Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA

Joel S. Bader*, Aruna Bansal† and Pak Sham†,‡

* CuraGen Corp., 555 Long Wharf Drive, New Haven, CT 06511, USA

† Gemini Genomics Ltd, 162 Science Park, Milton Road, Cambridge, CB4 0GH, UK

‡ Institute of Psychiatry, King's College, De Crespigny Park, London, SE5 8AF, UK

Abstract

Introduction Genetic factors underlying complex diseases are difficult to identify: many polymorphisms may contribute, each having a small effect and low penetrance. These factors may be identified by association studies of large populations, an alternative to family-based linkage studies. Allele frequency measurements of pooled DNA selected from population-level DNA repositories can reduce the costs of these studies. We provide guidance for selecting unrelated individuals for pooling and for comparing the power of studies based on pooled measurements to the power of individual genotyping, particularly for studies using single-nucleotide polymorphism (SNP) markers.

Materials and methods We used exact numerical calculations to set pooling criteria that maximized the power to detect association as a function of marker frequency, inheritance mode, and additive variance. Analytical approximations are also provided.

Results and discussion Power estimates are provided for two pooled DNA designs: the classification of individuals as affected or unaffected, analogous to a case-control design, and the optimized selection of individuals with extreme phenotypic values. Optimized selection is approximately fourfold more efficient than affected/unaffected classification. The optimal design for most markers is to pool the top and bottom 27% of individuals. Neglecting experimental measurement error, this design requires a population only 1.24-fold larger than that required for individual genotyping. When measurement error is included, the pooled DNA association test serves better as a pre-screen to identify candidate markers which then proceed to individual genotyping. This strategy can still provide a 100-fold savings over individual genotyping.

Keywords association, genotyping, linkage, QTL mapping, single-nucleotide polymorphism (SNP).

Introduction

Association studies detect markers in linkage disequilibrium with causative genetic polymorphisms. Single-nucleotide polymorphisms (SNPs) occur at a sufficient density to provide a suitable set of biallelic markers.[1] Linkage disequilibrium for these markers has been estimated to extend to 5000–100 000 nucleotides,[2–4] suggesting that many thousands of such markers are required for a full-genome scan.[5] These markers have an additional benefit: nucleotide substitutions in protein-coding regions, particularly those that change amino acid sequence, may be functional polymorphisms that directly affect phenotypic values. Association tests can require thousands to tens of thousands of individuals and have spurred the growth of population-level DNA repositories, unselected for any particular phenotype, as a resource whose costs may be shared across multiple studies.[6–8]

The most powerful methods for detecting the association between a marker and a phenotype require individual genotyping. Experimental savings come by testing allele frequency differences between DNA pooled from individuals selected

Correspondence to: Joel S. Bader, CuraGen Corp., 555 Long Wharf Drive, New Haven, CT 06511, USA.
E-mail: jsbader@curagen.com

UNCORRECTED PROOF

according to phenotypic value.[9–12] A conventional selection scheme for a disease phenotype is to classify individuals as affected or unaffected, analogous to a case-control study. Selection based on underlying disease-risk phenotypes, including quantitative measures such as blood pressure or body mass index is also possible, and may identify genetic markers for disease predisposition.

While there has been a limited discussion of optimized selection criterion for pooled DNA studies in the context of human genetics, association tests of DNA pooled on the basis of a quantitative phenotype are analogous to selection experiments for quantitative trait locus (QTL) mapping. Work in this field has shown that, under certain cost assumptions, it is optimal to genotype the upper and lower 27% of an unrelated population to estimate the effect of a marker on a quantitative phenotype.[13–15] Studies of sib-pair designs have yielded related findings.[16]

We applied similar techniques to provide optimized selection criteria for association studies of pooled DNA using the allele frequency difference between pools as a test statistic. We assumed that the samples were drawn from a pre-existing population-level DNA repository collected from individuals unselected for any particular phenotype, and that each individual has been measured for a particular phenotype of interest; the goal is to select pools to maximize the power of the test.

Assuming no experimental error in allele frequency measurements on pooled DNA, we determine the selection thresholds that maximize the power to detect association as a function of the frequency, phenotypic displacement, and inheritance mode of a functional polymorphism. The genetic parameters are also described in terms of a genotype relative risk model. Power calculations are then used to derive the repository size required to detect association at specified false-positive and false-negative rates. These calculations are performed at three decreasing levels of accuracy: exact numerical calculations using the true multinomial distribution of the test statistic; numerical calculations based on an approximate normal distribution of the test statistic; and analytical approximations accurate for complex traits where the polymorphism has a small effect on the phenotype.

