Memorial Sloan-Kettering Cancer Center
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology
& Biostatistics Working Paper Series

Year 2007 *Paper* 7

Sequential Quantitative Trait Locus Mapping in Experimental Crosses

Jaya M. Satagopan* Saunak Sen†
Gary A. Churchill‡

*Memorial Sloan-Kettering Cancer Center, satagopj@mskcc.org
†University of California - San Francisco, sen@biostat.ucsf.edu
‡The Jackson Laboratory, garyc@jax.org

http://biostats.bepress.com/mskccbiostat/paper7

Sequential Quantitative Trait Locus Mapping in Experimental Crosses

Jaya M. Satagopan, Saunak Sen, and Gary A. Churchill

Abstract

The etiology of complex diseases is heterogeneous. The presence of risk alleles in one or more genetic loci affects the function of a variety of intermediate biological pathways, resulting in the overt expression of disease. Hence, there is an increasing focus on identifying the genetic basis of disease by sytematically studying phenotypic traits pertaining to the underlying biological functions. In this paper we focus on identifying genetic loci linked to quantitative phenotypic traits in experimental crosses. Such genetic mapping methods often use a one stage design by genotyping all the markers of interest on the available subjects. A genome scan based on single locus or multi-locus models is used to identify the putative loci. Since the number of quantitative trait loci (QTLs) is very likely to be small relative to the number of markers genotyped, a one-stage selective genotyping approach is commonly used to reduce the genotyping burden, whereby markers are genotyped solely on individuals with extreme trait values. This approach is powerful in the presence of a single quantitative trait locus (QTL) but may result in substantial loss of information in the presence of multiple QTLs. Here we investigate the efficiency of sequential two stage designs to identify QTLs in experimental populations. Our investigations for backcross and F2 crosses suggest that genotyping all the markers on 60% of the subjects in Stage 1 and genotyping the chromosomes significant at 20% level using additional subjects in Stage 2 and testing using all the subjects provides an efficient approach to identify the QTLs and utilizes only 70% of the genotyping burden relative to a one stage design, regardless of the heritability and genotyping density. Complex traits are a consequence of multiple QTLs conferring main effects as well as epistatic interactions. We propose a two-stage analytic approach where a single-locus genome scan is conducted in Stage 1 to identify promising chromosomes, and interactions are examined using the loci on these chromosomes in Stage 2. We examine settings

under which the two-stage analytic approach provides sufficient power to detect the putative QTLs.

Sequential Quantitative Trait Locus Mapping in Experimental Crosses

Jaya M. Satagopan
Memorial Sloan-Kettering Cancer Center
satagopj@mskcc.org

Saunak Sen
University of California - San Francisco
sen@biostat.uscf.edu

Gary A. Churchill
The Jackson Laboratory
garyc@jax.org

March 1, 2007

Abstract

The etiology of complex diseases is heterogeneous. The presence of risk alleles in one or more genetic loci affects the function of a variety of intermediate biological pathways, resulting in the overt expression of disease. Hence, there is an increasing focus on identifying the genetic basis of disease by sytematically studying phenotypic traits pertaining to the underlying biological functions. In this paper we focus on identifying genetic loci linked to quantitative phenotypic traits in experimental crosses. Such genetic mapping methods often use a one stage design by genotyping all the markers of interest on the available subjects. A genome scan based on single locus or multi-locus models is used to identify the putative loci. Since the number of quantitative trait loci (QTLs) is very likely to be small relative to the number of markers genotyped, a one-stage selective genotyping approach is commonly used to reduce the genotyping burden, whereby markers are genotyped solely on individuals with extreme trait values. This approach is powerful in the presence of a single quantitative trait locus (QTL) but may result in substantial loss of information in the presence of multiple QTLs. Here we investigate the efficiency of sequential two stage designs to identify QTLs in experimental populations. Our investigations for backcross and F2 crosses suggest that genotyping all the markers on 60% of the subjects in Stage 1 and genotyping the chromosomes significant at 20% level using additional subjects in Stage 2 and testing using all the subjects provides an efficient approach to identify the QTLs and utilizes only 70% of the genotyping burden relative to a one stage design, regardless of the heritability and genotyping density. Complex traits are a consequence of multiple QTLs conferring main effects as well as epistatic interactions. We propose a two-stage analytic approach where a single-locus genome scan is conducted in Stage 1 to identify promising chromosomes, and interactions are examined using the loci on these chromosomes in Stage 2. We examine settings under which the two-stage analytic approach provides sufficient power to detect the putative QTLs.

Introduction

Complex diseases have a heterogeneous etiology. Risk alleles present in a genetic locus or their simultaneous presence in multiple loci may affect a variety of intermediate biological functions. Changes in any one or more intermediate functions results in the overt expression of disease. Modern genetic studies are, therefore, increasingly focusing on systematically evaluating specific disease pathways in an effort to identify disease risk factors in an efficient manner (Thomas 2005). Quantitative (or qualitative) measurements corresponding to a specific phenotype underlying the disease are obtained on all the study subjects. The genetic loci linked to the phenotypic traits are identified to determine those conferring disease risk through specific biological mechanisms. Suppose the disease of interest is heart disease. Some examples of intermediate phenotypic traits are cholesterol level, blood pressure, body mass index, and urinary free cortisol. A heart disease patient is very likely to have abnormal levels of at least one of these phenotypic traits. Genetic loci linked to these traits may provide insights into biological mechanisms underlying the etiology of heart disease. Due to recent developments in biotechnology, it is now becoming increasingly feasible to simultaneously examine a large number of phenotypes at the molecular level (for example, gene expressions). In this paper we consider identifying genetic loci related to a single quantitative phenotypic trait in experimental crosses. We focus on sequential methods to identify the putative trait loci (or, equivalently, disease loci) in an efficient manner.

Quantiative trait studies in experimental crosses proceed as follows. A desired cross (for example, backcross or F2) is obtained using two parental strains. Genotypes at several loci and the phenotypic trait are measured on multiple progeny from this cross. The quantitative trait loci (QTLs) linked to the trait are identified using a relevant analytic approach such as interval mapping (Lander and Botstein 1989). A one-stage genotyping approach is often utilized, whereby an ensemble of loci (markers) is genotyped on all the n available subjects (progeny). The genome is then scanned for the presence of QTLs by fitting a single QTL model at various loci. The genotyping cost (or genotyping burden) of this strategy is proportional to genotyping all the markers on the n subjects. Often the number of QTLs may be small relative to the number of markers genotyped. It may, therefore, be pragmatic to consider a strategy that would require fewer genotyping to identify the QTLs with adequate power.

Selective genotyping is a one-stage approach commonly used to minimize the genotyping burden (Lander and Botstein 1989; Darvasi and Soller 1992, 1994). This is an outcome-based sampling approach where all the markers are genotyped on subjects with extreme trait values alone. When a single QTL is associated with the phenotypic trait, genotyping only a quarter

of individuals from each extreme (i.e., one half of the n subjects) provides most of the linkage information compared to genotyping all the n subjects. However, selective genotyping may not be an efficient strategy when the variation in the trait is explained by multiple QTLs, at least one of which has a large effect (Sen et al. 2005). In such cases, the fraction of missing information can be as high as 50% if only one-half of the subjects having extreme trait values are genotyped. It is, therefore, important to devise a genotyping approach, whereby all the subjects (i.e., those with extreme as well as intermediate trait values) are genotyped, but without having to genotype every marker on every progeny, unless warranted otherwise.

Here we propose a two-stage sequential genotyping approach that requires fewer genotyping than a one-stage approach. Under the proposed method, several markers not related to the trait can be eliminated early on in the study by evaluating all the markers on only a subset of the available progeny. Only those markers showing promising evidence for association with the trait are further genotyped on the remaining progeny to identify the QTLs. Such cost-efficient two stage genotyping designs have been proposed to identify disease loci in human linkage analyses and population-based association studies (Elston 1994; Elston et al. 1996, 2007; Satagopan et al. 2002, 2004; Satagopan and Elston 2003; Wang et al. 2006). In this paper we develop this method for QTL mapping in experimental crosses. Any sequential approach can result in loss of power when the total sample size is fixed since a marker linked to the trait may be incorrectly eliminated in the first stage. This issue can be addressed by genotyping an appropriate subset of individuals in the first stage and by setting the corresponding significance level to identify the promising markers for further evaluation in the second stage. We show that the expected Fisher information of the two-stage design relative to a one-stage design has a simple form that can be used to identify the sample size and significance level for the first stage such that the loss of power and the genotyping burden relative to a one-stage design are minimized. Epistasis or interactions between multiple QTLs form an important characteristic of complex phenotypic traits. Conducting a genome scan using single QTL models may not be a powerful approach to identify all the QTLs (Broman and Speed 1999). The genome must be scanned using multi-locus models to simultaneously identify relevant QTLs, which can quickly become an arduous task. Two-stage analysis may be useful for circumventing this issue (Marchini et al. 2005). Here we examine a two-stage analytic approach where a genome scan using single locus models is conducted in the first stage to identify promising chromosomal regions. Interactions between loci in these regions alone are evaluated in the second stage using multi-locus models.

The goals of this paper are to investigate the power trade-offs between one-stage versus two-stage methods and to evaluate the optimal strategies. In the next section we describe the

characteristics of a one stage design by providing the overall significance level, power, and the genotyping burden of this approach. Next we develop the two stage design, derive simple equations to obtain the optimal design parameters, and evaluate the operating characteristics. These methods are first developed for a backcross, and subsequently described for an F2 intercross, assuming a single locus model. This is followed by an evaluation of two stage designs for a backcross when two unlinked QTLs are associated with the trait. Finally, we examine a two-stage analytic approach to identify multiple QTLs conferring main effects and epistatic effects on the trait.

One Stage Design

Consider a genome of interest with C chromosomes, and K_c markers on chromosome c ($1 \leq c \leq C$). The total number of markers is $K = \sum_{c=1}^{C} K_c$. Let L denote the length of the genome, and Δ represent the average genotyping density (i.e., distance between the markers), both in centiMorgan units. Under a one stage design all the K markers are genotyped on all the n available subjects. We first outline the general concept in the context of a backcross population using a single locus model, assuming that a single QTL is associated with the trait.

