http://www.ashley-pub.com

Report

The relative power of SNPs and haplotype as genetic markers for association tests

Joel S Bader

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

Identifying the polymorphisms that contribute to disease predisposition and drug response is a major goal of the post-genome era. Single nucleotide polymorphisms (SNPs) in disease-related genes are often used as candidates in the search for causative variations. Association tests based on haplotypes have also been suggested and, at times, have provided greater statistical power than tests based on the underlying SNPs. Here we review the statistical model traditionally used to describe association studies for complex traits and derive novel results for the relative power of SNP-based and haplotype-based tests of association. In the model, a set of independent SNP-based variations, some of which contribute to a measured phenotype, may be used as markers directly or may be organised into haplotype markers. Provided that the marker set includes all the causative SNPs, we find a simple rule for the relative power of SNP and haplotype markers: SNP-based tests have greater power when the number of causative SNPs (a subset of the total set of SNPs) is smaller than the total number of haplotypes. Furthermore, we find that regression tests for the simple main effect of each haplotype are generally more powerful than ANOVA tests applied to haplotype pairs. A review of recent literature supports our findings.

Keywords: *analysis of variance (ANOVA), association, haplet, haplotype, linkage disequilibrium, pharmacogenetics, quantitative trait locus (QTL), single nucleotide polymorphism (SNP)*

Pharmacogenomics (2001) ***2****(1):11-24*

1. Introduction

Identifying the genetic components of complex traits is one of the primary goals of the human genome project. Cancer, metabolic disorders, such as diabetes and obesity, cardiovascular disorders, such as hypertension and stroke, and psychiatric disorders typify these traits caused by multiple genetic and environmental factors. Often, the pharmacogenetics of drug response is also under the control of multiple genetic factors. A better understanding of complex traits would permit stratification of patient populations presenting a single disease phenotype into sub-classes whose disorders might have differing genetic components or different responses to particular therapeutics.

Studies identifying the underlying genetic variations that cause increased disease risk or affect drug response have depended on the availability markers spaced throughout the genome. Early physical maps of the human genome with markers spaced 10^6 - 10^8 nucleotides (nt) required family-based studies and linkage analysis to identify disease-related loci [1]. Although these types of studies have identified causative mutations for

monogenic disorders, they have not been as successful in identifying genetic components for complex, polygenic traits, which by definition lack the Mendelian inheritance patterns, strong penetrance and allelic homogeneity required for linkage methods.

More recently, SNPs have been suggested as an alternative marker set. These single nucleotide substitutions or deletions are typically biallelic variants and occur at sufficient density to permit whole-genome association studies in outbred populations [2]. Linkage disequilibrium (LD) is anticipated to extend 5000 - 100,000 nt [3-5], varying irregularly across the genome and implying that hundreds of thousands of individual SNPs will be required for a whole-genome scan. In order to correct for multiple hypothesis testing, a significance level of 10^{-8} - 10^{-9} has been suggested, which implies a sample size of several thousand individuals is required for adequate power to detect association [6-8]. Although the cost of genotyping can be reduced by testing allele frequency differences between pools of DNA collected from individuals with extreme phenotypes [9,10], these tests are necessarily less powerful than individual genotyping and require even larger population sizes.

While population sizes sufficiently large for full-genome scans may be obtained for unselected populations to study genetic risk factors for disease predisposition [11,12] and for specific diseases [13], these numbers are substantially larger than the enrolment of typical Phase I, Phase II and even Phase III trials. A reasonable strategy to reduce the sample sizes required for pharmacogenomic studies is to restrict attention to polymorphisms residing in a small set of candidate genes representing, perhaps, known drug targets and other genes in disease and drug response pathways. By sequencing a gene in 100 individuals and thereby scanning 200 chromosomes, one has 95% probability of identifying any genetic polymorphism with an allele frequency above 1.5%. These polymorphisms, usually SNPs, should include most of those that reside in the candidate region, affect the phenotype and have population-level significance. The SNP-based markers may also be used directly for association tests. Haplotypes or diploid haplotype pairs, constructed from the individual genotypes, constitute an alternative set of markers for an association test.

Haplotype-based tests have been suggested for use in clinical studies [14]. Nevertheless, haplotype-based tests require additional expense relative to SNP-based tests, including direct sequencing or computational inference to identify haplotypes, and for now preclude less costly tests of pooled DNA. With the interest in haplotype-based tests growing, experimentalists weighing the relative merits of SNP-based and haplotype-based tests or choosing between tests based on haplotypes or haplotype pairs may benefit from theoretical guidance based on realistic genetic models.

Here we analyse the relative power of association tests based on SNPs and haplotypes. Our focus is on quantitative traits characteristic of disease risk or clinical response and our goal is to provide simple, analytical estimates of the relative efficiency of SNP-based and haplotype-based tests. In Section 2, we provide a statistical model for a quantitative phenotype and review the power of SNP-based tests to detect association. We extend this analysis to haplotype-based regression tests and haplotype-pair analysis of variance (ANOVA) tests. The expected p-values and the sample sizes required to detect association are presented for each type of test. In Section 3, we review literature reports applying both SNP-based and haplotype-based tests to experimental and simulated data. We conclude in Section 4 with a summary and our recommendations.

2. Statistical model

The salient assumptions underlying the statistical model developed here are as follows:

- Every causative polymorphism in the candidate genetic region is a SNP (biallelic for notational simplicity). The SNP marker set includes each causative SNP as well as many other neutral polymorphisms.
- The causative SNPs make additive contributions to a quantitative phenotype; epistasis is negligible.
- The phenotypic variability due to any particular SNP is small relative to the overall phenotypic variability from residual genetic factors, including other markers in the set and environmental factors; selective pressure on SNPs or haplotypes is negligible.

This model describes any degree of LD by adjusting the total number of haplotypes. The extremes are complete disequilibrium, with two haplotypes, and complete equilibrium, with 2^T haplotypes, where T is

the total number of SNP markers. According to the third assumption, however, disequilibrium at the marker loci is not under phenotypic selection. This implies that pairs of alleles of different SNPs that shift the phenotypic value in the same direction are as likely to be found in the same haplotype as pairs of alleles that shift the phenotypic value in opposite directions.

Based on these assumptions, we derive an analytical expression for the relative power of SNP-based and haplotype-based tests. In Section 4, we discuss the limitations of these assumptions and indicate how generalising these assumptions would modify our results.

2.1 SNP-based phenotype model

A variance components model is used to describe the dependence of an individual's phenotype on its genotype [15]. This quantitative model may also be applied to a haplotype relative risk model for disease susceptibility in which the risks from haplotypes are multiplicative and each risk factor is proportional to an exponential of an underlying quantitative trait [16].