Results are depicted in terms of the repository sizes required for three types of experimental designs for detecting association with a quantitative phenotype: first, a pooled DNA test using a conventional affected/unaffected classification; second, a pooled DNA test of extreme individuals using optimized selection thresholds; third, individual genotyping of the entire population. We conclude with a discussion of the reduction in power of pooled DNA tests due to experimental measurement error and with suggestions for effective use of pooled DNA tests in practice.

Materials and methods

The calculation of optimized selection thresholds begins with a model for the genotype-dependent distribution of phenotypic values. A quantitative phenotype, denoted X, is standardized to have unit variance and zero mean. The phenotype is hypothesized to be affected by alleles A_1 and A_2, with frequencies p and $1 - p$, respectively, at a particular QTL. The population frequencies $P(G)$ for genotypes $G = A_1A_1$, A_1A_2 and A_2A_2 are assumed to obey the Hardy–Weinberg equilibrium. Using standard notation for a variance components model,[17] the effect μ_G of genotype G on phenotype X is a, d and $-a$, for genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. These displacements are each offset by subtracting $(2p - 1)a + 2p(1 - p)d$ to preserve the overall phenotypic mean of zero.

The inheritance mode of the QTL is represented by the displacement d of the heterozygote, for example purely recessive ($d = -a$), additive ($d = 0$), or dominant ($d = +a$) inheritance. The inheritance mode partitions the phenotypic variance due to the QTL into the additive variance σ_A^2 and the dominance variance σ_D^2, with

$$\sigma_A^2 + \sigma_D^2 = 2p(1 - p)[a - d(2p - 1)]^2 + 4p^2(1 - p)^2d^2.$$

This partitioning is important because, as will be seen below, pooled tests are sensitive primarily to the additive component of variance. Note that the additive component may be large, even when the inheritance is purely dominant or recessive. The contributions to the phenotype from remaining genetic and environmental factors are assumed to follow a normal distribution with residual variance σ_R^2,

$$\sigma_R^2 = 1 - (\sigma_A^2 + \sigma_D^2).$$

The genotype-dependent phenotype distributions for each genotype are

$$P(X|G) = (2\pi\sigma_R^2)^{-1/2}\exp[-(X - \mu_G)^2/2\sigma_R^2],$$

normal distributions centred at μ_G with width σ_R. The overall phenotype distribution is the weighted sum of the distributions from each genotype,

$$P(X) = \Sigma_G P(X|G)P(G).$$

For a complex trait in which the QTL makes a small contribution, the three underlying distributions may be unresolved in the observed $P(X)$.

This variance components model may be connected to an equivalent affected/unaffected genotype relative risk model by specifying a threshold phenotypic value X_T that classifies individuals as affected ($X > X_T$) or unaffected ($X < X_T$). The proportion r of the total population that is affected is the overall risk or disease prevalence; the probability that an individual with genotype G is affected, divided by the corresponding probability for an individual with genotype A_2A_2, is the genotype relative risk.

In the tests of pooled DNA considered here, a sample repository of total size N serves as the source of DNA to be selected

UNCORRECTED PROOF

for one of two pools; not every individual need be selected. The test statistic is the difference in the frequency that a particular allele, here always assumed to be A_1, occurs in the two pools. For a quantitative phenotype, it is natural to specify an upper threshold X_U and a lower threshold X_L as the selection criteria. Individuals with phenotypic values above X_U are selected for the upper pool; individuals with phenotypic values below X_L are selected for the lower pool; and individuals with phenotypic values between X_L and X_U are not pooled at all. The number of individuals selected for each pool is ρN. The fraction ρ expressed in terms of X_U is

$$\rho = \Sigma_G \Phi[-(X_U - \mu_G)/\sigma_R]P(G),$$

which is solved numerically to determine X_U. The genotypes of individuals selected by $X > X_U$ follow a multinomial distribution; the probability $\theta_U(G)$ that an individual selected for this pool has genotype G is $\Phi[-(X_U - \mu_G)/\sigma_R]P(G)/\rho$. A multinomial distribution is defined similarly for the lower pool,

$$1 = \Sigma_G \theta_L(G) = \rho^{-1}\Sigma_G \Phi[(X_L - \mu_G)/\sigma_R]P(G),$$

using the lower threshold X_L,

A pooling design based on an affected/unaffected classification is similar: affected individuals are selected for the upper pool; an equivalent number of suitably matched unaffected individuals are selected for the lower pool. The selection thresholds X_U and X_L are identical to the classification threshold X_T. The relative risk for genotype G, expressed in terms of the pooling threshold, is $[\theta_U(G)/P(G)]/[\theta_U(A_2A_2)/P(A_2A_2)]$.