Let g_i and y_i denote the genotype at the QTL and the phenotypic trait, respectively, in a backcross subject i ($1 = 1, \cdots, n$). Without loss of generality, the genotype at a locus is coded 0 or 1, each having marginal probability 1/2. The single locus model is given by:

$$y_i \;\; = \;\; \delta(2g_i - 1) + \epsilon_i \; , \tag{1}$$

where δ is the effect of the QTL. Further, $\epsilon_i \sim N(0, \sigma^2)$ is the random error, where σ^2 is the error variance. We shall assume $\sigma^2 = 1$. While σ^2 is estimated in a practical data analysis setting, the assumption of unit error variance during study design is simply a matter of scaling the QTL effect. Hence, δ/σ is interpreted as the standardized QTL effect or the effect size. In a design setting, the investigator strives to determine the power or sample size to detect a desired effect δ under some assumed σ^2. This is equivalent to designing a study to detect a desired effect size δ/σ. Hence, without loss of generality, we consider $\sigma^2 = 1$ throughout this paper.

The phenotypic trait is marginally distributed as Gaussian with mean 0 and variance $\tau^2 = 1 + \delta^2$. We assume that the parameter estimates and test statistics are calculated based on the interval mapping approach (Lander and Botstein 1989). In practice we observe the marker genotypes but not the QTL genotypes. Suppose we investigate evidence for the presence of a QTL at a locus flanked by two markers such that the recombination between the locus and

the left flanking marker is r_1. Let r be the recombination fraction between the flanking markers. Both r and r_1 can be calculated as a function of genetic distance using the Haldane mapping function (Ott 1991). Denote $\phi(.)$ as the probability density of a standard normal distribution. Suppose we are testing for the presence of a QTL at a particular genetic locus. Let q_i denote the conditional probability that the QTL genotype at that locus is 1, given the flanking marker genotypes. The values of r_1, r and, hence, q_i can be easily obtained and treated as known during the analysis of that locus. The log-likelihood corresponding to model (1), evaluated at the locus under investigation, is given by:

$$l(\delta; r_1) \;=\; \sum_{i=1}^{n} \log\{q_i\ \phi(y_i - \delta) + (1 - q_i)\phi(y_i + \delta)\}\ . \tag{2}$$

The score statistic, denoted Z, for testing the null hypothesis of no QTL ($\delta = 0$) is given by:

$$Z \;=\; \frac{l^{'}(\delta = 0; r_1)}{\sqrt{I(\delta = 0; r_1, n)}}\ , \tag{3}$$

where $l^{'}(\delta = 0; r_1)$ is the derivative of the log-likelihood with respect to δ, evaluated at $\delta = 0$. $I(\delta = 0; r_1, n)$ is the Fisher's information corresponding to δ, also evaluated at $\delta = 0$ using a sample of n individuals. Following the general theory of likelihoods and score statistics (Cox and Hinkley 1974), Z is distributed as $N(\delta\sqrt{I(\delta = 0; r_1, n)}, 1)$. [The mean of Z is outlined in the Appendix.] Hence, Z^2 has a χ^2 distribution with 1 degree of freedom and non-centrality parameter

$$\lambda(\delta, n) \;=\; I(\delta = 0; r_1, n)\delta^2\ . \tag{4}$$

Clearly, the non-centrality parameter is 0 under the null hypothesis. The information depends upon the distance between the flanking markers and the location of the QTL within the marker interval, and information is the least when the QTL is located in the middle of the marker interval (Sen et al. 2005). Therefore, thoughout this paper we consider designing a QTL study under the assumption that the QTL is located in the center of the marker interval. The Fisher's information is the expected value of the negative second derivative of equation (2) with respect to δ, given by $I(\delta = 0; r_1, n) = nQ_r$, where $Q_r = [1 - 4q(1-q)](1-r)$ and $q(1-q) = r_1^2(1-r_1)^2/\{r_1^2 + (1-r_1)^2\}^2$ [see Sen et al. 2005]. Here r and r_1 are the recombinations corresponding to Δ and $\Delta/2$ centiMorgan distances, respectively. Below we describe the overall significance level, power, and the associated genotyping burden.

Overall Significance Level: We calculate test statistics $Z^2(t)$ at every locus t on the genome. The overall significance level α is the probability that the genome-wide maximum test statistic,

$\max_t Z^2(t)$, exceeds a threshold $b^2(\alpha)$ under the null hypothesis. Denote $P_0(.)$ as the probability of an event under the null hypothesis of no QTL. The significance level can be approximated as (Feingold et al. 1993; Dupuis and Siegmund 1999):

$$\begin{aligned} \alpha &= P_0(\max_t Z^2(t) > b^2(\alpha)) \\ &= 1 - \exp\left\{-C[1 - \mathcal{X}_1^2(b^2)] - 0.04\ L\ b\ \phi(b)\ v(b\sqrt{0.04\Delta})\right\}\ , \end{aligned} \tag{5}$$

where $\mathcal{X}_s^2(x)$ is the cumulative probability of a χ^2 distribution with s degrees of freedom corresponding to critical value x, and $v(u) = \exp\{-0.583u\}$. We have denoted $b^2(\alpha)$ simply as b^2 in second step of the above equation. Although this approximation was derived based on the likelihood ratio statistic, it is applicable to the current setting where $Z(t)$ is the score statistic due to the asymptotic equivalence between $Z^2(t)$ and the likelihood ratio statistic evaluated at locus t (Cox and Hinkley 1974).

Power: Suppose our goal is to identify the QTL having effect δ on the trait with power $1 - \beta$ using a single locus model (equation 1). Consider the following two events. Event 1 corresponds to $\max_t Z^2(t)$ exceeding b^2 in the presence of a QTL, and Event 2 corresponds to test statistic Z^2 at the QTL position exceeding b^2. Let $P_A(.)$ denote the probability of an event under the alternative hypothesis. Clearly, $P_A(\text{Event 2}) \leq P_A(\text{Event 1})$. Our working definition of power is the probability of Event 2 under the alternative hypothesis. Therefore, the power $1 - \beta$ can be written as

$$\begin{aligned} 1 - \beta &= P_A(Z^2 > b^2(\alpha)) \\ &\approx 1 - \Phi[b(\alpha) - \sqrt{\lambda(\delta, n)}]\ , \end{aligned} \tag{6}$$

where $\Phi(.)$ is the cumulative probability of a standard normal distribution.

The sample size n, significance level α and power $1 - \beta$ are related through the following equation:

$$n = \frac{1}{\delta^2 \times [1 - 4q(1-q)] \times (1-r)} \times \left\{b(\alpha) - \Phi^{-1}(\beta)\right\}^2\ . \tag{7}$$

Genotyping Burden: Under a one stage design a total of K markers are genotyped on n subjects. Therefore, the genotyping burden or the amount of genotyping is:

$$T_1 = n \times K\ . \tag{8}$$

Two Stage Sequential Genotyping Design

First obtain the sample size n to detect a desired QTL effect with power $1-\beta$ and significance level α using equation (7) as if we were conducting a one-stage design. This will be our given sample size. Obtain the trait values of these n subjects. Our goal is to detect the QTL by genotyping these n subjects in an efficient manner without having to genotype all the markers on all the subjects. The two stage sequential genotyping design proceeds as follows. Given the trait values of these n subjects, genotype all the desired markers on random subset of $n\theta_1$ $(0 < \theta_1 < 1)$ individuals in Stage 1. Identify chromosomes significant at level α_1. The markers on these chromosomes are genotyped on the remaining $n(1-\theta_1)$ subjects in Stage 2 and tested using all the n individuals at level α_2. The sampling fraction θ_1 and the significance levels α_1 and α_2 are unknown, and form the two stage design parameters. This two stage approach involves fewer genotyping than a one stage design. Hence, if the cost of genotyping were linearly related to the amount of genotyping, this design would provide a cost-effective genotyping approach. However, any sequential approach would result in loss of power relative to a one stage approach. For an appropriate choice of θ_1, α_1, and α_2, it may be feasible to design a study involving substantially fewer genotyping but minimum loss of power, while ensuring an overall significance level of α. Below we investigate the operating characteristics of two stage designs to identify such θ_1, α_1 and α_2. Our investigations focus on addressing the following two questions: (1) Given the phenotypic traits of n individuals, what subgroup size should be used for genotyping in Stage 1? (2) How should the chromosomes be prioritized for further evaluation in the next stage so that the putative QTLs can be identified efficiently with minimum genotyping burden?

Overall Significance Level of the Two Stage Design

The overall significance level α is fixed at the outset, and is the probability of finding a false positive QTL i.e., the probability that the genome-wide maximum test statistic exceeds a significance threshold at the end of Stage 2 under the null hypothesis. Since the C chromosomes are unlinked, we can apply the Bonferroni correction and test each chromosome at level α/C. The significance level α_1 for testing a single chromosome in Stage 1 represents the probability that the maximum test statistic on a chromosome exceeds a critical value b_1^2 under the null hypothesis. The significance level in Stage 2, denoted α_2, is the conditional probability under the null hypothesis that the maximum test statistic on chromosome c exceeds a critical value b_2^2 at the end of Stage 2, given that the chromosome was declared significant in Stage 1. The following results are derived in the Appendix.

Result 1 *For any chromosome, the significance level for testing in Stage 1 is greater than the overall significance level:* $\alpha_1 > \alpha/C$.

Result 2 *The Stage 2 significance level is given by:* $\alpha_2 = \alpha/(C \times \alpha_1)$.

Result 3 *When the critical value for Stage 2 is equal to the critical value of a one stage design, i.e., when* $b_2^2 = b^2$*, the overall significance level of the two-stage design is at most* α.

Result 1 is consistent with intuition that the study may not have sufficient power to detect a QTL at significance level α/C when genotyping a fraction of $n\theta_1(< n)$ subjects. Hence, the significance testing must be conducted at level $\alpha_1 > \alpha/C$ at Stage 1 in order to have sufficient power to identify the chromosomes containing QTLs. Result 2 suggests that α_2 is determined once α_1 is known. Consequently, the only unknown parameters defining a two stage design are θ_1 (the sampling fraction for Stage 1), and α_1 (the per-chromosome significance level in Stage 1). Result 3 suggests that it is not necessary to derive a new critical value for testing the chromosomes in Stage 2 corresponding to significance level α_2. After the completion of Stage 1, the chromosomes genotyped in Stage 2 can be evaluated using the same critical value as that under one stage design in order to obtain a per-chromosome overall significance level of at most α/C.