In the variance components model, the quantitative phenotype denoted X is standardised to have zero mean and unit variance. Several quantitative trait loci, here modelled as biallelic markers or SNPs, are assumed to contribute to the phenotypic value. Individual SNPs may occur within the same gene and the total number of SNPs is T. A subset G of the total number are causative, defined as contributing an allele-specific shift to the phenotypic value. The alleles for a particular SNP γ, $\gamma = 1$ to T, are labelled $A_{\gamma 1}$ and $A_{\gamma 2}$, with respective frequencies p_γ and $1 - p_{\gamma 1}$ in an unselected population. Hardy-Weinberg equilibrium is assumed separately for each SNP (which does not preclude LD between SNPs) and the probabilities of the genotypes $A_{\gamma 1}A_{\gamma 1}$, $A_{\gamma 1}A_{\gamma 2}$ and $A_{\gamma 2}A_{\gamma 2}$ are therefore p_γ^2, $2p_\gamma(1 - p_\gamma)$ and $(1 - p_\gamma)^2$. The frequency of allele $A_{\gamma 1}$ for each individual is either 1, 0.5 or 0 and is denoted f_γ. The variance of f_γ is $\sigma^2_{f_\gamma}$, with

$$\sigma^2_{f_\gamma} = 1\times p_\gamma^{\,2} + (1/4)\times 2p_\gamma(1-p_\gamma) + 0\times(1-p_\gamma)^2 - p_\gamma^{\,2}$$
$$= p_\gamma(1-p_\gamma)/2.$$

The effect of allele $A_{\gamma 1}$ is assumed to be purely additive, a shift of $a_\gamma/2$ for each copy inherited. The shifts in phenotypic value are therefore $a_\gamma - \mu_\gamma$ for the $A_{\gamma 1}A_{\gamma 1}$ homozygote, $-\mu_\gamma$ for the heterozygote and $-a_\gamma - \mu_\gamma$ for the $A_{\gamma 2}A_{\gamma 2}$ homozygote, where the constant $\mu_\gamma = a_\gamma(2p_\gamma - 1)$ ensures that X has zero mean. This SNP contributes a phenotypic variance of σ^2_γ,

$$\sigma_\gamma^{\,2} = 2p_\gamma(1-p_\gamma)a_\gamma^{\,2}$$

to the total phenotypic variance of 1. For a polygenic trait, the variance σ_γ^2 contributed by any individual SNP is small compared to the residual variance $\sigma_R^{\,2} = 1 - \sigma_\gamma^{\,2} \approx 1$ from other genetic and environmental factors. The expected value of $\sigma_\gamma^{\,2}$ for a causative SNP is defined as $\sigma_G^{\,2}$,

$$\sigma_G^{\,2} = G^{-1}\sum_{\gamma=1}^{G}\sigma_\gamma^{\,2},$$

the mean of the individual variances. The fractional variance explained by all the SNPs together is $G\sigma_G^{\,2}$. This total may be much smaller than 1 if environmental factors and genetic factors outside the candidate region are important; it approaches the genetic heritability when the entire genome is considered.

Note that if the effect of a particular SNP is not purely additive, an additive effect can nevertheless be constructed by defining a_γ as half the difference in phenotypic shift between $A_{\gamma 1}$ and $A_{\gamma 2}$ homozygotes minus $(2p_\gamma - 1)d_\gamma$, where d_γ is the difference between the phenotype shift for heterozygotes and the midpoint of the shifts for homozygotes. This approach is generally valid for alleles with dominant, recessive, or multiplicative effects; it fails only for very rare recessive alleles and, correspondingly, for very common dominant alleles. However, in these extreme cases, the additive variance vanishes and association is difficult to detect without recourse to highly selected populations.

2.2 Haplotypes

The T individual SNPs may occur in up to 2^T distinct allelic combinations. Due to LD, however, a test population often exhibits only a smaller subset of H haplotypes. Using η to label the haplotype, $\eta = 1$ to H, the phenotypic shift for an individual with haplotypes η and η' is defined in analogy to the SNP shifts as $(a_\eta + a_{\eta'})/2$, where

$$a_\eta = \sum_{\gamma=1}^{G}[I(A_{\gamma 1}|\eta) - I(A_{\gamma 2}|\eta) - (2p_\gamma - 1)]a_\gamma$$

The indicator function $I(A_{\gamma 1}|\eta)$ has value 1 if haplotype η has allele $A_{\gamma 1}$ and is 0 otherwise.

Similarly, $I(A_{\gamma 2}|\eta) = 1$ if haplotype η has allele $A_{\gamma 2}$ and is 0 otherwise. The difference in these terms, either +1 or -1, less its mean value $2p_\gamma - 1$, multiplies a_γ to yield the phenotypic shift in haplotype η due to the phase of SNP γ and is summed over the G causative SNPs.

While the precise value of a_η depends on the particular alleles occurring in haplotype η, the distribution of values of a_η may be estimated by considering the term $I(A_{\gamma 1}|\eta) - I(A_{\gamma 2}|\eta)$ to be a random variable taking the value +1 with probability p_γ and the value -1 with probability $1 - p_\gamma$. This mean probability approximation recovers the SNP allele frequencies p_γ and ensures that the mean of a_η is zero. The variance $\mathrm{Var}(a_\eta)$ may be obtained under a random phase approximation in which the directions of the shifts a_γ are uncorrelated. With this assumption, the variance of the sum over SNPs is the sum of the individual variances even if the SNPs are in LD. The contribution of SNP γ to the variance of the haplotype phenotypic shift a_η is

$$\begin{aligned} & p_\gamma[1-(2p_\gamma-1)]^2 a_\gamma^{\,2} + (1-p_\gamma)[-1-(2p_\gamma-1)]^2 a_\gamma^{\,2} \\ &= 4p_\gamma(1-p_\gamma)a_\gamma^{\,2} \\ &= 2\sigma_\gamma^{\,2}. \end{aligned}$$

The final variance for the distribution of haplotype-dependent shifts a_η is

$$Var(a_\eta) = 2G\sigma_G^{\,2}$$

where σ_G^2 is the mean variance of causative SNPs as previously defined.

The mean phenotypic shift contributed by haplotype η is $p_\eta^{\,2} a_\eta + 2p_\eta(1-p_\eta)(a_\eta/2)$, or simply $p_\eta a_\eta$. The phenotypic variance contributed by this haplotype is defined as $\sigma_\eta^{\,2}$:

$$\begin{aligned} \sigma_\eta^{\,2} &= p_\eta^{\,2} a_\eta^{\,2} + 2p_\eta(1-p_\eta)(a_\eta/2)^2 \\ &= (1/2)p_\eta(1-p_\eta)a_\eta^{\,2}. \end{aligned}$$

When the number of haplotypes is large, the probability p_η for each haplotype is small and $\sigma_\eta^{\,2} \approx p_\eta a_\eta^{\,2}/2$. The mean value of $\sigma_\eta^{\,2}$ is defined as $\sigma_H^{\,2}$:

$$\sigma_H^{\,2} = H^{-1}\sum_{\eta=1}^{H} \sigma_\eta^{\,2} = H^{-1}\sum_{\eta=1}^{H} p_\eta a_\eta^{\,2}/2 = (G/H)\sigma_G^{\,2}$$

where we have assumed that p_η and a_η are uncorrelated. Note that the total haplotype-based phenotypic variance, $H\sigma_G^{\,2}$, equals the total SNP-based phenotypic variance, $G\sigma_G^{\,2}$.