The repository size N required to detect association between genotype G and either the quantitative phenotype X or the affected/unaffected classification depends on the desired type I error rate α and type II error rate β, the chosen test statistic, and the experimental design,[18] as well as on the underlying genetic model. For a one-sided test of a single marker, $\alpha = 1 - \Phi(z_\alpha)$ and $1 - \beta = \Phi(-z_{1-\beta})$, where $\Phi(z)$ is the cumulative probability distribution for standard normal deviate z. For a genome scan, the values $\alpha = 5 \times 10^{-8}$ ($z_\alpha = 5.33$) and $1 - \beta = 0.8$ ($z_{1-\beta} = -0.84$) have been suggested.[5] The null hypothesis is denoted H_0 with all μ_G equal to zero, and the alternative hypothesis is denoted H_1 with at least one non-zero μ_G.

An exact calculation of the repository size required to attain desired error rates for a specified genetic model proceeds as follows. First, a value of the pooling fraction ρ or the disease prevalence r is selected. A trial repository size N is specified, with the number of individuals n selected per pool set to the integer part of ρN or rN. Next, the probability $P_0(i,j,k)$ of selecting i individuals with genotype A_1A_1, j individuals with genotype A_1A_2, and k individuals with genotype A_2A_2, with $i + j + k$ equal to n, is tabulated using the multinomial distribution

$$P_0(i,j,k) = [n!/(i!j!k!)](p^2)^i(2p - 2p^2)^j(1 - 2p - p^2)^k.$$

The frequency of allele A_1 for this pool composition is $(2i + j)/2n$. The probability that two pools selected in this manner differ in frequency by at least Δp is calculated as the sum of $P_0(i,j,k)P_0(i',j',k')$ for all combinations of i,j,k and i',j',k' where

$$[2(i - i') + (j - j')]/2n \geq \Delta p.$$

Significance at level α is attained by increasing Δp until this sum is less than or equal to α. If not even the maximum value $\Delta p = 1$ is sufficient for significance at level α, then a larger value of N is selected for the current value of ρ and the calculation begins anew. Otherwise, multinomial probabilities for pool compositions are calculated under H_1 using

$$P_U(i,j,k) = [n!/(i!j!k!)]\theta_U(A_1A_1)^i\theta_U(A_1A_2)^j\theta_U(A_2A_2)^k$$

for the upper pool and a similar term $P_L(i',j',k')$, with θ_L replacing θ_U, for the lower pool. The probability that the allele frequency difference between the upper and lower pools is at least Δp is obtained as the sum of $P_U(i,j,k)P_L(i',j',k')$ for all compositions i,j,k and i',j',k' where $[2(i - i') + (j - j')]/2n \geq \Delta p$. If this probability is greater than or equal to β, the current N is feasible for type I error α and type II error β and a smaller value for N is attempted. This process continues until the smallest feasible N is found.

For the affected/unaffected design, this procedure is followed for each value of r. For the tail pool design, the smallest feasible value for N is calculated as a function of ρ, and the entire design is optimized by searching for the pooling fraction ρ with the smallest feasible N.

When each pool contains a large number of individuals and many copies of each allele, the distribution of allele frequencies for the pool approaches a normal distribution. The difference in allele frequencies between pools, which continues to serve as the test statistic, approaches a normal distribution as well. The pool sizes required to achieve specified error rates are obtained accurately in this case by approximating the multinomial distributions of allele frequencies as normal distributions. Under H_0, the mean of the test statistic is zero and the variance is $\sigma_0^2/n = p(1 - p)/n$, derived by noting that the variance of the frequency difference is twice the variance of the mean for a single pool of n individuals. The allele frequency variance for an individual is $p(1 - p)/2$, and averaging over the n individuals reduces the variance by the factor n.

Under H_1, the expected allele frequency difference Δp is

$$\Delta p = p_U - p_L = \Sigma_G[\theta_U(G) - \theta_L(G)]p_G,$$

where the genotype-dependent allele frequency p_G is 1 for $G = A_1A_1$, 0.5 for A_1A_2, and 0 for A_2A_2. The variance is σ_1^2/n, where σ_1^2 is obtained from the multinomial distribution,[19]

$$\sigma_1^2 = \Sigma_G[\theta_U(G) + \theta_L(G)]p_G^2 - (p_U^2 + p_L^2).$$

UNCORRECTED PROOF

The repository size N required for type I error α and power $1 - \beta$ is $n = [z_\alpha\sigma_0 - z_{1-\beta}\sigma_1]^2/\Delta p^2$.

For tail pools, ρ is then varied to find the smallest N.