Power of the Two Stage Design

The power P^* of the proposed two stage design is the probability that the test statistic at the QTL exceeds b^2 at the end of Stage 2. Let Z_1 and Z_2 denote the score statistics calculated at the QTL position in Stages 1 and 2 using $n\theta_1$ and n individuals, respectively. Further, $Z_1 \sim N(\sqrt{n\theta_1 Q_r}\delta, 1)$ and $Z_2 \sim N(\sqrt{nQ_r}\delta, 1)$. These score statistics are equivalent to the test statistics $S_1 = \sqrt{n\theta_1 Q_r} Z_1$ and $S_2 = \sqrt{nQ_r} Z_2$.

In a one-stage design we would observe the test statistic S_2 (equivalently, Z_2) at the QTL position. Under a two-stage design, the QTL region will be evaluated in Stage 2 only if the corresponding chromosome is selected in Stage 1. The test statistic at the QTL position obtained in Stage 2 is S_2. If the chromosome is not selected in Stage 1, we will only calculate the test statistic S_1. A chromosome is evaluated in Stage 2 if the maximum score statistic (or the maximum chi-squared statistic) on this chromosome exceeds b_1 (or b_1^2) in Stage 1. The power of Stage 1 is $1 - \beta_1$. Therefore, the chromosome harboring the QTL will be evaluated in Stage 2 with probability (at least) $1 - \beta_1$. With probability β_1, a test statistic at the QTL will be evaluated in Stage 1 but not in Stage 2. Therefore, the test statistic S at the QTL position obtained at the

end of the two-stage procedure is either S_1 or S_2, depending upon whether the QTL position has been evaluated in Stage 1 alone or in Stage 2 as well. Therefore, S is a mixture over S_1 and S_2, given by test statistic at the QTL position obtained under the two-stage procedure is $S = S_1\mathcal{I}(\max_t Z_t < b_1\sqrt{nQ_r\theta_1}) + S_2\mathcal{I}(\max_t Z_t > b_1\sqrt{nQ_r})$, where $\mathcal{I}(.)$ is an indicator function. The power of the two-stage procedure is $P^* = P_A(S > b\sqrt{nQ_r})$, and requires evaluating a double integral.

The expected value of S is $E(S) = \beta_1 E(S_1) + (1-\beta_1)E(S_2)$, which clearly depends upon $1-\beta_1$ and θ_1. The non-centrality parameter of S, denoted $\lambda_2(\delta, n)$, is the difference between the expected values of S evaluated under the alternative and null hypotheses. Since the expected value of S under the null hypothesis is 0, the non-centrality parameter is $E(S) = \lambda_2(\delta, n)$, written as:

$$\lambda_2(\delta, n) \;=\; \beta_1 \frac{\lambda(\delta, n\theta_1)}{\delta} + (1-\beta_1)\frac{\lambda(\delta, n)}{\delta}\;, \tag{9}$$

where $\lambda(.)$ is the non-centrality parameter given by equation (4). Note that $\lambda(\delta, n\theta_1)/\delta$ and $\lambda(\delta, n)/\delta$ are the expected values of S_1 and S_2, respectively. Since the n individuals are independent, the test statistic S_2 can be written as the sum of independent contributions from $n\theta_1$ individuals genotyped in Stage 1 and $n(1-\theta_1)$ individuals newly genotyped in Stage 2. Hence, $S_2 = S_1 + X$, where S_1 and X are independent. Further, X has a normal distribution with mean $n(1-\theta)Q_r\delta$ and variance $n(1-\theta)Q_r$. The variance of S is $Var(S) = n\theta_1 Q_r + n(1-\theta_1)Q_r(1-\beta_1)^2$. Suppose we fix θ_1. Intuitively $1-\beta_1 \to 1$ as $\alpha_1 \to 1$. The limiting case as $\alpha_1 \to 1$ will be a one-stage design since increasing number of chromosomes will be genotyped in Stage 2. For a fixed θ_1, $E(S) \to E(S_2)$ and $Var(S) \to V(S_2)$ as $1-\beta_1 \to 1$. These imply weak convergence of S to S_2 when $1-\beta_1 \to 1$ for a given θ_1. Therefore, identifying a two-stage design with non-centrality parameter close to that of a one-stage design i.e., $\lambda_2(\delta, n)$ close to $\lambda(\delta, n)$ would provide a design having power close to that of a one-stage design.

The relative non-centrality parameter or, equivalently, the relative information is defined as the ratio of the non-centrality parameters of the test statistics at the QTL locus under the two-stage and one-stage designs:

$$\begin{aligned} \frac{\lambda_2(\delta, n)}{\{\lambda(\delta, n)/\delta\}} &= \beta_1\theta_1 + 1 - \beta_1 \\ &= 1 - \beta_1(1-\theta_1)\;. \end{aligned} \tag{10}$$

The right hand side does not involve δ. Since $\beta_1 \le 1$ and $\theta_1 \le 1$, the relative information is ≤ 1. Equivalently, $P^* < 1-\beta$, where $1-\beta = P_A(S_2/\sqrt{n} > b)$ is the power of a one-stage design. Were we to identify θ_1 and α_1 (and, hence, $1-\beta_1$) such that $\beta_1(1-\theta_1)$ is small, then the relative information will be close to 1.

The Stage 1 sampling fraction, θ_1, can be written using equation (7) as:

$$\theta_1 \;=\; \frac{(b_1 - \Phi^{-1}(\beta_1))^2}{(b - \Phi^{-1}(\beta))^2} \; . \tag{11}$$

The overall significance level α (and, hence, b^2), and the overall power $1-\beta$ are specified at the outset, and b_1 depends upon α_1. Therefore, θ_1 depends upon α_1 and $1-\beta_1$. Further, given $1-\beta_1$, θ_1 does not depend upon the QTL effect. Finally, the relative information (equation 10) depends upon θ_1 and β_1.

Genotyping Burden

The total number of genotyping in a two stage design, denoted T_2, comprises of genotyping all the $K = \sum_{c=1}^{C} K_c$ markers in $n\theta_1$ subjects in Stage 1, and genotyping the markers from the promising chromosomes on the remaining subjects in Stage 2. Note that every null chromosome has probability α_1 of being evaluated in Stage 2. Further, every chromosome containing a QTL will be evaluated in Stage 2 with probability at least $1-\beta_1$. Let D denote the total number of chromosomes carrying a QTL. When there is a single QTL, D = 1. Therefore, T_2 is given by:

$$T_2 \;=\; n\theta_1 \sum_{c=1}^{C} K_c + n(1-\theta_1) \left\{ (1-\beta_1) \sum_{c=1}^{D} K_c + \alpha_1 \sum_{j=1}^{C-D} K_j \right\} \; . \tag{12}$$

The relative genotyping burden, T_2/T_1, is the ratio of the genotyping burdens of the two stage and one stage designs:

$$\frac{T_2}{T_1} \;=\; \theta_1 + (1-\theta_1) \left\{ (1-\beta_1) \frac{\sum_{c=1}^{D} K_c}{\sum_{j=1}^{C} K_j} + \alpha_1 \frac{\sum_{j=1}^{C-D} K_j}{\sum_{j=1}^{C} K_j} \right\} \; , \tag{13}$$

with θ_1 given by equation (11). Therefore, T_2/T_1 is independent of the model parameters for given α_1 and $1-\beta_1$. Clearly $T_2 \leq T_1$ i.e., fewer genotyping is undertaken in a two stage design relative to a one stage design.

Optimal Two Stage Design

The two stage design involves fewer genotyping than a one stage design, but incurs a loss of power. Therefore, our goal is to identify a two stage design so that the loss of power, $1-\beta-P^*$, is minimized. Equivalently, we shall identify a two-stage design such that the relative information (equation 10) is close to 1. In doing so, we must ensure that the overall significance level is maintained at a desired level α. The number of chromosomes C, the genome length L, the

desired effect δ, the average genotyping density Δ, the desired overall significance level α, and power $1-\beta$ are specified at the design stage. First obtain the sample size n to detect some desired QTL effect δ using equation (7) as if we were conducting a one-stage design. This will be our fixed sample size. The phenotypic traits will be measured on all these n subjects. Therefore, given the trait values of these n subjects, our goal is to identify α_1 and $1-\beta_1$ (and, hence, θ_1) so that (i) the overall significance level is α, (ii) $1-\lambda_2(\delta,n)/\lambda(\delta,n) < \epsilon$ for some small ϵ, and (iii) T_2/T_1 is minimum. Here ϵ is a user-specified quanitity indicating the acceptable amount of relative information loss. For example, setting $\epsilon = 0.05$ would indicate that the desired relative information is at least 95%. Following Results 1, 2, and 3, the overall significance level will be maintained at level α by choosing $\alpha_1 \in (\alpha/C, 1)$ and setting the Stage 2 critical value equal to that of a one stage design. The optimal parameters can be obtained through the following steps.

1. Fix a relative genotyping burden T_2/T_1.
2. For a value of $\alpha_1 \in (\alpha/C, 1)$, obtain the critical value b_1 by setting $C = 1$ and using L/C in place of L in equation (5).
3. Obtain $1-\beta_1$ from equation (13) using the Newton-Raphson method. In order to perform this calculation, θ_1 given by equation (11) must be substituted into equation (13).
4. Obtain the sampling fraction θ_1 from equation (11), and the relative information using equation (10).
5. Repeat the above steps for various values of $\alpha_1 \in (\alpha/C, 1)$, and find α_1 minimizing $1-\lambda_2(\delta,n)/\lambda(\delta,n)$.
6. Repeat this procedure for various values of T_2/T_1, and identify the smallest T_2/T_1 for which $1-\lambda_2(\delta,n)/\lambda(\delta,n) < \epsilon$.

Since equations (10) and (13) are independent of the QTL effect δ and equation (11) is independent of δ once β_1 is given, the optimal solution is also indepdent of δ, once n is given.