In the special case when $G = 1$ and only one of the SNPs has a non-zero phenotypic shift a_γ, each haplotype η will have a phenotypic shift a_η of either $2(1 - p_\gamma)a_\gamma$ or $-2p_\gamma a_\gamma$, depending on whether $A_{\gamma 1}$ or $A_{\gamma 2}$ is included. The corresponding values for $\sigma_\eta^{\,2}$ will be $p_\eta(1-p_\eta)\sigma_\gamma^{\,2}$ multiplied by either $p_\gamma/(1 - p_\gamma)$ or $(1 - p_\gamma)/p_\gamma$. Assuming that $A_{\gamma 1}$ is the minor allele with frequency p_γ, the haplotypes with the largest effect have variance

$$\sigma_\eta^{\,2} = p_\eta(1-p_\eta)[p_\gamma/(1-p_\gamma)]\sigma_\gamma^{\,2}.$$

Under the assumption of minimal selection, the SNP minor allele frequency should be close to 50% and the haplotype frequency should be closer to $1/H$. The variance then simplifies to $(1/H)\sigma_\gamma^{\,2}$, the same result as $(G/H)\sigma_G^{\,2}$ because $G = 1$.

2.3 Regression test for association

A suitable test statistic for association of either a SNP-based or haplotype-based marker with a quantitative phenotype is the coefficient b_1 for a regression model of the phenotypic value on the marker dose [17]:

$$X_i = b_1 \delta f_i + \varepsilon_i.$$

The N individuals included in the sample are specified by the index i. The difference between the marker frequency in individual i and in the total sample is δf_i and the residual ε_i is uncorrelated with δf_i. Under the null hypothesis, the variance of ε_i is equal to the total phenotypic variance of 1. Under the alternative hypothesis, the expected value for b_1 is

$$b_1 = \sigma_M / \sigma_f$$

where M is a generic index representing one of T total SNPs for a SNP-based test or one of H haplotypes for a haplotype-based test. The additive variance of the marker is $\sigma_M^{\,2}$, either $\sigma_\gamma^{\,2}$ for a SNP-based test or $\sigma_\eta^{\,2}$ for a haplotype-based test. The variance of the marker frequency is $\sigma_f^{\,2}$ and equals $p(1 - p)/2$ for a marker with frequency p under Hardy-Weinberg equilibrium. Since the variance of ε_i is close to 1 when $\sigma_M^{\,2}$ is small, the variance of the estimator for b_1, $\sigma_b^{\,2}$, is the same under the null hypothesis, $b_1 = 0$, and the alternative hypothesis, $b_1 > 0$, and

$$\sigma_b^2 = 1/(N\sigma_f^2)$$

for a one-sided test. Using a two-sided test would be more appropriate for a quantitative phenotype in which extreme high and low values are equally relevant, but would not materially affect any of our conclusions.

Combining the expected value for the regression coefficient with the standard deviation of the estimator, the expected p-value for a one-tailed test for a marker with additive variance σ_M, using a Bonferroni correction for M multiple tests, is

$$\mathrm{p-value} = 1-[\Phi(N^{0.5}\sigma_M)]^M. \quad \textbf{(1)}$$

The asymptotic expansion for $\Phi(z)$ yields

$$\mathrm{p-value} \approx M(2\pi N\sigma_M^2)^{-0.5}\exp(-N\sigma_M^2/2)$$

as an approximation valid for small p-values.

The relative significance of haplotype-based regression tests and SNP-based regression-tests may be obtained from this asymptotic expansion as

$$\mathrm{p-value}(HAP)/\mathrm{p-value}(SNP) = (H/G)^{1/2}(H/T)\exp[N\sigma_G^2(1-G/H)/2].$$

Setting the value of this ratio to 1 and solving for H yields an approximate expression for the number of haplotypes at which the significance for the haplotype-based test and the SNP-based test are identical:

$$H \approx G[1+2\ln(T/G)/(3+N\sigma_G^2)]$$

to lowest order. If all the SNPs are causative, G and T are identical, and the expected significance levels of SNP-based and haplotype-based tests cross when $H = G$. When T is larger than G, there is a small logarithmic correction and the value of H at the cross-over is slightly larger than G. This correction can be very small even when T is much larger than G because the logarithmic term $\ln(T/G)$ is further reduced by the term $N\sigma_G^2$, approximately equal to the square of the z-score corresponding to a significant finding (see below). For a test of ten markers, for example, a p-value of 0.005 corresponding to a final false-positive rate of 5% implies a z-score of 2.58 and $N\sigma_G^2 \sim (2.58)^2 \sim 7$. Thus, to a good approximation, the cross-over between the significance of SNP-based tests and haplotype-based tests occurs when the number of haplotypes is just larger than the number of causative SNPs.

We now calculate the population sizes required for adequate power to detect true associations. For a corrected final Type I error rate of α, the uncorrected p-value for a significant finding must be smaller than α/M. The Type II error rate β has no multiple testing correction. Defining the normal deviates $z_{\alpha/M} = \Phi^{-1}(1-\alpha/M)$ and $z_{1-\beta} = \Phi^{-1}(\beta)$, the resulting sample size required to detect a marker contributing phenotypic variance σ_M^2 with power 1-β is

$$N_{REGR} = (z_{\alpha/M} - z_{1-\beta})^2/\sigma_M^2. \quad \textbf{(2)}$$

A simplified approximation for the sample size may be obtained by noting that $z_{\alpha/M}$ is typically larger than $z_{1-\beta}$. When $\alpha = 0.05$, $M = 10$ and $1-\beta = 0.8$, for example, $z_{\alpha/M} = 2.58$ while $z_{1-\beta} = -0.84$. Neglecting $z_{1-\beta}$ relative to $z_{\alpha/M}$, which is equivalent to setting the power to 50%, yields

$$N \approx 2\ln(M/\alpha)/\sigma_M^2.$$

The asymptotic expansion $z_\alpha \sim 2\ln(1/\alpha)$, valid for small α, is the source of the logarithmic term $\ln(M/\alpha)$ in this expression.

The relative population sizes required for haplotype-based and SNP-based regression tests may be obtained from this asymptotic expansion as

$$N(HAP)/N(SNP) = (H/G)\ln(H/\alpha)/\ln(T/\alpha).$$

Again, the cross-over in power between haplotype-based tests and SNP-based tests occurs close to $H = G$ with a logarithmic correction shifting the exact location to a slightly larger value. A low-order approximation for the cross-over is

$$H \approx G\{1+\ln(T/G)/[1+\ln(T/\alpha)]\}.$$

Since $G \geq 1$ and $\alpha < 1$, the cross-over occurs in the region $G \leq H < 2G$.

Table 1: Summary of association tests.

Marker type	SNP	Haplotype	Haplotype pair
Test	Regression	Regression	ANOVA
Number of markers	G causative SNPs T total SNPs	H	H
Phenotypic variance explained by markers	$G\sigma_G^2$	$H\sigma_H^2$	$H\sigma_H^2$
Observed variance per causative marker	σ_G^2	$\sigma_H^2 = (G/H)\sigma_G^2$	σ_H^2
p-value for N individuals	$1-[\Phi(N^{0.5}\sigma_G)]^T$	$1-[\Phi(N^{0.5}\sigma_H)]^H$	$1-\{\Phi[2(NJ/H)^{0.5}\sigma_H]\}^C$ with J = 1, 1.5 or 2; $C = K(K\text{-}1)/2$; and $K \approx H(H+1)/2$
N for Type I error α and power 1-β	$(z_{\alpha/T} - z_{1-\beta})^2 / \sigma_G^2$	$(z_{\alpha/H} - z_{1-\beta})^2 / \sigma_H^2$	$(z_{\alpha/C} - z_{1-\beta})^2 H / 4J\sigma_H^2$

2.4 ANOVA test for haplotype association

ANOVA may also be used to test for association between haplotype pairs and a quantitative phenotype. In a typical ANOVA test, N individuals are sorted into $K = H(H+1)/2$ distinct haplotype pairs and the between-haplotype-pair phenotypic variance is compared to the within-haplotype-pair phenotypic variance. A significant finding in an ANOVA test is approximately equivalent to detecting a significant difference in mean phenotype value for at least one of the $C = K(K\text{-}1)/2$ possible pair-wise comparisons. The most significant finding will typically arise from the difference Δ in mean phenotypic value between the pair of genotypes with the most extreme positive and negative shifts.

