg.5200893

Keywords: association; linkage; quantitative trait locus (QTL); single-nucleotide polymorphism (SNP); DNA pooling; genotyping

Introduction

Association tests of outbred populations may have greater power than linkage analysis to identify the genetic variants contributing to complex human diseases.[1–4] While single-nucleotide polymorphisms (SNPs) occur at sufficient density to provide a suitable marker set,[5–10] individual genotyping remains costly. One method to reduce cost is to pool DNA from individuals with extreme phenotypic values and to measure the allele frequency difference between pools.[11–17] The power of pooled tests has been investigated for case–control studies.[18] More recently, pooled tests have been discussed for quantitative traits. In the absence of experimental error, the optimal design for an unrelated population is to compare frequencies between pools of the most extreme 27% of individuals ranked by phenotypic value, retaining 80% of the information of individual genotyping.[19] This result has been obtained more generally in the context of optimal inefficient statistics.[20] Experimental sources of error, primarily allele frequency measurement error, degrade the test power.[21] Recent applications[22,23] suggest that typical absolute measurement errors are 1–4%.

Population stratification poses a second challenge to pooled tests. Genomic control methods, developed to reduce stratification effects in genotype-based assoc-iation tests,[24–28] are not directly applicable to pooled tests.

Here we present optimized pooled DNA test designs, including family-based tests robust to stratification. Estimates of test power explicitly consider allele frequency measurement error. This distinguishes our treatment from prior theoretical work, permits the optimization of test design as a function of known parameters, and provides a bridge to experimentalists seeking practical guidance for whether to attempt and how to perform pooled association tests.

Methods

Sampling variance and concentration variance

Let p_i represent the frequency of allele A_1 for individual i, either 0, 1/2, or 1, and c_i represent the concentration of DNA contributed by this individual to a pool of n individuals. The allele frequency p^* for the pool is

$$p^* = \sum \frac{c_i p_i}{\sum c_j} = p + \sum \frac{(c_0 + \delta c_i)\delta p_i}{\sum c_0 + \sum \delta c_j}$$

*Correspondence: JS Bader; CuraGen Corporation, 555 Long Wharf Drive, New Haven, CT 06511 USA.
Tel: +1(203)974-6236, Fax: +33-(0)1-4501-2046,
E-mail: jsbader@curagen.com
Received 4 December 2001; revised 27 August 2002; accepted 28 August 2002

$$= p + \sum \frac{\left(\frac{1}{n} + \frac{\delta c_i}{nc_0}\right)\delta p_i}{1 + \frac{1}{nc_0}\sum \delta c_j}$$
$$\approx p + \sum \delta p_i \left(\frac{1}{n} + \frac{1}{n}\frac{\delta c_i}{c_0}\right)\left(1 - \frac{1}{n}\sum \frac{\delta c_j}{c_0}\right)$$
$$\approx p + \frac{1}{n}\sum \delta p_i + \sum \delta p_i \left(\frac{\delta c_i}{nc_0} - \sum \frac{\delta c_j}{n^2 c_0}\right)$$
$$\equiv p + \frac{1}{n}\sum \delta p_i + \frac{1}{n}\sum \delta p_i \delta c_i'$$

which defines the relative concentration error $\delta c_i'$. The terms δp_i and $\delta c_i'$ are uncorrelated, and each has expectation zero. Furthermore, the sum of the $\delta c_i'$ terms is constrained to be zero. The variance of p^* is

$$\mathrm{Var}(p^*) = \frac{1}{n^2}\sum_{i,j}\mathrm{Cov}(\delta p_i, \delta p_j) + \frac{1}{n^2}\sum_{i,j}\mathrm{Cov}(\delta c_i', \delta p_j)\mathrm{Cov}(\delta p_i, \delta p_j)$$
$$= \frac{1}{n^2}\sum_{i,j} r_{ij}\sigma_p^2 + \frac{\tau^2}{n}\sigma_p^2$$

We have used

$$\mathrm{Cov}(\delta p_i, \delta p_j) = \frac{p(1-p)}{2} r_{ij} = \sigma_p^2 r_{ij} \quad \text{and}$$
$$\mathrm{Cov}(\delta c_i', \delta c_j') = \tau^2\left(\delta_{ij} - \frac{1}{n}\right) \approx \tau^2 \delta_{ij},$$

with the concentration coefficient of variation defined as $\tau \equiv [\mathrm{Var}(c_i)]^{1/2}/c_0$ and the genotypic correlation between a pair of individuals defined as r_{ij}.

For the between-family design, a pool of n individuals contains n/s sibships of size s and genotypic correlation r, and

$$\mathrm{Var}(p^*) = \frac{sR}{n}\sigma_p^2 + \frac{\tau^2}{n}\sigma_p^2$$

with $R=(1/s)[1+(s-1)r]$. Since the individuals in the upper and lower pools are unrelated, $V_s+V_c=2\mathrm{Var}(p^*)$.