The normal approximation underestimates the repository size requirement relative to the exact results from the multinomial distribution. When the sum of the alleles in both pools is at least 60, the difference in repository sizes is no greater than 5%. We chose 60 alleles in both pools as the criterion for switching from the multinomial to the normal calculation. Standard algorithms were employed to perform the root search for X_U and X_L, the optimization, and the integration over the tail of a normal distribution.[20]

In the regime of typical complex traits, the effect of any single QTL is small, the residual variance σ_R^2 is nearly 1, and analytical results may be obtained by expanding Δp to second order in the effect size μ_G. This corresponds loosely to a perturbation theory for probability distributions.[21] The Δp expansion in turn requires a Taylor series expansion for $\Phi(z)$,

$$\Phi(z - \delta) = \Phi(z) - \delta(\mathrm{d}/\mathrm{d}z)\Phi(z) + (1/2)\delta^2(\mathrm{d}/\mathrm{d}z)^2\Phi(z),$$

truncated at second order. The first derivative is

$$(\mathrm{d}/\mathrm{d}z)(2\pi)^{-1/2}\int_{-\infty}^{z}\mathrm{d}z'\exp(-z^2/2) = (2\pi)^{-1/2}\exp(-z^2/2) \equiv y,$$

where y is the height of the normal distribution at normal deviate z, and the second derivative is

$$(\mathrm{d}/\mathrm{d}z)(2\pi)^{-1/2}\exp(-z^2/2) = -yz.$$

Summing these terms,

$$\Phi(z - \delta) = \Phi(z) - y\delta - (1/2)yz\delta^2.$$

Substituting this approximation into the expressions for $\theta(G)$ using $\delta = \mu_G/\sigma_R$ and $z = \Phi^{-1}(1 - \rho)$ yields for the tail design

$$p_U = P + (y/\rho\sigma_R)\{\Sigma_G P(G)p_G\mu_G\} + (y|z|/2\rho\sigma_R^2)\{\Sigma_G P(G)p_G\mu_G^2\}$$

and

$$p_L = p - (y/\rho\sigma_R)\{\Sigma_G P(G)p_G\mu_G\} + (y|z|/2\rho\sigma_R^2)\{\Sigma_G P(G)p_G\mu_G^2\}.$$

The corresponding expressions for the affected/unaffected pools, with $z = \Phi^{-1}(1 - r)$, are

$$p_U = P + [y/r\sigma_R]\{\Sigma_G P(G)p_G\mu_G\} + [y|z|/2r\sigma_R^2]\{\Sigma_G P(G)p_G\mu_G^2\}$$

and

$$p_L = p - [y/(1 - r)\sigma_R]\{\Sigma_G P(G)p_G\mu_G\} - [y|z|/2(1 - r)\sigma_R^2]\{\Sigma_G P(G)p_G\mu_G^2\}.$$

The required sums are

$$\Sigma_G P(G)p_G\mu_G = \sigma_A[p(1 - p)/2]^{1/2},$$

and

$$\Sigma_G P(G)p_G\mu_G^2 = (1/2)(1 - \sigma_R^2) - 4p^2(1 - p)^2ad + (2p - 1)\sigma_D^2/2 \approx \sigma_A^2/2.$$

The approximate value $\sigma_A^2/2$ for the second sum neglects the dominance variance and is exact for purely additive inheritance. It serves to simplify the final equations for Δp. Little error is made in the resulting Δp for two reasons: first, even with dominant or recessive inheritance, the additive variance is often larger than the dominance variance; second, this factor is part of a correction term that is already small.

The results for Δp are

$$\Delta p = 2^{1/2}y\sigma_0\sigma_A/\rho\sigma_R,$$

tail pools, and

$$\Delta p = [1 + \Phi^{-1}(1 - r)\sigma_A/2^{3/2}\sigma_0\sigma_R]y\sigma_0\sigma_A/2^{1/2}r(1 - r)\sigma_R,$$

affected/unaffected pools.

To the same order of approximation, σ_1^2 may be equated with σ_0^2, and the number of individuals required per pool is

$$n = [z_\alpha - z_{1-\beta}]^2\sigma_0^2/\Delta p^2.$$

The preceding three equations lead directly to our main results, Eqns 1 and 2.

The perturbation theory above is valid when the expansion parameters μ_G/σ_R are small, typically satisfied when $\sigma_A^2/2p(1 - p)$ is smaller than 1. In this regime, approximate genotype relative risks may be obtained from the Taylor series expansion for $\theta(G)$. To lowest order, the relative risk for the heterozygote is $1 + (d + a)y/r\sigma_R$, and for the A_1A_1 homozygote is $1 + 2ay/r\sigma_R$. For additive inheritance, $d = 0$, and the relative risk is multiplicative with allele dose when $ay/r\sigma_R$ is small.