The expected maximum difference Δ is obtained from the distribution of a_η as

$$\Delta = 2[\mathrm{Var}(a_H)]^{0.5} \text{ or } (8H\sigma_H^2)^{0.5}.$$

The variance for this test statistic is

$$\sigma^2 = \sigma_R^2[(1/n)+(1/n')]$$

where n and n' are the number of individuals in the total sample size of N in the two extreme classes. By earlier assumption, the residual phenotypic variance σ_R^2 is close to 1. Under the assumption of weak selection, each p_η is $1/H$. If the most extreme phenotypic shifts correspond to homozygous genotypes, then n and n' are both approximately N/H^2 and the variance is $\sigma^2 = 2H^2/N$. If the genotypes with extreme phenotype values are both heterozygous, the variance is H^2/N. The additive model suggests that homozygotes will be at least tied for the maximum phenotypic shift. The p-value for the comparison of extreme phenotypes is

$$\begin{aligned} \text{p-value} &= 1-[\Phi(\Delta/\sigma)]^C \\ &= 1-[\Phi(2\sigma_H N^{0.5} J^{0.5} / H^{0.5})]^C \end{aligned}$$

(3)

where the factor of C is the correction for multiple hypothesis testing and J = 1 if homozygotes are extreme, 2 if heterozygotes are extreme and 1.5 if one homozygote and one heterozygote are extreme.

As with the regression test, the residual variance is close to 1 and an expression yielding the required sample size is $1/\sigma^2 = (z_{\alpha/C} - z_{1\text{-}\beta})^2/\Delta^2$,

$$N_{ANOVA} = (z_{\alpha/C} - z_{1-\beta})^2 H / 4J\sigma_H^2.$$

(4)

The ratio $N_{\mathrm{ANOVA}}/N_{\mathrm{REGR}}$ of the sample size required for an ANOVA test, relative to that required for a series of H regression tests, is obtained from the ratio of **Equation 4** to **Equation 2**. An estimate for this ratio, valid when $z_{\alpha/C}$ and $z_{\alpha/H}$ are both large compared to $z_{1\text{-}\beta}$, is

$$N_{ANOVA} / N_{REGR} \approx (H/4J)\ln(C/\alpha)/\ln(H/\alpha).$$

The logarithmic dependence varies slowly and the factor $H/4J$ explains most of the relative efficiency. When the number of haplotypes is small, ANOVA is more powerful. A cross-over occurs near H = 4 if homozygotes are extreme and near H = 8 if heterozygotes are extreme. Beyond the cross-over, the regression test is more powerful.

2.5 Comparison of tests using SNPs and haplotypes

In this section, the analytical theory developed above and summarised in **Table 1** is used to explore the relative power of SNP-based regression tests, haplotype-based regression tests and haplotype-pair-based ANOVA tests.

For comparisons between SNP-based and haplotype-based tests, the total number of SNPs, *T*, is fixed at 20. The number of causative SNPs, *G*, is varied from 1 to 10 and the number of haplotypes, *H*, is varied from 1 to 100. Under these general constraints, chosen to characterise a typical quantitative trait locus (QTL), we examine two models. In the first model, with results depicted in **Figure 1**, the additive variance per SNP is fixed at 0.025, implying a total additive variance that varies from 0.025 (one causative SNP) to 0.25 (ten causative SNPs). In the second model, with results depicted in **Figure 2**, the total additive variance is fixed at 0.075, implying an additive variance per SNP that varies from 0.075 (one causative SNP) to 0.0075 (ten causative SNPs). These two models are identical when 3 of the 20 SNPs are causative.

In **Figure 1**, the top series of panels illustrates the expected significance for a fixed population size of 300 and the bottom series illustrates the population size required to attain a p-value of 0.05 (5% false-positive rate including the multiple-testing correction) and a power of 0.8 (20% false-negative rate) for the haplotype-pair ANOVA test (dot-dashed line), the haplotype regression test (dashed line) and the SNP regression test (solid line). The left panels correspond to one causative SNP, the middle panels correspond to three causative SNPs and the right panels correspond to ten causative SNPs.

The top middle panel shows that the SNP-based regression test yields a p-value of approximately 0.06 when the population size is fixed at 300. When there are only two haplotypes, the haplotype-pair ANOVA test yields the most significant finding. For three or four haplotypes, the haplotype regression test is the most significant. For five or more haplotypes, the SNP-based test is more significant than either haplotype-based test. The cross-over between SNP-based markers and haplotype-based markers occurs close to the number of causative SNPs, three in this example.