For a within-family design, the allele frequency difference between pools is

$$\Delta p^* = \frac{1}{n}\sum_i (1 + \delta c_i')\delta p_i - \frac{1}{n}\sum_j (1 + \delta c_j')\delta p_j,$$

where i and j label individuals in the upper and lower pools respectively, and

$$\mathrm{Var}(\Delta p^*) = \frac{2}{n^2}\sum_{i,i'}\mathrm{Cov}(\delta p_i, \delta p_{i'})\left[1 + \mathrm{Cov}(\delta c_i', \delta c_{i'}')\right] - \frac{2}{n^2}\sum_{i,j}\mathrm{Cov}(\delta p_i, \delta p_j)$$
$$= \frac{2(1-r)}{n}\sigma_p^2 + \frac{2\tau^2}{n}\sigma_p^2.$$

Expected allele frequency difference and non-centrality parameter

The genotype-dependent phenotype distribution is defined using a variance components model,

$$X_{ki} = Y_k + Y_{ki} + \mu_{ki}.$$

Family and individual effects are normally distributed with mean zero and variance

$$\mathrm{Var}(Y_k) = t - r\sigma_A^2 - u\sigma_D^2$$
$$\mathrm{Var}(Y_{ki}) = \sigma_R^2 - t + r\sigma_A^2 + u\sigma_D^2$$

The family index is k, the sib index is i, and the individual phenotypes X_{ki} are the sum of Y_k, the family effect excluding the QTL, Y_{ki}, the individual effect excluding the QTL, and μ_{ki}, the QTL effect $\mu(G_{ki})$ for sib i with genotype G_{ki}. The total phenotypic correlation between sibs is t. Both r and u relate to the genetic background shared between sibs, r being the genotypic correlation (1 for monozygotic twins, 1/2 for full sibs, 1/4 for half sibs) and u being the shared genotype expectation (1 for monozygotic twins, 1/4 for full sibs, 0 for half sibs).[29]

The phenotypic values X_{ki} and QTL effects μ_{ki} are re-expressed as family means and individual deviations from family means,

$$X_{k\bullet} = \frac{1}{s}\sum_i X_{ki}$$
$$\delta X_{ki} = X_{ki} - X_{k\bullet}$$
$$\mu_{k\bullet} = \frac{1}{s}\sum_i \mu_{ki}$$
$$\delta\mu_{ki} = \mu_{ki} - \mu_{k\bullet}.$$

The phenotypic variances excluding QTL effects are

$$\mathrm{Var}(X_{k\bullet} - \mu_{k\bullet}) = \frac{1}{s}\left[\sigma_R^2 + (s-1)(t - r\sigma_A^2 - u\sigma_D^2)\right] \equiv T\sigma_R^2$$
$$\mathrm{Var}(\delta X_{ki} - \delta\mu_{ki}) = (1-T)\sigma_R^2$$

When the QTL effects are small, $T \approx (1/s)[1+(s-1)t]$.

The probability that sibling 1 from family κ with genotypes $\mathbf{G}=(G_1,G_2,\ldots,G_s)$ is selected for the upper pool is $1-\Phi[(X'-\mu_G)/\sigma]$, where $\Phi(z)$ is the cumulative normal probability. The variable X under selection (with selection threshold X'), the QTL contribution μ_G, and $\sigma^2 \equiv \mathrm{Var}(X-\mu_G)$ depend on pooling design. For between-family pools, these are $x_{k\bullet}$, $\mu_{k\bullet}$, and $T\sigma_R^2$; for within-family pools, δX_{k1}, $\delta\mu_{k1}$, and $(1-T)\sigma_R^2$. Because the labeling of sibs is arbitrary, the fraction f of individuals selected for the upper pool is equal to the probability that sib 1 is selected,

$$f = \sum_{\mathbf{G}} \mathrm{PR}(\mathbf{G})\{1 - \Phi[(X' - \mu_G)/\sigma]\},$$

where Pr(**G**) is the probability of observing the sibship genotypes **G**. Numerical inversion provides X' as a function of f. When the QTL effect is small ($\mu_G < \sigma$), the linear approximation

$$\Phi[(X' - \mu_G)/\sigma] \approx \Phi(X'/\sigma) - (\mu_G/\sigma)\phi(X'/\sigma)$$

is accurate, where $\phi(z)=d\Phi(z)/dz$ is the normal probability density. This approximation yields $f=1-\Phi(X'/\sigma)$ because the terms linear in μ_G cancel in the sum over **G**.

The expected allele frequency of the upper pool is

$$E(\hat{\boldsymbol{p}}_U) = \frac{1}{f}\sum_{\mathbf{G}} \Pr(\mathbf{G}) p_G \cdot \{1 - \Phi[(X' - \mu_G)/\sigma]\},$$

where p_G represents the allele frequency of sib 1. Using the linear expansion for $\Phi[(X'-\mu_G)/\sigma]$ yields

$$E(\hat{\boldsymbol{p}}_U) = \sum_{\mathbf{G}} \Pr(\mathbf{G}) p_G + \frac{\phi(X'/\sigma)}{f\sigma}\sum_{\mathbf{G}} \Pr(\mathbf{G}) p_G \mu_G = p + \frac{\phi(X'/\sigma)}{f_\sigma} \mathrm{E}(p_G \mu_G).$$

An analogous expression for the lower pools gives a symmetric result, yielding

$$E(\hat{\boldsymbol{p}}_U - \hat{\boldsymbol{p}}_L) = \frac{2\phi[\Phi^{-1}(1-f)]}{f\sigma} E(p_G \mu_G)$$

where X'/σ has been replaced by $\Phi^{-1}(1-f)$.