We assess the operating characteristics of the two stage design under various parametric configurations to identify the optimal design. The overall significance level and power of a one stage design are $\alpha = 0.05$, and $1-\beta$ = 0.80, unless indicated otherwise. Figure 1 illustrates the optimal two stage design as a function of the relative genotyping burden for a hypothetical genome with C = 20 chromosomes, each of length 100 centiMorgans, and marker density Δ = 1 centiMorgan. It is evident that the relative information increases as the relative genotyping burden increases. The relative information is $\geq$ 95% ($\epsilon \leq$ 0.05) when the relative genotyping

burden $T_2/T_1 \geq$ 70%. Equivalently, a two stage design that utilizes only 70% genotyping burden results in less than 5% loss of information relative to a one stage design. Likewise, less than 10% information is lost by conducting a two stage design with 60% relative genotyping burden. This result holds regardless of the value of Δ (figures for other values of Δ are similar and, hence, not shown).

Table 1 provides the optimal two stage design parameters under various choices of Δ, and T_2/T_1 = 0.60 and 0.70. When T_2/T_1 = 70%, the optimal parameters are approximately α_1 = 0.21, $1-\beta_1$ = 0.90, and θ_1 = 0.60, and the maximum relative information is 0.96 (i.e., $\epsilon < 0.05$), regardless of the value of Δ. When the relative genotyping burden is 60%, the optimal two stage design parameters are approximately α_1 = 0.16, $1-\beta_1$ = 0.80, and θ_1 = 0.50, and the maximum relative information is 0.90 (i.e., $\epsilon \approx 0.10$) for all choices of Δ.

Figure 2 illustrates the behavior of the two stage design parameters when the relative genotyping burden is fixed at 70%. Choosing α_1 between 10% and 35% provides a relative information of at least 95%. The quadratic pattern for the relative information and $1-\beta_1$ can be explained as follows. Small values for α_1 ($< 10\%$) imply that few chromosomes will be declared significant at the end of Stage 1 for further evaluation in Stage 2. While this permits us to genotype more individuals in Stage 1 (θ_1 between 65% and 70%), small α_1 also implies small Stage 1 power $1-\beta_1$. Consequently, the relative information is small. When α_1 is large (for example, $\alpha_1 > 50\%$), this implies that more chromosomes will be declared significant in Stage 1 and, hence, genotyped in Stage 2. Therefore, when the genotyping burden is fixed, the fraction of individuals to be genotyped in Stage 1 is reduced, thus compromising the Stage 1 power to detect the chromosomes harboring the QTL. This, in turn, results in small relative information. Again, this result holds regardless of the value of Δ.

General Guideline: These results indicate that, as a general guideline, genotyping all the markers on θ_1 = 60% of the individuals in Stage 1, and genotyping the markers on chromosomes significant at α_1 = 20% level using the remaining individuals in Stage 2 results in minimal loss of information (ϵ between 0.05 and 0.10) and requires nearly 30% fewer genotyping than a one stage design. In particular, first obtain sample size n as if a one-stage design has been planned. Treating this as the fixed sample size, now conduct the study using a two-stage design with the above guidelines. Similar guidelines were obtained for association studies in humans (Satagopan et al. 2004). This general guideline is applicable to genomes of any size, and can be used to identify a QTL via single locus models.

Before proceeding further, it will be useful to understand why Δ does not impact the optimal

parameters. The critical value b and the sample size n, obtained at the outset using equation (7) to detect a desired QTL effect δ, will depend upon Δ. Suppose the desired QTL effect is δ = 0.20, and the overall power and significance level are $1-\beta$ = 0.80 and α = 0.05, respectively. The critical values under Δ = 20 and 1 are b = 3.90 and 3.95, respectively, yielding corresponding sample sizes of n = 700 and 580. Suppose we desire a relative genotyping burden of 0.60 under a two-stage design. When θ_1 = 0.50, the sample sizes are 350 and 290, respectively. When α_1 = 0.16, the critical values to test a single chromosome with Δ = 20 and 1 are b_1 = 2.51 and 2.56, respectively. The power of Stage 1 is $1-\beta_1$ = 0.80, relative information is approximately 0.90, and the relative genotyping burden is 0.60 under both the values of Δ. Hence, we have the same Stage 1 power and relative information under both the values of Δ precisely because the underlying sample sizes are different. However, the choice of optimal sampling fraction and significance level for Stage 1 do not depend upon the QTL effect and Δ once n is obtained corresponding to a given Δ and is treated as the fixed available sample size.

F2 Crosses

We now investigate optimal two stage genotyping for F_2 crosses having one of three possible genotypes at each locus (homozygous corresponding to one of the two parental types, or a heterozygote). Dominant and additive effects of QTLs can be examined using F2 crosses. At any locus, the dominant effect, denoted δ_d, is the difference between the trait value of the heterozygotes and the average trait value of the homozygotes. The additive effect, denoted δ_a, is the averge difference between the trait values of the homozygotes. Without loss of generality, the marker and QTL genotypes are coded as -1 (homozygous for one parental type), 0 (heterozygote), and 1 (homozygous for the other parental type), and have respective marginal probabilities 1/4, 1/2, and 1/4. Denoting $\mathcal{I}(.)$ as an indicator function, the phenotypic trait of individual i given the QTL genotype g_i is modeled as:

$$y_i \;=\; \left(-\delta_a - \frac{\delta_d}{2}\right)\mathcal{I}(g_i=-1) + \frac{\delta_d}{2}\mathcal{I}(g_i=0) + \left(\delta_a - \frac{\delta_d}{2}\right)\mathcal{I}(g_i=1) + \epsilon_i\ .$$

A single QTL model can be fit using the interval mapping approach, and a score statistic $Z(t)$ can be obtained at every locus t to test the null hypothesis $H_0 : \delta_a = 0 = \delta_d$ against the alternative $H_A : \delta_a \neq 0$ or $\delta_d \neq 0$. $Z(t)^2$ has a χ^2 distribution with 2 degrees of freedom. The non-centrality parameter under the alternative hypothesis is given in the Appendix. The overall significance level of a one stage design for an F2 cross can be approximated as (Dupuis and

Siegmund 1999):

$$\begin{aligned} \alpha &= P_0(\max_t Z_t^2 > b^2) \\ &= 1 - \exp\left\{ -\left(C + 0.03\, b^2\, L\, \nu(b\,\sqrt{0.06\Delta}) \right) \times \exp\left(-b^2/2\right) \right\} . \end{aligned} \tag{14}$$

The power $1-\beta$ of a one stage design is $P_A(Z > b^2)$, where Z is the test statistic calculated at the QTL position.

Under a two stage design, for $\alpha_1 \in (\alpha/C, 1)$, the critical value b_1^2 can be obtained from equation (14) by setting $C = 1$ and using L/C in place of L. The relative information and the relative genotyping burden have the same form as equations (10) and (13), respectively. The optimal two-stage design parameters can be obtained using the algorithm outlined for a backcross. The following result holds for small δ_a and δ_d when the sample size n is fixed (see Appendix for proof).

Result 4 *When n is fixed and δ_a and δ_d are small, θ_1 is independent of the QTL effects and phenotypic variance.*

Consider an experiment with a sample size of 320 F2 individuals segregating a single QTL with additive and dominance effects δ_a = 0.33 = δ_d (heritability = 5%), and the hypothetical genome consisting of C = 20 chromosomes, each of length 100 centiMorgans, with Δ = 1 centiMorgan. For this configuration, the power to detect the QTL is approximately $1-\beta$ = 80% at an overall significance level of α = 0.05 under a single locus model. Figure 3 illustrates the operating characteristics of the two stage design. The relative information is 95% (i.e., ϵ = 0.05) and the relative genotyping burden is 70% when θ_1 = 60% and α_1 = 0.20. Relative information of 95% is also attained when (θ_1, α_1) = (0.65, 0.12). This suggests that, while the optimal two stage design parameters can be obtained, the solution may not be unique. However, as a general guideline, genotyping all the markers on θ_1 = 60% of the individuals in Stage 1 and genotyping the chromosomes significant at level α_1 = 0.20 on the remaining subjects in Stage 2 provides 95% relative information with 70% relative genotyping burden. The operating characteristics of this general guideline is illustrated in Table 2 under a variety of parametric configurations. It is evident that the general guideline provides an optimal QTL mapping strategy, regardless of the QTL effects and the genotyping density. That the optimal design does not depend upon the QTL effects is consistent with Result 4. The relative information is at least 95% ($\epsilon < 0.05$) for a genome with C = 20 chromosomes. For a genome with C = 10 chromosomes, the relative information is between 93% and 95%. This general guideline is consistent with the optimal design recommendation identified for a backcross based on a single locus model.

Two QTL Models

The methods described so far consider a single locus model at various genomic locations to identify a QTL associated with the trait. Complex traits are very likely influenced by multiple QTLs. Suppose two QTLs are associated with the phenotypic trait. Let P_j denote the power to identify QTL j ($j = 1, 2$) using a single locus model. The power to detect both the QTLs is $P_1 \times P_2$. If $P_1 = 0.80 = P_2$, then the genome scan based on single locus model has only 64% power to detect both the loci. Genome scans using multi-locus models can provide a more powerful approach to identify the putative loci. Here we consider the simple case where D = 2 QTLs are associated with the phenotypic trait of a backcross individual, and examine the operating characteristics of a two stage design. Under a one-stage design, two-locus models are fit by considering every pair of loci on the genome. The null hypothesis of no QTL is tested against the alternative hypothesis of two QTLs. The two-stage design proceeds as follows. In Stage 1, all the markers are genotyped on a random subset of $n\theta_1$ individuals. A genome scan is conducted using two-locus models. On every chromosome c, we calculate test statistics corresponding to two-locus models where one locus is from chromosome c and the second locus is either from c or from a different chromosome. The maximum test statistic on chromosome c is the maximum of the test statistics from two-locus models where at least one locus is on c. Each chromosome-specific maximum test statistic is tested at significance level α_1. The markers on the significant chromosomes are genotyped on the remaining subjects. A genome scan based on two-locus models is conducted using these chromosomes and genotype data from all the n subjects to identify the two QTLs.