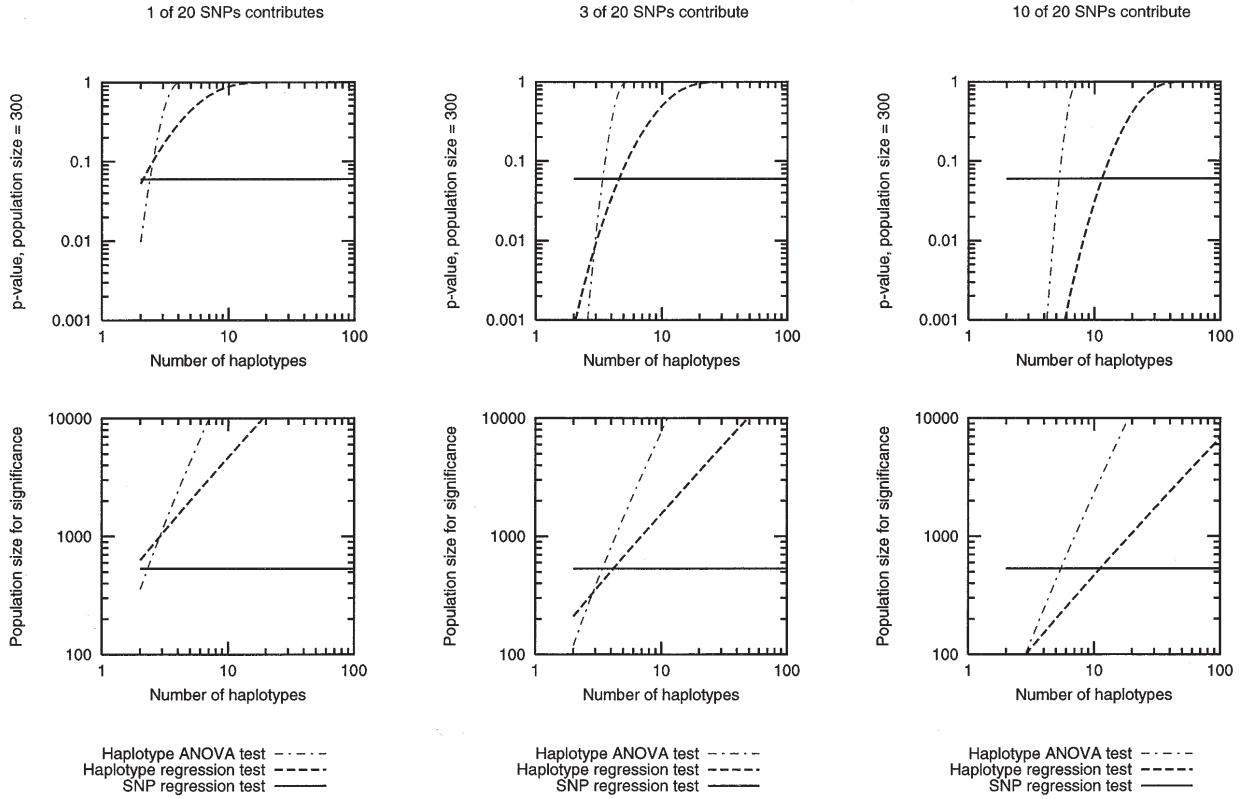
Because the additive variance per SNP is constant in **Figure 1**, the results for SNP-based markers do not depend on the number of causative SNPs; consequently, the solid lines showing the SNP results are in identical positions in the left, middle and right panels. The haplotype results do depend on the number of causative SNPs, as does the location of the cross-over between haplotype-based tests and SNP-based tests. In the left panel, for one causative SNP, the cross-over occurs between two and three haplotypes; in the middle panel, for three causative SNPs, the cross-over occurs between four and five haplotypes; and in the right panel, for ten causative SNPs, the cross-over occurs between ten and eleven causative SNPs. These results agree with the analytical approximation that the cross-over occurs when *H* is just larger than *G*.

Comparing the two haplotype tests, it is evident that the ANOVA test for haplotype pairs is usually less powerful than the regression test for haplotypes. The cross-over occurs when the number of haplotypes is close to three.

The bottom series of panels of **Figure 1** illustrates the same behaviour in terms of the population size required for significance. Within each panel, as the number of haplotypes increases from left to right, the population size required for haplotype-based tests increases while that required for SNP-based tests does not vary. Across the three panels, as the number of causative SNPs increases from one to ten, haplotype-based tests require smaller population sizes. This occurs because having more causative SNPs than haplotypes implies that some haplotypes must contain multiple causative SNPs, making these haplotypes easier to detect. As before, the cross-overs between SNP-based and haplotype-based tests occur when the number of haplotypes is close to the number of causative SNPs in the model.

In **Figure 2**, the top panels again depict p-values and the bottom panels again depict population sizes required for significance, and the left, middle and right panels again show results for one, three and ten causative SNPs. A constant additive variance is divided equally among the causative SNPs and SNP-based tests are consequently more powerful when the number of causative SNPs is smaller. Thus, the p-values and population sizes for SNP-based tests show an unfavourable increase from the left to middle to right panels. Within each panel, the results for SNP-based tests are not sensitive to the number of haplotypes included in the model, as was observed in **Figure 1**. Since the total variance arising from the

Figure 1: For comparisons between SNP-based and haplotype-based tests, the total number of SNPs is fixed at 20. The number of causative SNPs is 1 (left panels), 3 (middle panels) or 10 (right panels). The number of haplotypes, *H*, is varied from 1 to 100 within each panel. The additive variance per SNP is fixed at 0.025. The top series of panels illustrates the expected significance for a fixed population size of 300 and the bottom series illustrates the population size required to attain a p-value of 0.05 (5% false-positive rate including the multiple-testing correction) and a power of 0.8 (20% false-negative rate) for the haplotype-pair ANOVA test (dot-dashed line), the haplotype regression test (dashed line) and the SNP regression test (solid line). Haplotype-based tests and SNP-based tests cross in power when the number of haplotypes is just larger than the number of causative SNPs.

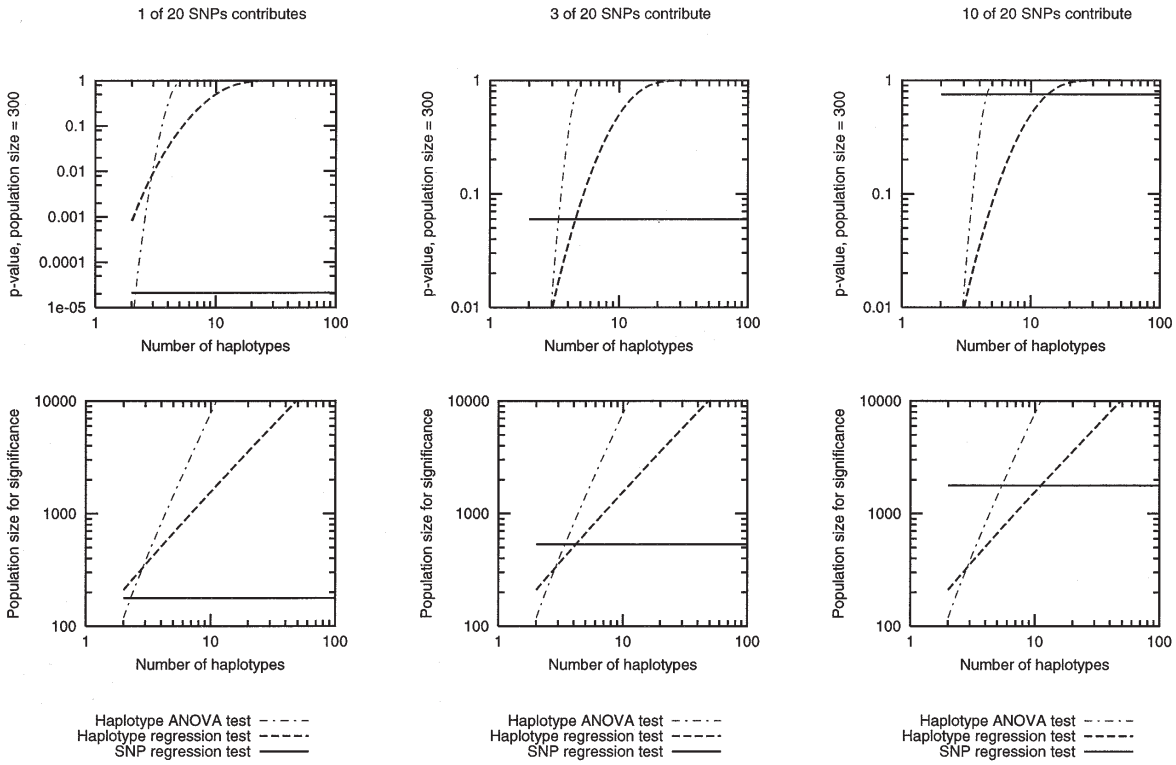


genetic markers is held constant, the haplotype-based tests are not sensitive to the number of causative SNPs and the haplotype results are constant in the left, middle and right panels (note the change of scale in the top left panel). Within each panel, the haplotype-based tests become less favourable as the number of haplotypes increases from left to right.