The expectation of the correlation between p and μ for an individual is

$$\begin{aligned} E(p\mu) &= p^2[a - (p-q)a - 2pqd] + 2pq \cdot \frac{1}{2} \cdot [d - (p-q)a - 2pqd] \\ &= pq[a - (p-q)d] \\ &= \sigma_p \sigma_A \end{aligned}$$

Similarly, the correlation between sibs i and j is $\mathrm{E}(p_i\mu_j)=r_{ij}\sigma_p\sigma_A$, where r_{ij} is their genotypic correlation. Summing over sibs yields either $R\sigma_p\sigma_A$ (between-family pools) or $(1-R)\sigma_p\sigma_A$ (within-family pools) for $\mathrm{E}(p_G\mu_G)$, with $R=(1/s)[1+(s-1)r]$ as before.

Selecting discordant-like sib-pairs is equivalent to selection based on $|\delta X_{ki}|$, and the within-family analytical results are directly applicable. For larger families, discordant-like families are pre-selected in decreasing rank order of the within-family phenotypic variance $\sum_s \delta X_{ks}{}^2$ summed over siblings s.

We have ascertained that the analytical results for the *NCP* are virtually indistinguishable from exact numerical results when the QTL effect is 5% or less of the trait variance. For larger effects, roughly when the effect size $\sigma_A{}^2$ approaches the minor allele frequency, the genotype-dependent phenotype distributions become resolved, transforming a complex trait into Mendelian trait amenable to traditional linkage analysis.

Analytical fit for the optimal pooling fraction

Optimizing the pooling fraction is equivalent to maximizing the objective function $I=2y^2/(f+f^2\kappa^2)$, where y is shorthand for $\phi[\Phi^{-1}(1-f)]$. Writing f as 1-$\Phi(z)$ and optimizing using $dI/dz=0$ yields

$$y \cdot (1 + 2f\kappa^2) - 2zf \cdot (1 + f\kappa^2) = 0.$$

We have used $y=\phi(z)$, $dy/dz=-yz$, and $df/dz=-y$.

When κ^2 is large, z is also large, and f may be replaced by its asymptotic expansion for large z, $f=y\cdot(z^{-1}-z^{-3})$. With this substitution, the optimum satisfies

$$\frac{z^3}{2y\kappa^2} = 1.$$

Taking the natural logarithm of both sides and equating exponents,

$$\frac{z^2}{2} + 3\ln z - \ln(\kappa^2\sqrt{2/\pi}) \equiv L(z) = 0.$$

When κ and z are both large, the term $3\ln z$ is asymptotically small, giving

$$z \sim \sqrt{\ln(2\kappa^4/\pi)} \equiv B(\kappa).$$

An improved fit is obtained by perturbation theory by writing

$$z = B(\kappa)[1 + b(\kappa)],$$

where $\lim_{\kappa\to\infty} b(\kappa)$. Substituting this expression for z into $L(z)$ and simplifying,

$$B^2 b + 3\ln[B(1+b)] = 0,$$

which gives the asymptotic form $b=(3/B^2)\ln B$, or

$$z \sim B - (3/B)\ln B.$$

For clarity, the functional dependence of B and b on κ has been suppressed.

Since the asymptotic behavior for large κ is not affected by introducing terms of lower order in κ, the fit can be improved for small κ without degrading the fit at large κ by writing

$$z = A - (3/A)\ln A + a_1, \text{ where}$$

$$A(\kappa) = \sqrt{a_2 + \ln\left(1 + a_3\kappa^2 + \frac{2}{\pi}\kappa^4\right)}.$$

The constants a_1, a_2, and a_3 are then selected to fit the exact numerical results at particular values of κ. Fitting the results z=0.612 at κ=0 and z=0.8047 at κ=1 provides the particular parameters

$$a_1 = -0.067,\ a_2 = 2,\ a_3 = 3.$$

Results

Consider a population of N/s families, each a sibship of size s (N total individuals). The genotypic correlation within a sibship is denoted r, r=1/4, 1/2, and 1 for half-sibs, full-sibs, and monozygotic twins, respectively. Sibships may also represent inbred lines, r being the the genetic correlation within each line. Sibs in different families are assumed to have uncorrelated genotypes.

To conduct a pooled DNA test for association of a particular allele A_1 with a quantitative trait, individuals are selected for an upper pool, comprising higher phenotypic values, and a lower pool, comprising lower phenotypic value, similar to designs for optimizing breeding value and for QTL mapping.[30–33] We restrict attention to balanced designs: each pool has fN individuals, with $f \leqslant 0.5$ defined as the pooling fraction. Balanced designs are favored when high and low phenotypes are treated symmetrically.[21]

We consider four designs: (i) unrelated individuals (s=1), in which the fN individuals having highest and lowest phenotypic values are selected for the upper and lower pools respectively; (ii) between-family, in which all s sibs from the fN/s families having highest and lowest mean phenotypic values are selected for the upper and lower pools; (iii) within-family, in which the s' sibs having highest and lowest phenotypic values within each family are selected for the upper and lower pools, with f=s'/s; (iv) within-family with pre-selection of discordant families, in which a fraction f' of families with greatest within-family phenotypic variance are selected, $Var = \sum_2 (X_s - \bar{X})^2$ where X_s is the phenotype of sib s and $\bar{X}$ is the family mean, then the extreme high and low sib within each selected family are selected for the upper and lower pool, with f=f'/N.