If individual genotypes are measured for the N individuals in the population, the regression coefficient b_1 in the regression model

$$X = b_1(p_G - p) + \varepsilon$$

is a suitable test statistic. The residual contribution ε to the phenotype has mean zero and is uncorrelated with p_G. Under H_0, b_1 has mean zero and variance:

$$\mathrm{Var}(b_1|H_0) = N^{-1}\mathrm{Var}(X)/\mathrm{Var}(p_G) = 1/N[p(1 - p)/2].$$

Under H_1, the expected value and the variance of b_1 are

UNCORRECTED PROOF

$$E(b_1|H_1) = \mathrm{Cov}(X,p_G)/\mathrm{Var}(X) = \sigma_A[p(1-p)/2]^{1/2}$$

and

$$\mathrm{Var}(b_1|H_1) = N^{-1}\mathrm{Var}(\varepsilon)/\mathrm{Var}(p_G) = \sigma_R^2/N[p(1-p)/2].$$

The repository size required for a one-sided test of b_1 with Type I error α and power $1-\beta$ is

$$N = [z_\alpha\,\mathrm{Var}(b_1|H_0)^{1/2} - z_{1-\beta}\,\mathrm{Var}(b_1|H_1)^{1/2}]^2/[E(b_1|H_1)]^2,$$

which is presented in simplified form as Eqn 3.

Results and discussion

We consider two experimental designs using DNA pooled from individuals selected from a pre-existing repository of N samples: affected/unaffected pools, with DNA pooled from n affected and n unaffected individuals; and tail pools, with DNA pooled from the n most extreme individuals at each tail of the phenotype distribution.

For the affected/unaffected design, the expected number of affected individuals is $n = rN$, and an additional n suitably matched controls are selected from the remainder of the population. An analytical approximation for the repository size is:

$$N_{\mathrm{aff/unaff}} = [z_\alpha - z_{1-\beta}]^2[\sigma_R^2/\sigma_A^2]\cdot 2r(1-r)^2/\{y_r^2[1 + \Phi^{-1}(1-r)\sigma_A/2^{3/2}\sigma_R p^{1/2}(1-p)^{1/2}]^2\}, \quad (1)$$

where y_r is the height of the standard normal distribution at $\Phi^{-1}(\rho)$ (see Materials and methods for derivation). Repository size requirements are minimized with a prevalence of 50%, much larger than values realistic for complex disorders.

The tail pools are parameterized by the fraction $\rho = n/N$ of population N selected for each pool. An analytical approximation for the repository size is

$$N_{\mathrm{tail}} = [z_\alpha - z_{1-\beta}]^2[\sigma_R^2/\sigma_A^2]\cdot \rho/2y_\rho^2, \quad (2)$$

where y_ρ is the height of the standard normal distribution at $\Phi^{-1}(\rho)$ (see Materials and methods for derivation). The design is optimized by selecting ρ to minimize $\rho/2y_\rho^2$ and hence N_{tail}. The optimal fraction, 27.03%, is independent of all remaining parameters.

The repository size required to achieve the same error rates using individual genotyping is

$$N_{\mathrm{indiv}} = [z_\alpha - z_{1-\beta}\sigma_R]^2/\sigma_A^2, \quad (3)$$

based on a regression model of phenotypic value on allele dose (see Materials and methods for derivation).

Results of the analytical approximations are shown in Fig. 1 with individual genotyping serving as a reference. The tail design, with ρ = 27% of the population selected for each pool, requires a repository only 1.24-fold larger than required for individual genotyping. It is also robust to variation in ρ near its optimum, as values from 19% to 37% drop the efficiency no more than 5%. In contrast, for 10% disease prevalence, the affected/unaffected design requires a repository 5.3-fold larger than that required for individual genotyping and is 4-fold less efficient than the tail design.

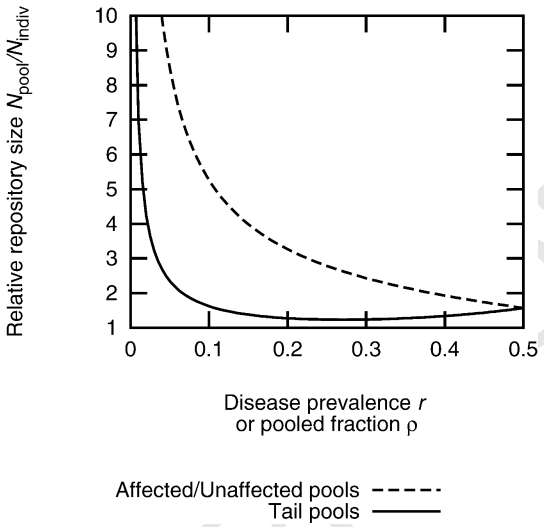

Figure 1 The repository size required to detect association for a QTL for a complex trait is shown for pooled DNA designs relative to individual genotyping designs having equivalent type I and type II error rates. The ratio $N_{\mathrm{aff/unaff}}/N_{\mathrm{indiv}}$ for affected/unaffected pools (dashed line) is shown as a function the disease prevalence r, while the ratio $N_{\mathrm{tail}}/N_{\mathrm{indiv}}$ (solid line) is shown as a function of the fraction ρ of the total population selected for each pool. The optimum value of $N_{\mathrm{tail}}/N_{\mathrm{indiv}}$ is 1.24 and occurs at ρ = 27.03% selected for each pool.