In **Figure 2**, as in **Figure 1** before, the cross-over in power between haplotype-based tests and SNP-based tests occurs when the number of haplotypes is just larger than the number of causative SNPs. In the left panels with one causative SNP, the cross-over occurs between two and three haplotypes; in the middle panels with three causative SNPs, the cross-over occurs between four and five haplotypes; and in the right panels with ten causative SNPs, the cross-over occurs between ten and eleven haplotypes. Furthermore, the cross-over between the haplotype-pair ANOVA test and the haplotype regression test occurs between two and three haplotypes, also as was seen in **Figure 1**.

We note finally that, in a previous investigation of association tests in a different context, we compared results from analytic approximations as in **Table 1**, from numerical solutions to the non-linear equations appropriate for the asymptotically normally distributed test statistic and from exact numerical calculations based on the true underlying multinomial distribution of genotype frequencies sampled from a population [18]. When population sizes required for specified selectivity and sensitivity were calculated by the three methods, the relative differences were usually quite small, below 5%. In the examples provided here, the analytical approximations are

Figure 2: Same as **Figure 1**, except the total additive variance is fixed at 0.075, implying an additive variance per SNP that varies from 0.075 (1 causative SNP) to 0.0075 (10 causative SNPs). The middle panels of **Figure 1** and **Figure 2** share identical parameters. Haplotype-based tests and SNP-based tests cross in power when the number of haplotypes is just larger than the number of causative SNPs.



effectively expansions ordered by the additive variance per marker and are likewise expected to be accurate to 5%, except when the additive variance per marker is large, on the order of 0.05 or greater. Effects of this magnitude occur only at the extremes of the examples provided: in **Figure 1**, the rightmost panel for haplotype-based markers with five or fewer haplotypes, and in **Figure 2**, the leftmost panel for SNP-based markers. In the present context of complex traits, more accurate descriptions of these extreme cases are not necessary because the strong genetic effects approach major gene effects that are readily detected even in small populations.

3. Review

As the use of SNPs and SNP-based haplotypes for association tests is still in its early stages, the number of published reports comparing these methods is limited. Here we focus on three recent examples: two experimental studies and one simulation study.

The first example is an analysis of SNPs around the apolipoprotein E (*APOE*) gene by Martin *et al.* [19]. The APOE-4 allele has been demonstrated to increase the risk for developing late-onset Alzheimer's disease (AD) [20-23]. As part of a multifaceted study, Martin *et al.* compared the p-values obtained for linkage of six SNPs with AD using one-locus, two-locus and three-locus tests. In one series of tests, the single causative mutation was included in the marker set. A single-locus test, corresponding to a SNP-based test, readily identified this mutation. Multiple-locus tests, corresponding to haplotype-based tests, did not enhance the significance substantially. Only in a second series of tests, excluding the causative mutation from the marker set, did they find a clear advantage for the multiple-locus tests. The likelihood-based method they employed made use of familial information [24] and is therefore not an exact

equivalent of the association tests considered here. Nevertheless, their results are consistent with our predictions when the causative mutation is included as a marker: since there are many haplotypes but only a single causative mutation, the single-locus test should provide greater power than the multiple-locus tests.

A second example concerns studies of the β_2-adrenergic receptor (β_2AR). This G-protein coupled receptor is expressed in airway smooth muscle cells and mast cells and is the target of bronchodilating β-agonists, such as isoprenaline, salmeterol and albuterol used in the treatment of asthma [25]. Polymorphisms at codons 16 (Arg to Gly) and 27 (Gln to Glu) have been associated with response to β-agonist treatment [26-29]. Between the β_2AR transcription start site and the intronless coding region is a 5′-leader cistron that encodes a 19-aa peptide. Polymorphisms in this region have been shown to affect β_2AR expression [30]. To understand the relevance of these and other polymorphisms in β_2AR, Liggett and co-workers undertook an association study focusing on the relationship between SNPs, haplotypes and response to the bronchodilator albuterol [31].

This most recent Liggett study examined chromosomes from 23 Caucasians, 19 African-Americans, 20 Asians and 15 Hispanic-Latinos. A total of 13 polymorphic sites occurred in a region including ~ 700 nt of ORF and ~ 1100 nt of 5′ UTR, including the 5′-leader cistron. While 12 total haplotypes were identified, only four had frequency above 5% in any ethnicity and only three of these occurred at 2% frequency or greater in the Caucasian population. In these three haplotypes, 10 of the 13 SNPs were variable. The SNPs and haplotypes were then tested for association with albuterol response, adjusted for sex and baseline severity, in a population of 121 Caucasian patients with moderate asthma. A haplotype association test was performed using ANOVA for the five haplotype pairs observed in the treated population and SNP main effects were tested using ANOVA for SNP genotypes with p-values corrected for multiple hypothesis testing. While the haplotype-based test yielded a significant finding at a p-value of 0.007, none of the SNP-based tests was significant at a p-value of 0.05. The parameters we use to analyse these findings are $H = 3$ haplotypes, $T = 10$ of the 13 SNPs which vary in these haplotypes and $C = 10$ possible pairwise comparisons between the five haplotype pairs.

Using **Equation 3**, the characteristic haplotype contribution to the phenotypic variance, σ_H^2, may be estimated from the haplotype-based ANOVA to be 0.063. Had haplotype-based regression been performed instead of ANOVA, we predict using **Equation 1** that a p-value of 0.008 would have been observed. Although the small number of haplotypes suggests strong LD between SNPs, sequence data presented by Martin and co-workers demonstrates that correlation between SNPs extends no further than one or two SNPs, in accord with their observation that no SNP correlated perfectly with any haplotype. We next estimated the expected p-value from a SNP-based regression test. The p-value resulting from **Equation 1**, corrected for multiple hypothesis testing, depends on the number of SNPs assumed to be causative. If only two of the SNPs are causative, the estimated p-value is significant at 0.0036. If five are causative, however, the p-value is 0.15 and if all ten are causative, the p-value is 0.49. Thus, with multiple causative SNPs, the SNP-based associations may not rise to significance. Therefore, the Liggett study may be consistent with simple additive effects from multiple causative SNPs; unique or non-additive interactions between SNPs may not be required to explain the data. A larger population might be required to verify the existence of non-additivity because this series of experiments, with insufficient power to detect the simple main effect of individual SNPs, is unlikely to have sufficient power to detect the interaction terms in an ANOVA model. Similarly, although a model including haplotype main effects and haplotype-haplotype interactions would be expected to yield significance for the main effects, it is unlikely that the interaction terms would be significant.

The final example is a report by Long and Langley of a series of simulations designed to assess the power of various association studies [32]. Although the details of the simulation model, including the use of haploid rather than diploid genomes for estimates of the power of haplotype-based association studies, are different from the model we consider, the essence of the model is the same: multiple polymorphic markers exist in LD with each other and with a quantitative trait nucleus that is not under selective pressure. Long and Langley report, based on their simulations, that tests which consider each single marker in turn have power similar to or greater than haplotype-based tests. We reach the same conclusion with our analytical results,

provided that the total number of haplotypes is larger than the total number of SNPs.