Given the genotypes at the two QTLs, the trait of a backcross subject i is modeled as:

$$y_i \;=\; \delta_1(2g_{i1} - 1) + \delta_2(2g_{i2} - 1) + \epsilon_i \; . \tag{15}$$

The score statistic, denoted $Z(t_1, t_2)$, can be calculated based on the above model at any two genomic locations t_1 and t_2 to test the null hypothesis of no QTL (δ_1 = 0 = δ_2) at those two loci. The square of test statistic has a chi-square distribution with 2 degrees of freedom and non-centrality parameter:

$$\lambda(\delta_1, \delta_2, n) \;=\; n[1 - 4q(1 - q)](1 - r)(\delta_1^2 + \delta_2^2) \; . \tag{16}$$

The overall significance level α is approximated as (Dupuis and Siegmund 1999):

$$\begin{aligned} \alpha \;&=\; P_0(\max_{t_1,t_2} Z^2(t_1, t_2) > b^2) \\ &=\; 1 - \exp\left\{ -\frac{(C + L/\Delta)^2}{2} \times \left[\tau(\sqrt{\Delta b^2 \zeta/2}) \right]^2 \times \left[1 - \mathcal{X}_2^2(b^2) \right] \right\} \; , \end{aligned} \tag{17}$$

where ζ = 2 and $\tau(x) = \exp\{-2\sum_{s=1}^{\infty}\Phi(-x\sqrt{s})/s\}$.

The relative information and the relative genotyping burden have the same form as equation (10) and (13). The optimal θ_1, $\alpha_1 \in (\alpha/C, 1)$ (and the corresponding $1-\beta_1$) can be obtained as described earlier. We examined the optimal two-stage design under various parametric configurations when $D = 2$ QTLs are associated with the trait. Heritability or the proportion of variation in the trait explained by the two QTLs is $h^2 = (\delta_1^2 + \delta_2^2)/(1 + \delta_1^2 + \delta_2^2)$. The optimal two-stage design was examined for various values of $\delta_1^2 + \delta_2^2$ (equivalently, h^2). The results indicate that regardless of the heritability h^2 and the genotyping density Δ, 95% or more relative information and, hence, near-optimal power is obtained when (θ_1, α_1) = (0.60, 0.20). The relative information is at least 90% when (θ_1, α_1) = (0.60, 0.10). The operating characteristics of these optimal two stage designs are described in Table 3 under various parametric configurations. Our investigations suggest that, as a general guideline, genotyping all the chromosomes on θ_1 = 60% of the subjects in Stage 1 and genotyping the chromosomes significant at α_1 = 20% provides most of the information and utilizes only 70% of the genotyping relative to a one-stage design. This is consistent with the general guideline obtained for backcross and F2 crosses based on a single QTL model.

Two Stage Analytic Approach

Complex traits may be a consequence of multiple QTLs conferring effects individually (main effects) or solely through epistasis (interaction effects), or both. Conducting a genome scan using single QTL models may not provide adequate power to detect all the QTLs. Fitting multi-locus models to identify the QTLs can, however, be a tedious task, particularly under dense genotyping. A sequential analytic strategy can be employed to identify multiple QTLs in such cases. Here we examine a two-stage analytic approach when a dense set of markers is genotyped on all the n available individuals.

We consider the case where the trait is generated through the following model consisting of the main effects and interaction between two unlinked QTLs:

$$y_i \;=\; \delta_1(2g_{i1}-1) + \delta_2(2g_{i2}-1) + \delta_3(4g_{i1}g_{i2}-1) + \epsilon_i \; . \tag{18}$$

Here δ_3 is the interaction effect. The phenotypic trait has marginal mean 0 and marginal variance $\tau^2 = 1+\delta_1^2+\delta_2^2+3\delta_3^2+\delta_1\delta_3+\delta_2\delta_3$. In our investigations below, we assume $\delta_1 = \delta_2$. The heritability based on the above model is $h^2 = h_{12}^2 + h_3^2$, where $h_{12}^2 = (\delta_1^2+\delta_2^2)/\tau^2$ and $h_3^2 = \delta_3(\delta_1+\delta_2+3\delta_3)/\tau^2$. The quantity $\eta_{12} = h_{12}^2/h^2$ can be interpreted as the fraction of heritability conferred solely by the

main effects of the two QTLs. Therefore, $\eta_{12} = 1$ implies that $\delta_3 = 0$, and $1 - \eta_{12} = h_3^2/h^2 = 1$ implies that $\delta_1 = 0 = \delta_2$.

Suppose we conduct a one-stage genome scan using single QTL models. As described in the previous section, the power to detect the two QTLs under this approach is $P_1 \times P_2$. Now suppose that we conduct a one-stage genome scan using two QTL models, where two main effects terms and a pairwise interaction term are considered as in equation (18). In this setting, there is no QTL under the null hypothesis, and there are two QTLs under the alternative. The test statistic, denoted Z^2, calculated at the two QTL positions using the above model has a χ^2 distribution with 3 degrees of freedom and non-centrality parameter given by:

$$\lambda(\delta_1, \delta_2, \delta_3, n) \;\; = \;\; n \times \left\{\delta_1^2 + \delta_2^2 + 3\delta_3^2 + \delta_1\delta_3 + \delta_2\delta_3\right\} \; . \tag{19}$$

The overall significance level α is given by equation (17) with $\zeta = 4/3$. The power of the one-stage genome scan using two-QTL models is $1 - \beta = 1 - \Phi[b - \sqrt{\lambda(\delta_1, \delta_2, \delta_3, n)}]$. The computational burden of this one-stage approach is the total number of chromosomes that will be considered for the pair-wise genome scan. Since two-locus models will be fit using markers on all the chromosomes, the computational burden is $CB_1 = C$, the total number of chromosomes.

Consider the following two-stage analytic approach. In Stage 1, conduct a genome scan using single QTL models and test each chromosome at significance level α_1. In Stage 2, conduct a two-locus scan by applying model (18) to pairs of loci on the chromosomes identified as being significant in Stage 1 and test the maximum test statistic at significance level α_2. Let $1 - \beta_1$ and $1 - \beta_2$ denote the powers to detect the chromosomes containing two QTLs in Stage 1. Different test statistics are calculated in the two stages. The test statistic in Stage 1 is based on a single QTL model, while that in Stage 2 is obtained using a two-QTL model. The power, P^*, of the two-stage approach is the joint probability that the chromosomes containing the QTLs are identified in Stage 1, and the two locus test statistic Z^2, calculated at the QTL positions, exceeds b^2 in Stage 2. Therefore,

$$\begin{aligned}
P^* &= P_A(\text{the two QTL chromosomes are selected in Stage 1 and } Z^2 > b^2) \\
&= P_A(\text{the two QTL chromosomes are selected in Stage 1}) \times \\
&\quad P_A(Z^2 > b^2 | \text{QTL chromosomes are selected in Stage 1}) \\
&= (1 - \beta_1) \times (1 - \beta_2) \times P_A(Z^2 > b^2 | \text{QTL chromosomes are selected in Stage 1}) \\
&= (1 - \beta_1) \times (1 - \beta_2) \times P_A(Z^2 > b^2 | \text{test statistic is calculated at the two QTL positions}) \\
&= (1 - \beta_1) \times (1 - \beta_2) \times (1 - \beta) \; .
\end{aligned} \tag{20}$$

A two locus model will be fit at the two putative loci only when the relevant chromosomes are selected in Stage 1. Therefore, P^* can be substantially smaller than $1 - \beta$. The computational

burden of the proposed two stage design is $CB_2 = \sum_{j=1}^{D}(1-\beta_j)+\alpha_1(C-D)$. Clearly, $CB_2 \leq CB_1$. Therefore, the relative computational burden is given by:

$$\frac{CB_2}{CB_1} = \frac{\sum_{j=1}^{D}(1-\beta_j)}{C} + \alpha_1 \times \left(1 - \frac{D}{C}\right) . \quad (21)$$

Our goal is to identify the two stage parameter α_1 so that (i) the overall significance level is α, (ii) the loss of power, $1 - \beta - P^*$, is smaller than ϵ for some desired ϵ, and (iii) the relative computational burden is minimized.

Since $\delta_1 = \delta_2$, we have $1-\beta_1 = 1-\beta_2$. The optimal two-stage parameters can be identified as follows. The values of α, $1-\beta$, C, L, n, and the desired effect size δ_1 (= δ_2), and δ_3 are specified when designing the study. The value of ϵ is specified by the user.

1. Choose $\alpha_1 \in (\alpha/C, 1)$.
2. Obtain critical value b_1^2 from equation (5) by setting $C = 1$ and using L/C in place of L on the right hand side.
3. Calculate the power of Stage 1, $1 - \beta_1$, using equation (6), noting that the test statistic at a QTL locus in Stage 1 has a marginal χ^2 distribution with 1 degree of freedom and non-centrality parameter $n \times \delta_1^2/(\tau^2 - 1 - \delta_1^2)$.
4. Calculate $1 - \beta - P^* = 1 - \beta - (1 - \beta_1)^2 \times (1 - \beta)$. Check if this value is $\leq \epsilon$.
5. Calculate the relative computational burden CB_2/CB_1 using equation (21).
6. Repeat the above procedure for various choices of α_1 to identify that α_1 providing $1 - \beta - P^* \leq \epsilon$ and the smallest relative computational burden.

We examined the power to detect two QTLs under three methods: one-stage genome scan using single locus models, one-stage genome scan using two-locus models with main effects and a pair-wise interaction term, and the proposed two-stage analytic approach. Consider a hypothetical genome with C = 20 densely genotyped chromosomes, each of length 100 centi-Morgans. The desired overall significance level is α = 0.05. Table 4 gives the power of the three methods. The power of the two-stage analytic approach is provided for α_1 = 0.10, 0.20, and 0.25. The results indicate that over a broad range of values of h^2, the two-stage approach provides sufficient power to identify both the QTLs so long as a reasonable fraction of heritability is explained by the main effects. As a general rule, a two-stage approach with α_1 = 0.25 provides sufficient power to detect both the QTLs when $\eta_{12} \geq 0.75$. The loss of power is $\epsilon \leq 10\%$. This result holds

for all of the configurations examined, regardless of the value of h^2. The power of this design is given in Column 8 of Table 4, and should be compared with the power of the one-stage genome scan using two-locus models given in Column 9. When α_1 = 0.25, the relative computational burden is 32% i.e., this two-stage analytic approach requires evaluating 68% fewer two-locus models to identify the two QTLs than a one-stage approach. Column 5 provides the power to detect the two loci using a one-stage genome scan with single locus models. Comparing this with Columns 6, 7, and 8, it is evident that the proposed two-stage analytic approach is substantially more powerful than a one-stage single locus genome scan.