The effect of varying the inheritance mode is shown in Fig. 2 for tail pools. In this example, the type I error is 5×10^{-8}, the type II error is 0.2, and the displacement a is 0.25 in units of the phenotypic standard deviation. The heterozygote displacement d varies from $-a$, pure recessive inheritance, to $+a$, pure dominant inheritance. Results are shown for three frequencies of allele A_1: P = 0.5, 0.1 and 0.01. Solid lines correspond to exact numerical calculations. In the top panel showing the repository size N, filled circles correspond to analytical approximations, Eqn 1, and are virtually indistinguishable from exact calculations. When P = 0.5, A_1 and A_2 have equal frequencies, the additive variance is 0.03125, and the dominance variance is 0 regardless of inheritance mode. Since the population requirements depend primarily on the additive variance, N is independent of the inheritance mode. For allele frequencies below 0.5, the additive variance increases from left to right and the population requirements decreases. The maximum population is required when $d = a/(2p-1)$, which always falls outside the range depicted. The bottom panel depicts the corresponding values of ρ from the numerical calculations. The optimal pooling fractions fall in a narrow range from 24.5% to 27.5%, close to the analytical approximation of 27.03%.

UNCORRECTED PROOF

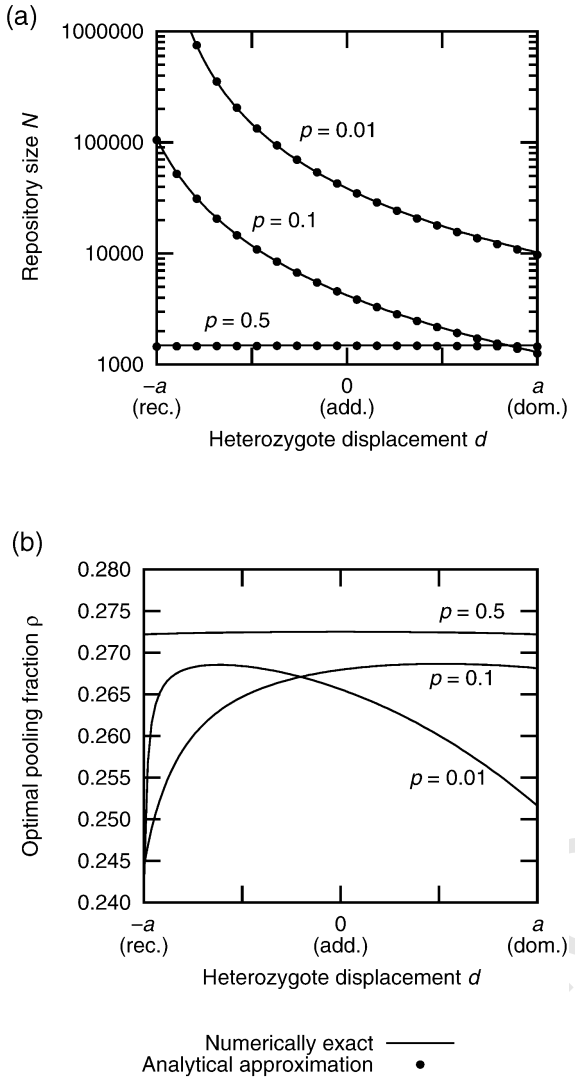

Figure 2 The effect of varying the inheritance mode is shown for tail pools. The type I error is 5×10^{-8}, the type II error rate is 0.2, and the displacement a is 0.25 in units of the phenotypic standard deviation. The displacement d of heterozygotes varies from $-a$, pure recessive inheritance, to $+a$, pure dominant inheritance. Three allele frequencies are shown, P = 0.5, 0.1 and 0.01. Solid lines correspond to exact numerical calculations. (a) The repository size N is shown. Filled circles corresponding to analytical approximations, Eqn 1, are virtually indistinguishable from exact calculations. (b) The optimal pooling fraction ρ from numerical calculations falls in a narrow range from 24.5% to 27.5%, close to the analytical approximation of 27.03%.