Long and Langley also investigate the effects of increasing marker density relative to a parameter $4Nc$, a measure of the extent of LD along a chromosome. Once the marker density is comparable to the inverse of this length, the simulation results suggest that it is more powerful to increase the number of individuals genotyped than to increase the number of markers tested. Our findings are similar, with the extent of LD expressed in terms of the relative number of SNPs and haplotypes.

4. Conclusion and expert opinion

We have examined the power of association studies using regression tests and ANOVA to identify SNP-based and haplotype-based markers for quantitative traits. Our primary result is that SNP-based tests are more powerful when the number of causative SNPs is smaller than the number of haplotypes and the SNP marker set includes the causative polymorphisms. Since only one or two causative SNPs might be expected in a typical QTL and the number of haplotypes is almost certainly larger, SNP-based tests may be preferred when genotyping studies are preceded by a search for high-frequency genetic variants in candidate regions.

Simple formulae estimate the sample size requirements and p-values for SNP-based and haplotype-based tests. When haplotypes are the preferred marker set, results derived from analytical theory indicate that ANOVA tests of haplotype pairs should only be used when the number of haplotypes is very small. When the number of haplotypes increases beyond four or five, a regression test of haplotype main effects has greater power.

These predictions agree with literature reports. We also support simulation findings that increasing the sample size of a study is more important than increasing the number of SNPs once the density of SNPs is comparable to the length scale of LD.

We now revisit the main assumptions of our model and describe how generalisations may shift the balance between SNPs and haplotypes. The first assumption is that every causative polymorphism is in the marker set. If this assumption is violated, detecting an association requires LD between the causative polymorphism and the SNPs or haplotypes available as markers and the relative power of the two marker sets depends on the details of the genetic model. At one extreme, a non-causative SNP marker may be in complete LD with a causative polymorphism missing from the marker set, essentially reproducing the results obtained here that tend to favour SNPs as markers. At the other extreme, it is possible to specify a model in which no individual SNP is linked to a causative polymorphism while haplotypes show complete linkage, favouring haplotype markers by necessity. Real-word examples will fall between these extremes and the experience of Martin *et al.* indicates that a reasonable compromise may be to use multiple-locus tests, with the number of loci included being a function of the decay of LD across individual SNPs.

In this context, we note that LD implicit in the existence of haplotypes also indicates that, under the assumptions of the model, causative SNPs that are in LD with each other may show enhanced correlation with a quantitative phenotype. This effect is the basis of our results for the power of haplotype-based markers but, for simplicity, was neglected in obtaining results for SNP-based markers. Including these effects for SNP-based markers enhances their effective power. Thus, the results reported here may underestimate the power of SNP-based markers.

The second assumption, purely additive contributions, may be violated in either direction: the combined effect of a pair of polymorphisms may be enhanced or diminished relative to their simple sum. An enhanced effect suggests that the haplotype bearing this set of polymorphisms will show greater association than expected under an additive model, which may indicate a preference for haplotype markers over SNP markers. Similarly, haplotypes bearing polymorphisms that combine for a diminished effect may indicate a preference for SNP markers.

A related epistatic effect is the presence of dominant or recessive alleles at a single SNP locus. Barring extreme cases of rare recessive alleles, it is possible to describe the phenotypic contribution in terms of an effective additive model. Similarly, when haplotypes are examined, the contribution from a single non-additive locus may be described in terms of an effective additive contribution and the conclusions reached here should still be valid. If there are multiple loci with dominant and recessive alleles, the relative

power of SNPs and haplotypes may depend on the LD between the different loci.

The third set of assumptions, small contributions and lack of selection, may be violated when a single SNP or haplotype is responsible for a major contribution to the phenotypic value. As discussed previously, large contributions totalling 5% or more of the overall phenotypic variance are readily detected by association in population sizes in the low hundreds and may even be amenable to traditional family-based linkage studies. One would expect that both SNP markers and haplotype markers would easily show significant association in these cases.

The assumed absence of selection in the model is consistent with studies of late-onset diseases and drug-response phenotypes and, for our purposes, is equivalent to assuming that LD is uncorrelated with SNP effects. It is unlikely that selection is completely absent and in early-onset diseases selection may be very strong. One outcome of strong selection is that a polymorphism making a large, deleterious phenotypic contribution may have low population frequency. Since the additive variance of a low-frequency marker is approximately proportional to the product of its frequency and the square of its contribution, higher-frequency polymorphisms making smaller contributions may be easier to detect. This conclusion applies to SNPs as well as haplotypes.

Consider, however, pairs of causative polymorphisms occurring in a haplotype carrying multiple unfavourable polymorphisms. Negative selection will tend to eliminate the extreme multiple-hit haplotypes from the population. These are the haplotypes required for greatest significance of haplotype-based tests and removing them reduces the power of the haplotype marker set. In contrast, the individual causative alleles may still have a considerable population frequency, much as heterozygote populations are maintained in diseases caused by a recessive allele at a single locus. When the haplotypes conferring an extreme phenotypic value are eliminated at a faster rate than the underlying SNPs, the causative SNPs may show greater significance than haplotypes and SNP-based tests will tend to be favoured over haplotype-based tests. To summarise, our base assumption is lack of selection and consequently no correlation between LD and SNP effects; when selection is present, it may create a form of LD that segregates deleterious mutations onto different haplotypes and favours SNPs over haplotypes for association tests.

A more complete examination of the points raised above might be required to reach more general conclusions. Epistatic effects, the presence of LD induced by selection pressure and the optimal marker set when the causative polymorphisms are not included in the set of known SNPs are three areas where a more complete theoretical treatment or simulation studies could provide additional guidance. We anticipate that the optimal marker set may obtained by identifying linkage groups containing a small number of SNPs, essentially sub-haplotypes or 'haplets', and using these as a set of polymorphic markers. This would be similar to the approach of Martin *et al.* except that the number of markers grouped for a multi-locus test would depend on estimates of the local extent of LD rather than on a global estimate.

If epistatic effects are large, models including SNP-SNP or haplotype-haplotype interaction terms may be necessary. The haplotype regression test, for example, may be considered an ANOVA test for haplotype main effects; including the haplotype-haplotype interaction terms, resulting in an analysis of covariance (ANCOVA) or multivariate analysis of variance (MANOVA) test applied to haplotypes, should be equivalent to the haplotype-pair ANOVA test used here. For the additive models we considered, the test of main effects is usually more powerful than the test including the interaction terms. It would be interesting to explore how large the interaction terms must be for their inclusion to yield a more powerful test.

Secondary assumptions that are less likely to affect the main conclusions can also be readily investigated. For example, in the examples shown in **Figures 1** and **2**, each causative SNP contributed equally to the total additive variance. In reality, a power-law-like distribution should govern the effect size: most loci make small contributions, while a few loci make large contributions. A significant finding of association depends on the most extreme additive variance in the SNP marker set rather than the mean variance among the causative markers. Similarly, a significant finding for the haplotype marker set depends on the haplotype with the most extreme additive variance rather than on the mean additive variance per haplotype, as was used here. Thus, it may be appropriate to include a more detailed description of the distributions of additive variance or the underlying distributions of phenotypic shifts, allele frequencies and haplotype frequencies. Nevertheless, since the

extreme value distribution has only a slow logarithmic divergence from the population mean, the error made in using the mean values is not likely to affect the conclusions.