Discussion

In this paper we have outlined two-stage sequential methods for QTL mapping in experimental crosses. Recent developments in molecular technology enables us conduct dense genotyping. Sequential genotyping methods can provide a cost-efficient strategy to search for QTLs without having to genotype all the markers on all the study subjects. Our investigations show that when a single QTL is associated with the phenotypic trait, genotyping all the markers on only 60% of the individuals in Stage 1 and genotyping the markers on chromosomes significant at 20% level using additional individuals in Stage 2 provides near-optimal power to identify the putative locus and utilizes only 70% genotyping burden relative to a one stage design. When two QTLs are associated with the trait, this guideline continues to provide an efficient approach to identify the QTLs when both Stages 1 and 2 involve genome scan using two locus models. The optimal parameters are independent of the heritability and genotyping density. When planning a two-stage design, the sample size n is initially calculated to detect a QTL with some desired power and significance level, assuming a genotyping density of Δ centiMorgans, as if one were conducting a one-stage design. Genotyping is then conducted using a two-stage approach by treating this n as the fixed available sample size. Therefore, while the sample size will depend upon several parameters including Δ, the choice of sampling fraction and significance level for Stage 1 do not depend upon the QTL effect and Δ once the sample size is fixed. Our investigations based on a single QTL model indicate that the general guideline is applicable to both backcross as well as F2 crosses. If the cost of genotyping is linear in the total number of markers, this two stage design provides a cost-effective genotyping strategy. These guidelines have parallels to optimal two-stage genotyping strategies for association studies in human population.

Note that once the promising markers are genotyped on additional individuals in Stage 2, the analysis of these markers is based on the entire samples i.e., the individuals genotyped in Stage

1 as well as Stage 2. While Stage 2 may be viewed as a replication study and the analysis may be conducted solely based on Stage 2 data, this would result in an inefficient genetic mapping strategy relative to a joint analysis of Stage 1 and Stage 2 samples (Skol et al. 2006). This is consistent with intuition since the joint analysis utilizes a larger sample size than an analysis solely based on Stage 2 data. This article implicitly assumes that the cost of genotyping a single marker is the same in the two stages. Wang et al. (2006) examine differential costs for genotyping a large set of markers in Stage 1 and a smaller subset in Stage 2, conclude that two-stage designs provide an optimal genotyping strategy by examining the power function, and provide optimal two-stage design parameters under different cost settings.

The contribution to the likelihood from each individual is based on a mixture model. A reviewer pointed out that the asymptotic normality of the score statistic and the asymptotic chi-squared distribution of the likelihood ratio statistic may be violated for mixture models. This violation occurs when the mixing proportion q_i is 1/2 for all the individuals, resulting in a singular information matrix (see for example, Quinn et al. 1987). Consider a backcross. Note that $q_i = 1/2$ only when the flanking markers recombine, and $q_i = (1 - r_1)^2/(1 - r)$ or $r_1^2/(1 - r)$ when they do not recombine. In fact, if we observe that q_i = 1/2 for all the subjects, this is an indication that all the study subjects have recombining flanking markers, which should not occur in a well-designed study. A randomized study should yield both recombinant and non-recombinant pairs of markers. Hence, $q_i \neq 1/2$ for all the backcross subjects in a well-designed study. Any chromosome with genotyped markers should not have this issue, and we consider testing for the presence of a QTL at specific locations in intervals flanked by a pair of genotyped markers. Similar argument holds for other crosses as well. Hence, the expected information is non-singular and the asymptotic distributions of the score statistic and the likelihood ratio statistic are valid.

We have used the approximation to the tail probability of the test statistic given by equation (5) in our calculations. This approximation is valid for equally spaced markers, and the tail probability would be over-estimated when the inter-marker distances are not the same (Malley et al. 2002). Our main focus here is study design, where the investigator typically makes some assumption about average inter-marker distances to evaluate the sample size and power. The approximation to the significance level would be reasonable for addressing design considerations under such settings. Since the tail probability is over-estimated when inter-marker distances vary, the resulting sample size n would be conservative. While the approximation to the significance level (equation 5) may be reasonable for study designs, it may be pragmatic to calculate p-values using a resampling approach during data analysis. Resampling methods have been described by Churchill and Doerge (1994) and Malley et al. (2002) for a one-stage design, and by Lin

(2006) for a two-stage design. In this paper we have defined power as the probability that the test statistic at the QTL position(s) exceeds a desired critical value. This is our working definition of power, and may have some limitations. Alternatively, one may define power as the probability that the QTL position is within the confidence interval for the position corresponding to the maximum test statistic. Such a definition may guarantee better likelihood of detecting the QTL than our working definitions. Two-stage designs under this alternative definition of power is yet to be examined, and has not been attempted here.

The number of QTLs is unknown at the outset. Further, it is unknown whether multiple loci are associated with the trait solely through main effects or confer effects through interactions. It is, therefore, natural to conduct an initial genome scan using single locus models to identify the promising chromosomes in Stage 1. Multi-locus models can then be fit using the markers on these chromosomes in Stage 2 to identify the QTLs. Our investigations indicate that this approach provides substantial computational efficiency to identify multiple QTLs with sufficient power so long as at least 75% of the heritability is due to the main effects of the QTLs. This is consistent with intuition that a genome scan based on single locus models in Stage 1 may not have sufficient power to detect the chromosomes containing the putative QTLs when two loci confer substantial interaction effect. Posterior summaries for the presence of QTL at a locus (for example, Sen and Churchill 2001) obtained using hierarchical models can provide a more powerful genome scan strategy to identify QTLs. Such methods can be used for genome scanning in Stage 1 to identify the promising chromosomes. Further research is required to assess the efficiency of two stage designs that employ such analytic approach in Stage 1 and the resulting cost-efficiency and gains in computational burden, and has not been attempted here.

The proposed two stage design is conceptually similar to group sequential designs used in clinical trials (Jennison and Turnbull 2000). In a group sequential trial, a set of individuals are recruited into the trial and randomly assigned to have or not have the treatment. Outcomes from the treatment and control groups are compared, and the study proceeds to the next stage if the comparison is not statistically significant. Otherwise, the trial is stopped. Under the proposed two stage design, the study proceeds to the next stage if a chromosome is declared significant. While one or a few treatments are evaluated in a group sequential clinical trial, a QTL study involves testing multiple genetic loci. The overall significance error in a group sequential clinical trial is written as a sum of type I errors over multiple stages. Under the proposed two stage design the probability of not finding a false positive (equation 22 in the Appendix) is written as the sum of the corresponding probabilities in successive stages. This provides insight into the choice of

significance level α_1 for Stage 1, and the choice of significance level α_2 and the corresponding critical value in Stage 2, as given by Results 1, 2, and 3.

DNA pooling (Churchill et al. 1993) is another cost-effective genotyping strategy whereby n given subjects are grouped into m pools consisting of k individuals each such that $n = m \times k$. DNA is isolated *en masse* from each pool. Thus, the total number of genotyping per marker is $m < n$. Two-stage and multi-stage designs involving DNA pooling in Stage 1 and individual genotyping of the promising loci in subsequent stages have been examined for population-based association studies in humans (Zuo et al. 2006; Prentice and Qi 2007). Further research is needed to examine the optimal properties of two-stage or multi-stage QTL mapping studies involving DNA pooling. Cost-effective genetic mapping methods have also been investigated from phenotyping perspectives, particularly when multiple phenotypes are measured and/or when phenotyping is more expensive than genotyping. Jin et al. (2004) proposed a selective phenotyping strategy using a criterion that maximizes the genetic diversity of the phenotyped subjects. Medugorac and Soller (2001) considered cost-information tradeoffs under selective phenotyping with a main trait of interest and a correlated trait. This article focuses solely on the case where genotyping, but not phenotyping, cost can be substantial. The role of two-stage designs under selective phenotyping remains to be examined, and has not be considered here.

Our investigations where a single locus or multi-locus model is used in both stages make use of the fact that the non-centrality parameter is a mixture of the corresponding quantities in the two stages, thus providing a direct approach to evaluate optimal two stage design without the need to examine the power. Two stage designs have been used for mapping disease susceptibility loci in experimental organisims and human population. Sugiyama et al. (2001) used a novel two stage design for mapping traits associated with salt-induced hypertension in 250 rats. A selective genotyping approach was used in Stage 1, by genotyping all the markers on 92 mice with extreme trait values to identify the promising chromosomal regions. In Stage 2, the recombinant regions were densely genotyped on all the 250 mice if the flanking markers recombined. The rational behind this approach is that the genotypes of the intermediate loci are completely known if the flanking markers do not recombine. The design proposed in this article assmes that, once a chromosome is declared significant in Stage 1, all the markers on this chromosome are genotyped on additional individuals in Stage 2. Alternatively, one may genotype markers from only those intervals having a significant LOD score, instead of having to genotype all the markers on that chromosome. This approach can further reduce the genotyping burden. The proposed two stage design can be extended to encompass designs of this kind. However, the significance level b_2^2 for Stage 2 may be different from b^2, and further work is needed to obtain

the appropriate b_2^2. Maraganore et al. (2005) employed a novel approach to identify genetic loci associated with Parkinson disease using family-based as well as population-based case-control samples. A dense set of single nucleotide polymorphisms (SNPs) were genotyped on discordant siblings in Stage 1. The promising SNPs were genotyped in Stage 2 on unrelated case-control samples. Data on these promising SNPs from the discordant siblings and case-control samples were used to identify the putative disease loci. The rational behind this approach is that the use of discordant siblings reduces population stratification issues and, hence, false positive associations, and the use of case-control samples provides improved power to identify the putative loci. Likewise, different crosses (such as backcross, F2 or recombinant inbred lines) may be used in different stages to fine map the disease loci. The direct approach for evaluating power loss can provide a framework for investigating the operating characteristics of such designs. The efficiency of two stage designs employing different samples or crossses in different stages remains to be explored. Computer programs written using the R programming language (http://cran-r-project.org) are available from the authors to perform the two-stage design calculations described in this article.