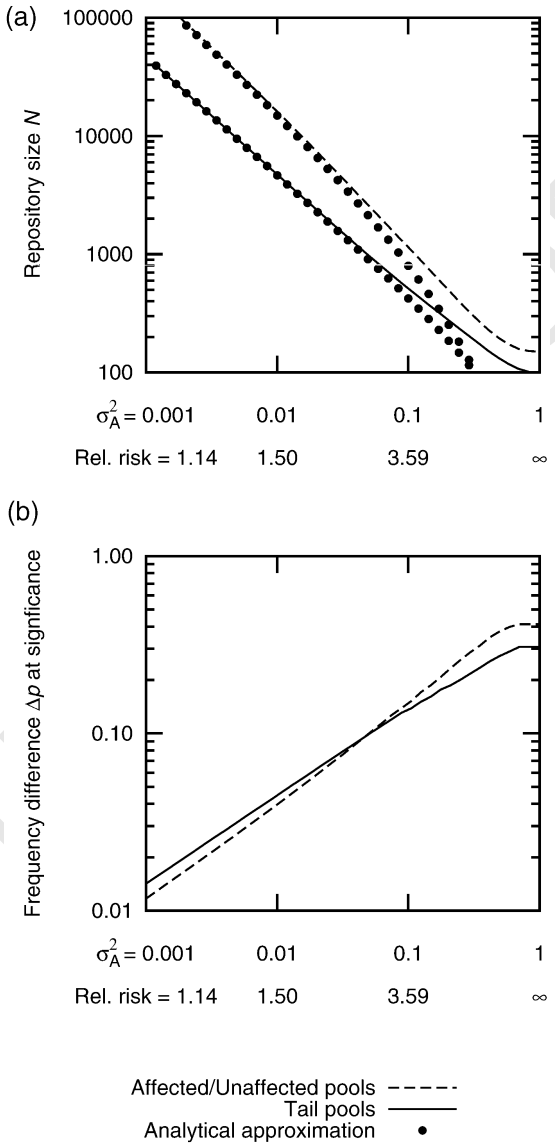

Figure 3 (a) Exact numerical results for the repository size N required to achieve a type I error rate of 5×10^{-8} and type II error rate of 0.2 are shown for affected/unaffected pools (dashed line) and tail pools (solid line) as a function of the additive variance, also presented as the genotype relative risk for a heterozygote, for an allele with frequency 0.1 and purely additive inheritance. Analytical approximations (solid circles), Eqns 1 and 2, are indistinguishable from the exact results when the genotype relative risk is smaller than 2. The disease prevalence r is 10% for the affected/unaffected pools, and 27% of the population is selected for each of the tail pools. (b) The frequency difference at the significance threshold is shown for the same parameters. This threshold determines the measurement accuracy required for association tests based on pooled DNA.

The effect of varying the additive variance directly, or equivalently the genotype relative risk for an allele of known frequency, is shown in Fig. 3. The top panel of Fig. 3 shows that analytical approximations for N from Eqns 1 and 2 (solid circles) are nearly indistinguishable from the exact numerical results (dashed and solid lines) when the genotype relative risk is below a factor of 2–3. Type I and II error rates are 5×10^{-8} and 0.2, respectively, and the allele frequency is 0.1. The bottom panel shows the corresponding allele frequency difference that must be measured for a significant finding with a test of pooled DNA. For example, alleles carrying a 1.5-fold heterozygote relative risk, corresponding to an additive variance of 0.01, have a raw frequency difference of 0.04 at significance: the upper pool has an allele frequency of 0.12 and the lower pool a frequency of 0.08. The population size required to achieve significance is 4700, with 1270 individuals selected per pool.

UNCORRECTED PROOF

This analysis assumes that allele frequency measurement error is negligible. Allele frequencies measured by most technologies, including PCR amplification,[22] kinetic PCR,[23] denaturing high performance liquid chromatography,[24] single-strand conformation polymorphism,[25] pyrophosphate sequencing,[26] and mass spectrometry,[27] are typically reported with standard errors in the range 0.01 to 0.02. Assuming a measurement error of 0.01, the measurement error in the frequency difference is larger by a factor of $\sqrt{2}$, yielding a final error of 0.014. Based on the measurement error, the allele frequency difference of 0.04 in the example above corresponds to a z-score of 2.86 and a type I error rate of 0.002.

While this error rate is much larger than the error rate of 5×10^{-8} required for a whole-genome scan, a practical solution is to employ pooled allele frequency measurements as a pr-screen; candidate associations identified by the pr-screen may then be confirmed by individual genotyping of the entire population, or possibly just the extreme tails. Setting a type I error rate for the pr-screen of 0.01 (z-score of 2.33), corresponding to an allele frequency difference of 0.033, implies a 100-fold savings over an equivalent study that does not employ a pre-screen.