A suitable statistic for a two-sided test for each design is

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

where the estimated frequencies of allele A_1 in the upper and lower pools are denoted $\hat{p}_U$ and $\hat{p}_L$. The denominator is $\mathrm{Var}(\hat{p}_U - \hat{p}_L) = V_S + V_C + V_M$. The sampling variance V_s represents the unavoidable error in estimating the allele frequency frequency from a finite sample. The concentration variance V_C arises from sample-to-sample DNA concentration variance within a pool. The measurement variance is V_M=$2\varepsilon^2$, where ε is the experimental allele frequency measurement error for each pool. We assume that the three sources of variation are independent, justified when DNA samples are treated uniformly. Other sources of error, for example errors arising from unequal amplification of alleles, may also be included in this statistical framework.[34]

Under the null hypothesis, Z^2 has a χ^2 distribution with one degree of freedom. Under the alternate hypothesis, the tested marker is assumed to be a bi-allelic quantitative trait locus (QTL) with alleles A_1 and A_2 occurring at frequencies p and $(1-p) \equiv q$. For between-family tests, the alleles are also assumed to be in Hardy–Weinberg equilibrium in a random-mating population. The variance of the allele frequency per individual is $\sigma_p^2 = pq/2$, and the estimated allele frequency is $\hat{p} = (\hat{p}_U + \hat{p}_L)/2$. The estimated variance of the allele frequency per individual, denoted $\hat{\sigma}_p^2$, equals $\hat{p}(1-\hat{p})/2$.

The mean phenotypic effects for genotypes G=A_1A_1, A_1A_2, and A_2A_2 are m_G=a,d and $-a$, respectively. The dominance ratio d/a describes the inheritance mode with values -1, 0, and 1 for pure recessive, additive, or dominant inheritance. The proportion of trait variance accounted for by the QTL is denoted σ_Q^2,

$$\sigma_Q^2 = 2pq[a - d(p-q)]^2 + [2pqd]^2 = \sigma_A^2 + \sigma_D^2.$$

The mean QTL effect is m=$(p-q)a$+$2pqd$. Phenotypic values are assumed to be normally distributed for each genotype with mean μ_G=$\mathrm{m}_G - m$ and residual variance $\sigma_R^2 = 1 - \sigma_Q^2$ arising from other genetic and environmental factors. The distribution of phenotypic values in the population is a mixture of three normal distributions with overall mean 0 and variance 1. The total phenotypic correlation between sibs from genetic factors (including the QTL) and environmental factors is denoted t.

The non-centrality parameter (NCP),

$$NCP = [E(\hat{p}_U = \hat{p}_L)]^2 / \mathrm{Var}(\hat{p}_U - \hat{p}_L),$$

measures the information provided by a pooled DNA test. The notation $\mathrm{E}(\hat{O})$ is the expectation of an observable $\hat{O}$. Below we evaluate the NCP numerator, providing accurate analytical results when possible and simulation results otherwise. We calculate the NCP denominator analytically for the null hypothesis. For the alternative hypothesis, the expected allele frequencies for each pool are displaced symmetrically from p to $p \pm \delta p$ (see Methods), and the value of the denominator decreases by a small value proportional to $(\delta p/p)^2$. We make a conservative approximation by ignoring this change and using the null hypothesis denominator throughout. The NCP then equals $(z_{\alpha/2} - z_{1-\beta})^2$, where α and β are the type I and II error rates for the two-sided test and $z_\gamma \equiv \Phi^{-1}(1-\gamma)$ with Φ the cumulative normal probability. Maximising the NCP optimises the test.

The denominator of the *NCP* (see Methods) is

$$V_S+V_C+V_M=\frac{2J\hat{\sigma}_p^2}{Nf}+\frac{2\tau^2\hat{\sigma}_p^2}{Nf}+2\varepsilon^2=\frac{2\hat{\sigma}_p^2}{Nf}\cdot(J+\tau^2)\cdot\left[1+\frac{Nf\varepsilon^2}{(J+\tau^2)\hat{\sigma}_p^2}\right]$$
$$=\frac{2\hat{\sigma}_p^2}{Nf}\cdot(J+\tau^2)\cdot(1+f\kappa^2)$$

where τ is the coefficient of variation for DNA concentration; $R=\frac{1}{s}\cdot[1+(s-=1)r]$ relates family-based genotypic variance components to pairwise correlations; and J is 1 for pools of unrelated individuals, sR for the between-family design, and $(1-r)$ for both within-family designs. Typically τ is less than 10% and τ^2 may be ignored relative to J. The term κ, denoted the scaled measurement error, is defined

$$\kappa\equiv\varepsilon/[(J+\tau^2)\hat{\sigma}_P^2/N]^{1/2}$$

and is independent of QTL effect.

The numerator of the *NCP* is (see Methods)

$$[\mathrm{E}(\hat{p}_U-\hat{p}_L)]^2=\frac{4\sigma_A^2\sigma_p^2\{\phi[\Phi^{-1}(1-f)]\}^2}{\sigma_R^2f^2}\cdot F$$

where $\phi(z)$ is the normal density and F is 1 for pools of unrelated individuals, R^2/T for between-family pools, and $(1-R)^2/(1-T)$ for within-family pools without pre-selection. For the within-family design using discordant-like pre-selection, $F=(1-r)^2/2(1-t)$ for sib-pairs (expressions for larger sibships are unwieldy). The term R is defined above, and $T=\frac{1}{s}\cdot[1+(s-1)t]$ relates family-based phenotypic variance components to pairwise correlations.