Appendix

Mean of the score statistic: Single QTL Model

Here we describe the mean of the score statistic Z for a single QTL model based on the likelihood given by equation (2). The first derivative of the log-likelihood with respect to δ and evaluated at $\delta = 0$ is given by $l^{'}(\delta = 0; r_1, n) = \sum_{i=1}^{n} y_i(2q_i - 1)$. The expected value of $l^{'}(\delta = 0; r_1, n)$ based on n individuals with independently distributed phenotypic traits and independent and identically distributed genotypes is given by:

$$\begin{aligned} E\{l^{'}(\delta = 0; r_1, n))\} &= nE\{(2q_i - 1)E(y_i|q_i)\} \\ &= n\delta E\{(2q_i - 1)^2\} \\ &= n\delta E\{1 - 4q_i(1 - q_i)\} \\ &= n\delta \sum P(m_{1i}, m_{2i})\{1 - 4P(g_i = 1|m_{1i}, m_{2i})P(g_i = 0|m_{1i}, m_{2i})\} \ . \end{aligned}$$

The summation in the last row is over the four possible flanking marker genotypes $(m_{1i}, m_{2i}) = \{(1,1),(1,0),(0,1),(0,0)\}$. Since the QTL is assumed to be located in the center of the marker interval, it can be easily seen that the expected value is $n\delta(1-r)[1-4q(1-q)] = \delta I(\delta = 0; r_1, n)$. Hence, the mean of the score statistic is $E(Z) = \delta\sqrt{I(\delta = 0; r_1, n)}$.

Derivation of Results 1 - 4

Results 1 - 4 are derived below. These results are applicable to any QTL model (for example, one-QTL or two-QTL models) and any cross (for example, backcross or F2).

Proof of Results 1 and 2:

On any chromosome c, let $W_1^{(c)}$ and $W_2^{(c)}$ denote the chromosome-wide maximum test statistic calculated using $n\theta_1$ and n individuals, respectively. We observe $W_1^{(c)}$ in Stage 1 and $W_2^{(c)}$ in Stage 2 (if the chromosome is evaluated in Stage 2). The significance level in Stage 1 is the probability that $W_1^{(c)}$ exceeds critical value b_1^2, under the null hypothesis. The significance level in Stage 2, denoted α_2, is the probability that $W_2^{(c)}$ exceeds b_2^2 in Stage 2 conditional upon the fact that $W_1^{(c)}$ exceeded b_1^2 in Stage 1. The probability of finding no false positive association on

a chromosome c under the null hypothesis is:

$$\begin{aligned}
1-\frac{\alpha}{C} &= P_0(\text{there is no false positive association at the end of the study}) \\
&= P_0(\text{there is no false positive association in Stage 1}) + \\
&\quad P_0(\text{there is no false positive association in Stage 2} \\
&\quad\quad \text{and there is a false positive association in Stage 1}) \\
&= P_0(W_1^{(c)} < b_1^2) + P_0(W_2^{(c)} < b_2^2, W_1^{(c)} > b_1^2) \\
&= P_0(W_1^{(c)} < b_1^2) + P_0(W_2^{(c)} < b_2^2 | W_1^{(c)} > b_1^2) \times P_0(W_1^{(c)} > b_1^2) \\
&= 1-\alpha_1 + (1-\alpha_2) \times \alpha_1 \; . \qquad (22)
\end{aligned}$$

Since α_1 and $\alpha_2 \in (0,1)$, this equation indicates that $1-\alpha_1 < 1-\alpha/C$ i.e., $\alpha_1 > \alpha/C$, proving Result 1. Further, it follows from the above equation that $\alpha_2 = \alpha/(C \times \alpha_1)$, proving Result 2.

Proof of Result 3:

The overall significance level of the two-stage design is the probability that a chromosome is declared significant in Stage 1 and Stage 2. Denoting b_1^2 and b_2^2 as the critical values in Stages 1 and 2, respectively, the overall significance level of the two-stage design is defined as $P_0(W_1^{(c)} > b_1^2, W_2^{(c)} > b_2^2)$. Further, we can write $P(W_2^{(c)})$ as

$$\begin{aligned}
P_0(W_2^{(c)} > b_2^2) &= P_0(W_2^{(c)} > b_2^2, W_1^{(c)} < b_1^2) + P_0(W_2^{(c)} > b_2^2, W_1^{(c)} > b_1^2) \\
&\geq P_0(W_2^{(c)} > b_2^2, W_1^{(c)} > b_1^2) \; .
\end{aligned}$$

Let b^2 denote the critical value of a one-stage design. The overall significance level of a one-stage design is α/C, and is defined as the probability that the test statistic $W_2^{(c)}$ exceeds b^2 under the null hypothesis i.e., $\alpha/C = P_0(W_2^{(c)} > b^2)$. Our goal is to design a two-stage procedure to have overall significance level of at most α/C. Equivalently, we need the right hand side of the above equation to be at most α/C. From the above equation it can be easily seen that when the critical of Stage 2 is set to equal b^2, we have $P_0(W_2^{(c)} > b^2, W_1^{(c)} > b_1^2) \leq \alpha/C$, proving Result 3.

Proof of Result 4:

Under a one-stage design, the power to detect a QTL under model is given by $1-\beta = P_A(Z > b_2^2)$, where Z is a random variable distributed as chi-square with 2 degrees of freedom and non-centrality parameter $\lambda \doteq \lambda(\delta_a, \delta_d, n)$. The probability density of Z is given by (Johnson, Kotz, and

Balakrishnan 1995)

$$f(z) \;=\; \exp\{-\lambda/2\} \sum_{r=0}^{\infty} \left(\frac{\lambda}{2}\right)^r \; \frac{1}{r!} \; g(z; 1+r, 1/2) \;,$$

where $g(z; 1+r, 1/2)$ is the probability density of a gamma distribution with shape $1+r$ and scale $1/2$, having cumulative distribution:

$$G(z; 1+r, 1/2) \;=\; \int_0^z \frac{1}{\Gamma(1+r)} \; \left(\frac{1}{2}\right)^{1+r} \; u^r \; \exp\{-2u\} \; du \;.$$

When the sample size is fixed and the QTL effects δ_a and δ_d are small, resulting in small non-centrality parameter λ, $\beta = P_A(Z < b_2^2)$ can be written as:

$$\begin{aligned} \beta &= \int_0^{b_2^2} f(z) \; dz \\ &= \exp\{-\lambda/2\} \sum_{0}^{\infty} \left(\frac{\lambda}{2}\right)^r \; \frac{1}{r!} \; G(b_2^2; 1+r, 1/2) \\ &= \left(1 - \frac{\lambda}{2}\right) \times \left(G_1 + \frac{\lambda}{2} G_2\right) \;. \end{aligned}$$

The second step follows by taking the integral inside the summation, since $G(.)$ is a cumulative distribution and hence is in the interval $(0, 1)$. The third step utilizes the fact that λ is small and, hence, the first two terms are used to approximate the infinite sum. In the last step $G_1 = G(b_2^2; 1, 1/2)$ and $G_2 = G(b_2^2; 2, 1/2)$. For small λ, it can be easily seen that $\beta < G_1$. Therefore, solving the above quadratic equation for a given β provides:

$$\lambda \;=\; \frac{(G_2 - G_1) + \sqrt{(G_2 - G_1)^2 + G_2(G_1 - \beta)}}{G_2}$$

The right hand side depends upon the power $1-\beta$ and the critical value b_2^2, and is independent of the QTL effects. Similar expression can be written for $\lambda(\delta_a, \delta_d, n\theta_1)$, the non-centrality parameter of Stage 1, as a function of Stage 1 power $1 - \beta_1$ and the corresponding critical value b_1^2. Define $G_{21} = G(b_1^2; 2, 1/2)$ and $G_{11} = G(b_1^2; 1, 1/2)$. Therefore:

$$\begin{aligned} \theta_1 &= \frac{\lambda(\delta_a, \delta_d, n\theta_1)}{\lambda(\delta_a, \delta_d, n)} \\ &= \frac{(G_{21} - G_{11}) + \sqrt{(G_{21} - G_{11})^2 + G_{21}(G_{11} - \beta_1)}}{(G_2 - G_1) + \sqrt{(G_2 - G_1)^2 + G_2(G_1 - \beta)}} \times \frac{G_2}{G_{21}} \end{aligned}$$

The right hand side, and hence θ_1, is independent of the QTL effects.

Non-centrality parameter of F2 cross

The test statistic at the QTL position of an F2 cross has a χ^2 distribution with 2 degrees of freedom. The non-centrality parameter can be obtained using the second derivative of the log-likelihood. The QTL is in the middle of a marker interval of length Δ centi Morgans. Using symbolic calculations, the non-centrality parameter is given by $A\delta_a^2 + D\delta_d^2$, where:

$$\begin{aligned} A &= [1 - 4q(1-q)](1-r) \\ D &= A^2 \times \frac{a_1}{a_2} \\ a_1 &= 6r^4 - 12 * r^3 + 10r^2 - 4 * r + 1 \\ a_2 &= 8r^4 - 16r^3 + 12r^2 - 4r + 1 \end{aligned}$$

References

Broman K and Speed T (1999). A review of methods for identifying qtls in experimental crosses. In F. Seillier-Moiseiwitsch (Ed.), *Statistics in genetics and molecular biology, Volume 33 of IMS Lecture Notes - Monograph Series*, pp. 114–142. Institute of Mathematical Statistics and American Mathematical Society, Hayward, CA.

Churchill G. A and Doerge R. W (1994). Empirical threshold values for quantitative trait mapping. *Genetics 138*, 963–971.

Churchill G. A, Giovannoni J. J, and Tanksley S. D (1993). Pooled-sampling makes high-resolution mapping practical with DNA markers. *Proceedings of the National Academy of Sciences 90*, 16–20.

Cox D and Hinkley D (1974). *Theoretical Statistics*. Chapman and Hall, London.

Darvasi A and Soller M (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics 85*, 353–359.

Darvasi A and Soller M (1994). Optimum spacing of genetic markers for determining linkage between marker loci and quantitative trait locus. *Theoretical and Applied Genetics 89*, 351–357.

Dupuis J and Siegmund D (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics 151*, 373–386.

Elston R. C (1994). P value, power, and pitfalls in the linkage analysis of psychiatric disorders. In E. Gershon and C. Cloninger (Eds.), *Genetic Approaches to Mental Disorders: Proceedings of the Annual Meeting of the American Psychopathological Association*, pp. 3–21. American Psychiatric Press, Washington D.C.