Finally, although the subject of this report is individual genotyping, association tests using pooled DNA are also possible. A pooling study requires approximately 25% more individuals than an individual genotyping study [18], an increase in cost that may be insubstantial compared to the cost reductions arising from pooling. Current technologies for pooled DNA can measure allele frequencies but not haplotype frequencies. This factor in favour of SNP-based markers might be the most important consideration for many studies.

Acknowledgements

Joel Bader wishes to thank A Bansal and P Sham for helpful discussions and the reviewers for comments that improved the manuscript.

Bibliography

Papers of special note have been highlighted as:
- • of interest
- •• of considerable interest

1. OTT J: *Analysis of Human Genetic Linkage (Third Edition).* Johns Hopkins University Press, Baltimore, USA (1999).

2. RISCH N, MERIKANGAS K: **The future of genetic studies of complex human diseases.** *Science* (1996) **273**:1516-1517.
•• A primary reference for dense-map association studies.

3. KRUGLYAK L: **Prospects for whole-genome linkage disequilibrium mapping of common disease genes.** *Nature Genet.* (1999) **22**:139-144.

4. COLLINS A, LONJOU C, MORTON NE: **Genetic epidemiology of single-nucleotide polymorphisms.** *Proc. Natl. Acad. Sci. USA* (2000) **96**:15173-15177.

5. 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.

6. FULKER DW, CHERNY SS, SHAM PC, HEWITT JK: **Combined linkage and association analysis of quantitative traits.** *Am. J. Hum. Genet.* (1999) **64**:259-267.
•• Family-based tests.

7. ABECASIS GR, CARDON LR, COOKSON WOC: **A general test of association for quantitative traits in nuclear families.** *Am. J. Hum. Genet.* (2000) **66**:279-292.
•• An important extension of [6].

8. SHAM PC, CHERNY SS, PURCELL S, HEWITT JK: **Power of linkage *versus* association analysis of quantitative traits, by use of variance components models, for sibship data.** *Am. J. Hum. Genet.* (2000) **66**:1616-1630.

9. DANIELS JK, HOLMANS P, WILLIAMS NM *et al.*: **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.

10. 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.
• DNA pooling for dichotomous traits.

11. GULCHER J, STEFANSSON K: **Population genomics: laying the groundwork for genetic disease modeling and targeting.** *Clin. Chem. Lab. Med.* (1998) **36**:523-527.

12. SANDERS R: **Mining the Swedish clinical archives to develop pharmacogenomic tests.** *Mol. Diagn.* (1999) **4**:319-325.

13. RABINOW P: **French DNA, trouble in purgatory.** University of Chicago Press, Chicago, USA (1999).

14. JUDSON R, STEPHENS JC, WINDEMUTH A: **The predictive power of haplotypes in clinical response.** *Pharmacogenomics* (2000) **1**:15-26.

15. FALCONER DS, MACKAY TFC: **Introduction to quantitative genetics.** Prentice Hall, New York, USA (1996).

16. TERWILLIGER J, OTT J: **A haplotype based 'haplotype relative risk' approach to detecting allelic associations.** *Hum. Hered.* (1992) **42**:337-346.

17. SNEDECOR GW, COCHRAN WG: *Statistical Methods (Eighth Edition).* Iowa State University Press, Ames, USA (1989).

18. BADER JS, BANSAL A, SHAM PC: **Efficient tests of association for disease-risk phenotypes using pooled DNA.** (2001) (Manuscript Submitted).

19. MARTIN ER, LAI EH, GILBERT JR *et al.*: **SNPing away at complex disease: analysis of single-nucleotide polymorphisms around *APOE* in Alzheimer's disease.** *Am. J. Hum. Genet.* (2000) **67**:383-394.
• Direct fine-mapping of the *APOE* polymorphism.

20. PERICAK-VANCE MA, BEBOUT JL, GASKELL PC *et al.*: **Linkage studies in familial Alzheimer's disease: evidence for chromosome 19 linkage.** *Am. J. Hum. Genet.* (1991) **48**:1034-1050.

21. CORDER EH, SAUNDERS AM, STRITTMATTER WJ *et al.*: **Gene dose of apolipoprotein E Type 4 allele and the risk of Alzheimer's disease in late onset families.** *Science* (1993) **261**:921-923.

22. SAUNDERS AM, SCHMADER KE, BREITNER JC *et al.*: **Apoplipoprotein E ε4 allele distributions in late-onset Alzheimer's disease and in other amyloid-forming diseases.** *Lancet* (1993) **342**:710-711.

23. STRITTMATTER WJ, SAUNDERS AM, PERICAK-VANCE MA, SALVESEN GS, ENGLISH J, ROSES AD: **Apolipoprotein E: high avidity binding to a bA amyloid and increased frequency of Type 4 isoform in familiar Alzheimer's disease.** *Proc. Natl. Acad. Sci. USA* (1993) **90**:1977-1981.

24. CLAYTON D: **A generalization of the transmission/disequilibrium test for uncertain haplotype transmission.** *Am. J. Hum. Genet.* (1999) **65**:1170-1177.
• Haplotype-based tests.

25. *The Pharmacological Basis of Therapeutics, Ninth Edition.* Goodman LS, Hardman JG, Limberd LE, Molinoff PB, Ruddon RW, Gilman AG (Eds.), McGraw Hill, New York, USA (1996).

26. TAN S, HALL IP, DEWAR J, DOW E, LIPWORTH B: **Association between beta 2-adrenoceptor polymorphism and susceptibility to bronchodilator desensitisation in moderately severe stable asthmatics.** *Lancet* (1997) **350**:995-999.

27. TAYLOR DR, DRAZEN JM, HERBISON GP, YANDAVA CN, HANCOX RJ, TOWN GI: **Asthma exacerbations during long term beta agonist use: influence of beta(2) adrenoceptor polymorphism.** *Thorax* (2000) **55**:762-767.

28. CHONG LK, CHOWDRY J, GHAHRAMANI P, PEACHELL PT: **Influence of genetic polymorphisms in the beta2-adrenoceptor on desensitization in human lung mast cells.** *Pharmacogenetics* (2000) **10**:153-162.

29. LIGGETT SB: **The pharmacogenetics of beta2-adrenergic receptors: relevance to asthma.** *J. Allergy Clin. Immunol.* (2000) **105**:S487-S492.

30. MCGRAW DW, FORBES SL, KRAMER LA, LIGGETT SB: **Polymorphisms of the 5′ leader cistron of the human beta2-adrenergic receptor regulate receptor expression.** *J. Clin. Invest.* (1998) **102**:1927-1932.

31. DRYSDALE CM, MCGRAW DW, STACK CB *et al.*: **Complex promoter and coding region β_2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness.** *Proc. Natl. Acad. Sci. USA* (2000) **97**:10483-10488.
• A haplotype-based marker gave greater significance than the underlying SNPs.

32. LONG AD, LANGLEY CH: **The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits.** *Genome Res.* (1999) **9**:720-731.
• Simulation study of SNP-based and haplotype-based markers.

Joel S Bader
CuraGen Corp., 555 Long Wharf Drive, New Haven, CT 06511 USA
Tel.: +1 203 974 6236; Fax: +1 203 401 3351;
E-mail: jsbader@curagen.com