The resulting analytical result for the *NCP*, valid for small QTL effect, is

$$NCP=\frac{N\sigma_A^2}{\sigma_R^2}\cdot\frac{F}{J+\tau^2}\cdot\frac{2\{\phi[\Phi^{-1}(1-f)]\}^2}{f+f^2\kappa^2}.$$

The first of the three factor is identical to the *NCP* for an association test performed by individual genotyping on a population of N unrelated individuals; the second factor, with τ=0, is the correction for individual genotyping a population of N/s families each having s sibs and then performing either a between-family test, with $F/J=R/sT$, or a within-family test, with $F/J=(s-1)R/s(1-T)$. The third factor represents the fraction of information retained when the association test is performed by pooling instead of individual genotyping, and maximising this factor with respect to f provides the optimal pooling fraction. With no measurement error, κ=0, tests are optimised with f=0.27 and 80% of the information is retained.[19] As ε increases, the maximum information that can be retained is determined entirely by the single collective term κ.

Expressions for F, J, and κ^2 are summarised in Table 1, and we now provide examples of each family-based design.

Results for between-family designs are depicted in Figure 1 for populations of sib-quads, sib-pairs, and unrelated individuals, each population having 1000 total individuals. The optimal pooling fraction, indicated by an arrow, shifts to lower values as the number of sibs per family decreases. The optimal fraction and the information retained also shift to lower values as the minor allele frequency decreases, with results shown for frequencies 0.1 and 0.01. The raw measurement error is 0.01, and the pooling frac-

Table 1 The non-centrality parameter for family-based pooled DNA designs[a]

Design	F	J
Unrelated individuals	1	1
Between-family	R^2/T	sR
Within-family	$(1-R)^2/(1-T)$	$1-r$
Within-family, discordant pre-selection[b]	$(1-r)^2/2(1-t)$	$1-r$

[a]The non-centrality parameter (*NCP*) is $[E(\hat{p}_U-\hat{p}_L)]^2/\mathrm{Var}(\hat{p}_U-\hat{p}_L)$. The numerator is $F\cdot\left(4\sigma_A^2\sigma_p^2\{\phi\left[\Phi^{-1}(1-f)\right]\}^2/\sigma_R^2f^2\right)$, where F is provided in this table, f is the pooling fraction σ_A^2 and σ_R^2 are the additive and residual variance for a QTL with allele frequency p, σ_p^2 is $p(1-p)/2$, $\phi(z)$ is the normal probability density and $\Phi(z)$ is the cumulative normal probability. The denominator of the *NCP* is $\left[2(J+\tau^2)\sigma_p^2/Nf\right]+2\varepsilon^2$, where J is provided in this table, τ is the coefficient of variation for DNA sample concentrations in the pool, N is the total number of individuals before selection, and ε is the raw measurement error. The combined expression for the *NCP* is $(N\sigma_A^2/\sigma_R^2)\cdot[F/(J+\tau^2)]\cdot\{2(\phi[\Phi^{-1}(1-f)])^2/(f+f^2\kappa^2)\}$, where κ^2 is $N\varepsilon^2/[(J+\tau^2)\sigma_p^2]$ and κ is termed the scaled error. Each sibship has s sibs with genotypic correlation r and phenotypic correlation t; R and T are $(1/s)[1+(s-1)r]$ and $(1/s)[1+(s-1)t]$, respectively.
[b]analytical results are for sib-pairs only. For larger families see numerical results (Figure 3).

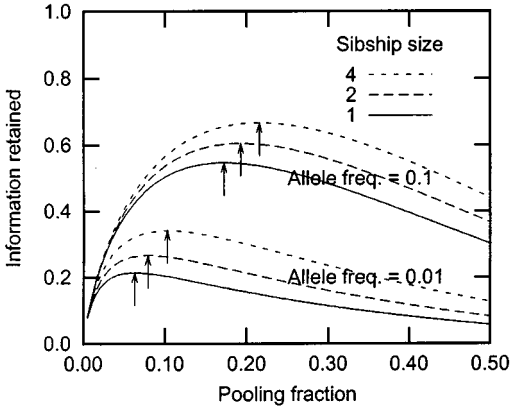

Figure 1 The information retained by the between-family pooled test design, expressed as a fraction of the information from a between-family test based on individual genotyping, is depicted sibships of size 4, 2, and 1, each population having 1000 total individuals. The optimal pooling fraction, indicated by an arrow, shifts to lower values as the number of sibs per family decreases. The optimal fraction and corresponding information retained also shift to lower values as the minor allele frequency decreases, with results shown for frequencies 0.1 and 0.01. The raw measurement error is 0.01.

tion and information retained would decrease for larger ε (see Figure 4 for examples of changing ε).

For within-family designs, the optimal pooling fraction (top panel) and information retained (bottom panel) are shown in Figure 2 as a function of κ for sibship sizes of 2–5, 6, 8, 16 and 32. For sibships through 5, it is always optimal to select just the highest and lowest sib. For larger families and small measurement error, the top and bottom quarters of the sibs are pooled and 80% of the information is retained. The pooling fraction and information retained decrease as the scaled measurement error increases.

Within-family tests can be improved by pre-selection of discordant-like families. In Figure 3, the optimal fraction of families to select (top panel) and information retained (bottom panel) are displayed for sibship sizes 2 through 6 as a function of the scaled measurement error κ (results from computer simulation). The pooling fraction and information retained decrease as κ increases. Pre-selection has the greatest benefit for sib-pairs: for the smallest values of κ, only 56% of families are selected, retaining 80% of the information; had all families been used, only 60% of the information would have been retained. Pre-selection is less beneficial for trios and larger sibships.