This experimental limitation sets a threshold for the effect size that may be identified in a pooled DNA pre-screen. The relationship between the expected value of Δp and the parameters of the genetic model for a SNP with purely additive inheritance is

$$\Delta p = 2.44 \times [z_\alpha/(z_\alpha - z_{1-\beta})]p(1-p)a,$$

where the initial factor of 2.44 arises from the optimized pooled tail design, z_α and $z_{1-\beta}$ corresponds to the type I and II errors that would be obtained neglecting measurement error, and a is the phenotypic displacement as before. For use in a pre-screen with a P-value of 0.01 from measurement error alone, $z_\alpha = 2.33$ is reasonable. To retain at least 95% of the true associations, β should be no greater than 0.05, with $z_{1-\beta} = -1.64$. These parameters yield Δp equal to $1.43 \times p(1-p)a$, or $p(1-p)a = 0.023$ for the 0.033 frequency difference threshold. For a minor allele frequency of 0.1, this corresponds to a displacement a of 0.26 and an additive variance of 0.012; for allele frequencies of 0.5, the displacement is 0.092 and the additive variance is 0.0042. Thus, the pre-screen retains the power to detect markers with additive variance down to 0.5% to 1.5%, depending on the marker frequency.

In conclusion, we have compared the efficiencies of tests for association using DNA pooled according to an affected/unaffected design and an optimized tail design. The optimal fraction for tail pooling is 27%; rare alleles, which are more difficult to detect in general than more frequent alleles contributing the same phenotypic variance, have a slightly lower optimal fraction. The tail design is approximately 4-fold more efficient than an affected/unaffected design, suggesting that quantitative phenotypes are preferable to qualitative classifications when DNA sample collections are compiled. Excluding effects of measurement error, the repository size required for a pooled DNA study is only 1.24-fold larger than that required for an individual genotyping study with the same type I and II error rates. When the effects of measurement error are included, tests of allele frequency differences between pools may still be valuable as a pre-screen that can reduce the number of markers for individual genotyping by a factor of 100.

References

1 Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000; **405**: 847–856.
2 Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 1999; **22**: 139–144.
3 Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci USA* 2000; **96**: 15173–15177.
4 Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. Linkage disequilibirum in the human genome. *Nature* 2001; **411**: 199–204.
5 Risch NJ, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
6 Rabinow P. *French DNA: Trouble in Purgatory*. Chicago, University of Chicago Press, 1999.
7 Hagmann MU. K. plans major medical DNA database. *Science* 2000; **287**: 1184.
8 Jeffords JM, Daschle T. Political issues in the genome era. *Science* 2001; **291**: 1249–1251.
9 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.
10 Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G. Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Gen* 1997; **61**: 734–747.
11 Daniels JK, Holmans P, Williams NM, Turic D, McGuffin P, Plomin R, Owen MJ. A simple method for analysing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *Am J Hum Genet* 1998; **62**: 1189–1197.
12 Risch NJ, 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.
13 Hill WG. Design and efficiency of selection experiments for estimating genetic parameters. *Biometrics* 1971; 27: 293–311.
14 Kimura M, Crow JF. Effect of overall phenotypic selection on genetic change at individual loci. *Proc Natl Acad Sci USA* 1978; 75: 6168–6171.
15 Ollivier L, Messer LA, Rothschild MF, Legault C. The use of selection experiments for detecting quantitative trait loci. *Genet Res, Camb* 1997; 69: 227–232.
16 Zhao H, Zhang H, Rotter JI. Cost-effective sib-pair designs in the mapping of quantitative-trait loci. *Am J Hum Gen* 1997; 60: 1211–1121.
17 Falconer DS, MacKay TFC. *Introduction to Quantitative Genetics*. Boston, Addison-Wesley, 1996.
18 Snedecor GW, Cochran WG. *Statistical Methods*, 8th edn. Ames, Iowa, Iowa State University Press, 1989.
19 Beyer WH, ed. *CRC Standard Mathematical Tables*, 27th edn. Boca Raton, Florida, CRC Press, 1984.
20 Press Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in C, The Art of Scientific Computing*, 2nd edn. Cambridge, UK, Cambridge University Press, 1997.

UNCORRECTED PROOF

21 Chandler D. *Introduction to Modern Statistical Mechanics*. New York, Oxford University Press, 1987.
22 Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Gen Res* 1998; 8: 111–123.
23 Germer S, Holland MJ, Higuchi R, High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Gen Res* 2000; **10**: 258–266.
24 Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, O'Donovan MC. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum Gen* 2000; **107**: 488–493.
25 Sasaki T, Tahira T, Suzuki A, Higasa K, Kukita Y, Baba S, Hayashi K. Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. *Am J Hum Gen* 2001; **68**: 214–218.
26 Alderborn A, Kristofferson A, Hammerling U. Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. *Genome Res* 2000; **10**: 1249–1258.
27 Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, Braun A. 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.

UNCORRECTED PROOF

Author Query Form

Journal: Genescreen

Article: 036

Dear Author,

During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.

Many thanks for your assistance.

Query Refs.	Query	Remarks
1	**define 'QTL'**	