In Figure 4, the optimal pooling fraction (top panel) and information retained (bottom panel) using between-family pools and within-family pools with discordant-like pre-selection are displayed for a population of 500 sib-pairs (1000 individuals) as a function of the raw measurement error ε. Results are shown marker frequencies 0.5 and 0.01. With no measurement error, the optimal pooling fraction of 0.27 retains 80% of the information in each case. As measurement error increases, the optimal pooling fraction and information retained both decrease.

The information loss increases for rarer alleles and is worse for the within-family test than for the between-

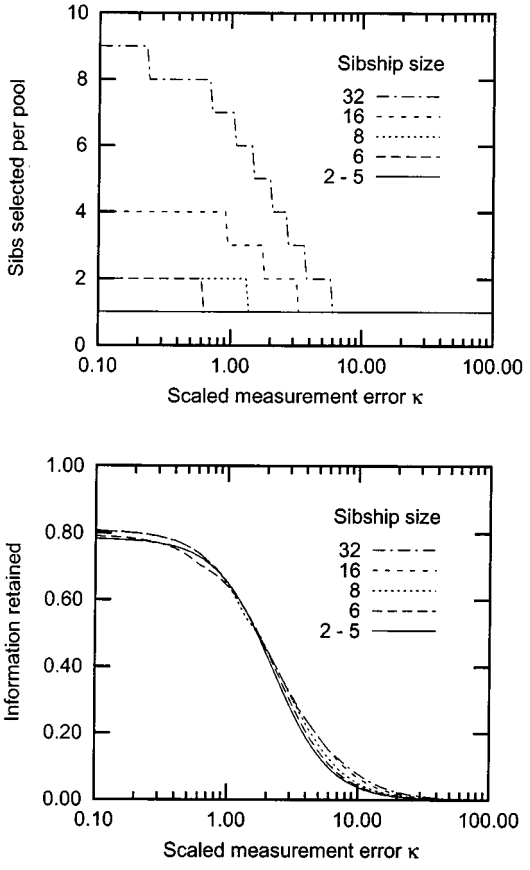

Figure 2 The optimal number of sibs to select from each family (top panel) and the information retained relative to individual genotyping (bottom panel) are shown for sibship sizes 2–5, 6, 8, 16 and 32 as a function of the scaled measurement error κ. For sibships through 5, it is always optimal to select just the highest and lowest sib.

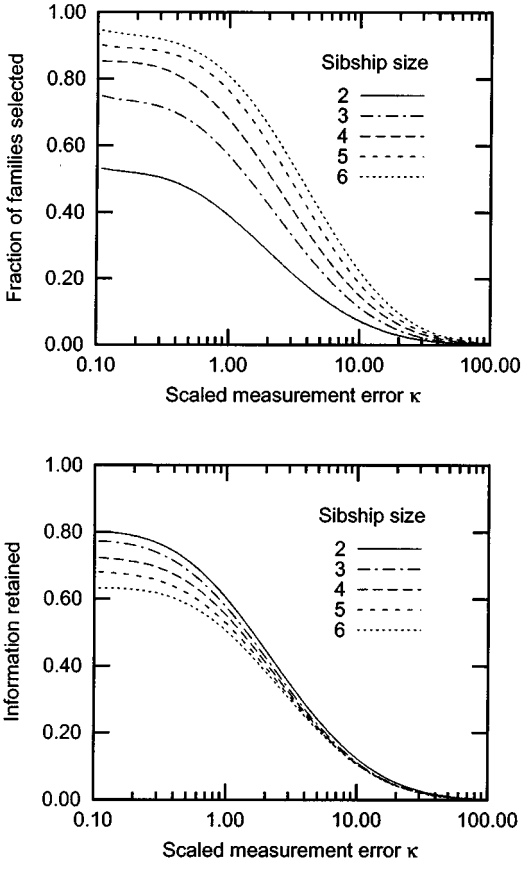

Figure 3 The optimal fraction of families to select (top panel) and information retained (lower panel) are displayed for sibships of size 2 through 6 as a function of the scaled measurement error κ.

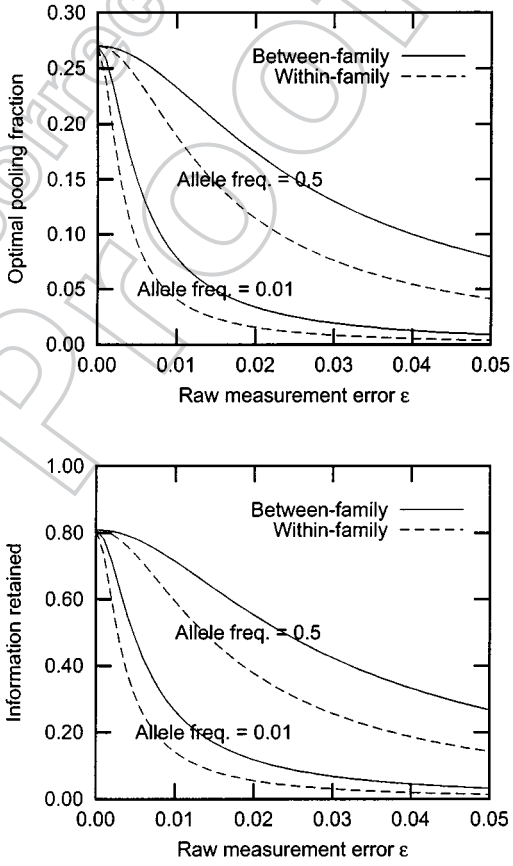

Figure 4 The optimal pooling fraction (top panel) and information retained (bottom panel) for between-family and within-family tests of a population of 500 sib-pairs are shown as a function of raw measurement error for marker frequencies 0.5 and 0.01. The within-family tests include pre-selection of discordant-like families.

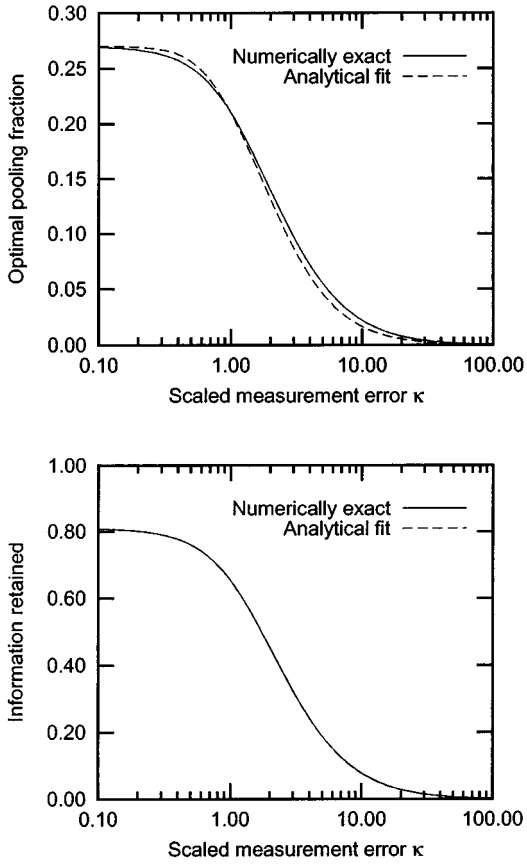

Figure 5 The optimal pooling fraction (top panel) and the information retained (bottom panel) from exact numerical calculations (solid line) and an analytical fit (dashed line) are displayed as a function of the normalised measurement error κ. The fit coincides with the exact results for the information retained.

family test. This behaviour can be deduced from the scaled error κ^2, which is inversely proportional to the allele frequency sampling variance. Since the sampling variance is $3\times$ smaller within-family vs. between-family, κ^2 is $3\times$ larger, $4N\varepsilon^2/p(1-p)$ vs. $4N\varepsilon^2/3p(1-p)$, and more information is lost. The inverse dependence of κ^2 on minor allele frequency explains the decrease in power for rare alleles.

Because the allele frequency difference between sibs is uncorrelated from their allele frequency mean, the between-family and within-family tests are independent estimators of σ_A even when individuals contribute their DNA under both designs. The *NCP* of a combined test is the sum of the *NCP*s for each test and it too follows a χ^2 distribution with 1 degree of freedom. In practice, estimates for σ_A may obtained by inverting the expressions for $\mathrm{E}(\hat{p}_U - \hat{p}_L)$ provided in Table 1, then weighting each estimator by the inverse of its variance.

Population stratification may be indicated by a difference between the estimates for σ_A from a between-family and within-family test. In the absence of stratification, the difference follows a normal distribution with variance

$$\mathrm{Var}[\hat{\sigma}_{A+} - \hat{\sigma}_{A-}] = V_+ \cdot [f_+^2 T\sigma_R^2/4\gamma_+^2 R^2\hat{\sigma}_p^2] + V_- \cdot [f_-^2(1-T)\sigma_R^2/4\gamma_-^2(1-R)^2\hat{\sigma}_p^2]$$

where the '+' and '−' subscripts refers to the between-family and within-family designs respectively, $\gamma_\pm = \phi[\Phi^{-1}(1-f_\pm)]$, and V represents the total variance, V_S+V_C+V_M, for each design. When stratification is indicated, the between-family estimate of $_A$ may be unreliable but the within-family estimate remains robust.

A universal calibration curve for pooled test design is provided in Figure 5, with the optimal pooling fraction (top panel) and information retained (bottom panel) displayed as a function of κ. An accurate analytical fit to the numerically exact results is (see Methods)

$$f = 1 - \Phi[A - (3/A)\ln A - 0.067], \text{ with}$$

$$A(\kappa) = \sqrt{2 + \ln\left(1 + 3\kappa^2 + \frac{2}{\pi}\kappa^4\right)}.$$

The local maxima of the pooling fraction fitting error $f_{\text{fit}} - f_{\text{exact}}$ occur at κ=0.5 (fitting error=+0.006) and at κ=3.5 (fitting error=−0.01). The fitting error for the information retained vanishes on the scale of the figure. The experimental measurement error ε corresponding to the scaled error κ depends on the population structure and marker frequency. For example, for a population of 500 cases, 500 matched unrelated controls, and 10% marker frequency, ε=0.0067κ is the raw error corresponding to κ.

Discussion

Based on the pooled designs described above, we outline a QTL mapping study using 100 000 markers. For 80% power to detect a QTL with 1% additive variance and no more than 100 false-positives from pooled tests (the false-positives may be resolved using individual genotyping), an *NCP* of 17 is required. We assume pooling of discordant sib-pairs to protect against stratification effects. At the scaled error κ=1 where the pooled tests are still close to maximum power, the pooling fraction would be 21%, 65% of the information of a population would be retained, and a population of 2600 individuals would be required. The raw measurement error corresponding to κ=1 for this population size is 0.005 for an allele with 50% frequency and 0.002 for an allele with 5% frequency, 5× to 10× more precise than achieved by current-day instrumentation.

To account for lower precision, we set κ=10, which from Figure 5 is seen to retain 7.7% of the information and corresponds to a pooling fraction of 1.6%. In this case, the total population size would be 22 000; the precision required for a pooled test would be 0.017 for an allele with 50% frequency and 0.007 for an allele with 5% frequency. This is currently feasible if repeated measures are used to decrease the effective measurement error.

Pooled tests perform worse for within-family tests and rare alleles, and may therefore be difficult to apply to disease-risk variants under negative selection pressure. The loss of power may be less severe for pharmacogenetic studies of variants affecting drug response, where selection pressure is absent, and for test crosses of model organisms or agricultural species whose marker frequencies are under experimental control.

The analysis provided here for quantitative traits may be extended to threshold characters yielding dichotomous classifications of a population. For case–control classification, the disease prevalence corresponds to the pooling fraction *f*. When the quantitative character is available for measurement, it is approximately 4× more efficient to compare unrelated individuals with extremely high *vs* extremely low characters than to compare the derived cases *vs* controls.[19]

In summary, we have derived the optimal pooling fractions for within-family and between-family tests of association. With ideal instrumentation, 80% of the information is retained and the optimal pooling fraction is 27%. As allele frequency measurement error increases, the optimal pooling fraction and the information retained both decreases. The information loss is more severe for low-frequency alleles and for within-family tests. The optimal pooling fraction depends on a single parameter representing the measurement error, and a universal calibration curve provides optimised designs as a function of this parameter.

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Ott J: *Analysis of Human Genetic Linkage*. 3rd edn. Baltimore: Johns Hopkins University Press, 1999; pp 000. (1)
3 Sham PC, Cherny SS, Purcell S *et al*: Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Gen* 2000; **66**: 1616–1630.
4 Ardlie KG, Kruglyak L, Seielstad M: Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 2000; **3**: 299–309.
5 Collins FS, Guyer MS, Chakarvarti A: Variations on a theme: cataloguing human DNA sequence variation. *Science* 1997; **274**: 1580–1581.
6 Abecasis GR, Noguchi E, Heinzmann A *et al*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Gen* 2001; **68**: 191–197.
7 Reich DE, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.
8 Patil N, Berno AJ, Hinds DA *et al*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294**: 1719–1723.
9 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
10 Dawson E, Abecasis GR, Bumpstead S *et al*: A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 2002; **418**: 544–548.
11 Barcellos LF, Klitz W, Field LL *et al*: Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Gen* 1997; **61**: 734–747.
12 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 Gen* 1998; **62**: 1189–1197.
13 Shaw SH, Carrasquillo MM, Kashuk C *et al*: Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res* 1998; **8**: 111–123.
14 Stockton DW, Lewis RA, Abboud EB *et al*: A novel locus for Leber congenital amaurosis on chromosome 14q24. *Hum Gen* 1998; **103**: 328–333.
15 Suzuki K, Bustos T, Spritz RA: Linkage disequilibrium mapping of the gene for Margarita Island ectodermal dysplasia (ED4) to 11q23. *Am J Hum Gen* 1998; **63**: 1102–1107.
16 Fisher PJ, Turic D, Williams NM *et al*: DNA pooling identifies QTLs on chromosome 4 for general cognitive ability in children. *Hum Mol Gen* 1999; **8**: 915–922.
17 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.

18 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**: 1273–1288.
19 Bader JS, Bansal A, Sham P: Efficient SNP-based tests of association for quantitative phonotypes using pooled DNA. *Genescreen* 2001; **1**: 143–150.
20 Mosteller F: On some useful 'inefficient' statistics. *Annals of Mathematical Statistics* 1946; **17**: 377–408.
21 Jawaid A, Bader JS, Purcell S *et al*: Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur J Hum Gen* 2002; **10**: 125–132.
22 Beutow KH, Edmonson M, MacDonald R *et al*: High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci USA* 2001; **98**: 581–584.
23 Grupe A, Germer S, Usuka J *et al*: In silico mapping of complex disease-related traits in mice. *Science* 2001; **292**: 1915–1918.
24 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 788–808.
25 Pritchard JK, Rosenberg NA: Use of unliked genetic markers to detect population stratification in association studies. *Am J Hum Gen* 1999; **65**: 220–228.
26 Pritchard JK, Stephens M, Rosenberg NA *et al*: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.
27 Zhang S, Zhao H: Quantitative similarity-based association tests using population samples. *Am J Hum Gen* 2001; **69**: 601–614.
28 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 Gen* 2001; **68**: 466–477.
29 Falconer DS, MacKay TFC: *Introduction to quantitative genetics*. Boston: Addison-Wesley, 1996; pp 000. (2)
30 Hill WG: Design and efficiency of selection experiments for estimating genetic parameters. *Biomerics* 1971; **27**: 293–311.
31 Kimura M, Crow JF: Effect of overall phenotypic selection on genetic change at individual loci. *Proc Natl Acad Sci USA* 1978; **75**: 6168–6171.
32 Darvasi A, Soller M: Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* 1994; **138**: 1365–1373.
33 Ollivier L, Messer LA, Rothschild MF *et al*: The use of selection experiments for detecting quantitative trait loci. *Genet Res, Camb* 1997; **69**: 227–232.
34 Le Hellard S, Ballereau SJ, Visscher PM *et al*: SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res* 2002; **30**: e74, (electronic preprint).

EJHG		
Manuscript No.	253_01	
Author		
Editor		
Master		
Publisher		

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

QUERIES: to be answered by AUTHOR/EDITOR

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	Reference 2 – please supply page span	
2	Reference 29 – please supply page span	