Elston R. C, Guo X, and Williams L. V (1996). Two-stage global search designs for linkage analysis using pairs of affected relatives. *Genetic Epidemiology 13*, 535–558.

Elston R. C, Lin D, and Zheng G (2007). Multi-stage sampling for genetic studies. *Annual Review of Genomics and Human Genetics 00*, 00–00.

Feingold E, Brown P. O, and Siegmund D (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identify by descent. *American Journal of Human Genetics 53*, 234–251.

Jennison C and Turnbull B. W (2000). *Group sequential methods with application to clinical trials*. Chapman and Hall, New York.

Jin C, Lan H, Attie A. D, Churchill G. A, and Yandell B. S (2004). Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics 168*, 2285–2293.

Johnson N, Kotz N, and Balakrishnan N (1995). *Continuous univariate distributions, Volume 2*. John Wiley and Sons, New York.

Lander E. S and Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics 121*, 185–199.

Lin D. Y (2006). Evaluating statistical significance in two-stage genomewide association studies. *American Journal of Human Genetics 78*, 505–509.

Malley J. D, Naiman D. Q, and Bailey-Wilson J. E (2002). A comprehensive method for genome scans. *Human Heredity 54*, 174–185.

Maraganore D. M, de Andrade M, Lesnick T. G, et al. (2005). High-resolution whole-genome association study of parkinson disease. *American Journal of Human Genetics 77*, 685–693.

Marchini J, Donnelly P, and Cardon L. R (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics 37*, 413–417.

Medugorac I and Soller M (2001). Selective genotyping with a main trait and correlated trait. *Journal of Animal Breeding and Genetics 118*, 285–295.

Ott J (1991). *Analysis of Human Genetic Linkage*. The Jhons Hopkins University Press, Baltimore.

Prentice R. L and Qi L (2007). Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation. *Biostatistics 7*, 339–354.

Quinn B. G, McLachlan G. J, and Hjort N. L (1987). A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *Journal of the Royal Statistical Society – Series B 49*, 311–314.

Satagopan J. M and Elston R. C (2003). Optimal two-stage genotyping in population-based association studies. *Genetic Epidemiology 25*, 149–157.

Satagopan J. M, Venkatraman E. S, and Begg C. B (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics 60*, 589–597.

Satagopan J. M, Verbel D. A, Venkatraman E. S, Offit K. E, and Begg C. B (2002). Two-stage designs for gene-disease association studies. *Biometrics 58*, 163–170.

Sen S and Churchill G. A (2001). A statistical framework for quantitative trait mapping. *Genetics 159*, 371–387.

Sen S, Satagopan J. M, and Churchill G. A (2005). Quantitative trait locus study design from an information perspective. *Genetics 170*, 447–464.

Skol A, Scott L. J, Abecasis G. R, and Boehnke M (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics 38*, 209–213.

Sugiyama F, Churchill G. A, Higgins D. C, Johns C, et al. (2001). Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics 71*, 70–77.

Thomas D. C (2005). The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention 14*, 557–559.

Wang H, Thomas D. C, Peer I, and Stram D. O (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genetic Epidemiology 30*, 356–358.

Zuo Y, Zuo G, and Zhao H (2006). Two-stage designs in case-control association analysis. *Genetics 173*, 1747–1760.

Table Captions.

Table 1: Optimal two stage design parameters for a backcross under a single locus model with 60% and 70% relative genotyping burden. The optimal parameters are derived for a hypothetical genome with 20 chromosomes, each of length 100 centiMorgans. Column 1 (Δ) provides the average marker distance in centiMorgan units. Column 2 (RGB) is the relative genotyping burden. Column 3 (α_1) is the Stage 1 significance level for testing each chromosome. Column 4 ($1-\beta$) is the Stage 1 power. Column 5 (θ_1) is the sampling fraction for Stage 1. Column 6 (RI) is the relative information. These results suggest that genotyping all the chromosomes on $\theta_1 \approx$ 60% of the subjects in Stage 1 and genotyping the chromosomes significant at level $\alpha_1 \approx$ 0.20 on the remaining individuals in Stage 2 provides nearly 95% of the information relative to a one stage design.

Table 2: Two-stage design parameters for an F2 cross based on a single locus model under the general guideline of θ_1 = 0.60 and α_1 = 0.20. Column 1 (C) is the number of chromosomes of the hypothetical genome. The length of each chromosome is assumed to be 100 centiMorgans. Column 2 gives the additive (δ_a) and dominance (δ_d) effects of a single QTL and the heritability $h^2 = \kappa^2/(1+\kappa^2)$, where $\kappa^2 = 3/16 \times (\delta_a + \delta_d/2)^2 + 1/16 \times (\delta_d/2)^2 + 3/16 \times (\delta_a - \delta_d/2)^2$. Column 3 ($n$) is the total sample size so that the power of the one stage design to detect a QTL with heritability h^2 is approximately 80% at 5% significance level. Column 4 (Δ) is the average marker distance in centiMorgan units. Column 5 ($1-\beta_1$) is the Stage 1 power. Column 6 (RGB) is the relative genotyping burden of the two stage design. Column 7 (RI) is the relative information of the two stage design.

Table 3: Characteristics of a two stage design based on a two-QTL model for a backcross segregating two QTLs. The two stage design parameters are provided under the guidelines (θ_1, α_1) = (0.60, 0.20) and (0.60, 0.10), indicated in Column 1. The length of each chromosome is assumed to be 100 centiMorgans. Column 2 (C) is the number of chromosomes. Column 3 gives the heritability ($h^2 = \kappa^2/(1+\kappa^2)$), where $\kappa^2 = \delta_1^2 + \delta_2^2$. The value of κ^2 is given in parentheses. Column 4 is the total sample size n, so that the power of a one stage design is approximately 80%. Column 5 (Δ) is the average marker density in centiMorgan units. Column 6 ($1-\beta_1$) is the Stage 1 power. Column 7 (RGB) is the relative genotyping burden. Column 8 (RI) is the relative information.

Table 4: Characteristics of the two stage analytic approach to detect two QTLs. The power of a one-stage genome scan using two-QTL models is $1-\beta \approx$ 0.90. Column 1 is the

heritability h^2, with sample size n given in parentheses. Column 2 is η_{12}, the fraction of heritability conferred solely by the main effects of the two QTLs. Column 3 is the main effect of the two QTLs. Column 4 is the interaction effect. Column 5 is the power of a one-stage analysis to detect both the QTLs based on a genome scan with single-locus models. Columns 6, 7, and 8 are the power of the two-stage analytic approach under $\alpha_1 =$ 0.10 (under column denoted (a)), 0.20 (b), and 0.25 (c). Column 9 is $1 - \beta$. The relative computational burdens are approximately 18%, 28%, and 32% for $\alpha_1 = 0.10$, 0.20, and 0.25, respectively, regardless of h^2, n, and η_{12}.

Table 1.

Δ	RGB	α_1	$1-\beta_1$	θ_1	RI
20	0.60	0.16	0.80	0.50	0.903
	0.70	0.21	0.90	0.60	0.961
10	0.60	0.16	0.80	0.50	0.903
	0.70	0.21	0.90	0.60	0.961
5	0.60	0.16	0.80	0.50	0.903
	0.70	0.21	0.90	0.60	0.961
1	0.60	0.15	0.80	0.51	0.902
	0.70	0.21	0.90	0.60	0.961
0.10	0.60	0.17	0.80	0.50	0.900
	0.70	0.21	0.90	0.60	0.960

Table 2.

C	h^2 (δ_a, δ_d)	n	Δ	$1-\beta_1$	RGB	RI
20	0.01 (0.15, 0.15)	1500	20	0.90	0.69	0.96
		1550	1	0.88	0.70	0.95
	0.01 (0, 0.33)	825	20	0.90	0.69	0.96
		950	1	0.87	0.69	0.95
	0.04 (0.33, 0)	500	20	0.90	0.69	0.96
		475	1	0.87	0.69	0.96
	0.05 (0.33, 0.33)	310	20	0.91	0.69	0.96
		320	1	0.88	0.70	0.95
10	0.01 (0.15, 0.15)	1400	20	0.88	0.71	0.95
		1420	1	0.84	0.71	0.94
	0.01 (0, 0.33)	760	20	0.87	0.71	0.95
		880	1	0.84	0.71	0.94
	0.04 (0.33, 0)	460	20	0.88	0.71	0.95
		440	1	0.84	0.71	0.93
	0.05 (0.33, 0.33)	290	20	0.88	0.71	0.95
		295	1	0.84	0.71	0.94

Table 3.

(θ_1, α_1)	C	h^2 $(\delta_1^2 + \delta_2^2)$	n	Δ	$1-\beta_1$	RGB	RI
(0.60, 0.20)	20	0.25 (0.33)	360	20	0.96	0.71	0.98
			370	1	0.92	0.71	0.97
		0.15 (0.18)	1200	20	0.96	0.71	0.98
			1230	1	0.92	0.71	0.97
	10	0.25 (0.33)	320	20	0.93	0.74	0.97
			335	1	0.88	0.73	0.95
		0.15 (0.18)	1100	20	0.94	0.74	0.97
			1130	1	0.88	0.73	0.95
(0.60, 0.10)	20	0.25 (0.33)	360	20	0.92	0.67	0.97
			370	1	0.87	0.67	0.95
		0.15 (0.18)	1200	20	0.92	0.67	0.97
			1230	1	0.87	0.67	0.95
	10	0.25 (0.33)	320	20	0.88	0.70	0.95
			335	1	0.82	0.70	0.93
		0.15 (0.18)	1100	20	0.89	0.70	0.95
			1130	1	0.82	0.70	0.93

Table 4.

h^2 (n)	η_{12}	δ_1 (= δ_2)	δ_3	$P_1 \times P_2$	(a)	(b)	(c)	$1-\beta$
0.15	1	0.30	0	0.41	0.85	0.88	0.88	0.90
(260)	0.90	0.28	0.03	0.26	0.78	0.83	0.85	0.89
	0.80	0.27	0.05	0.20	0.77	0.82	0.84	0.90
	0.75	0.26	0.06	0.15	0.70	0.78	0.80	0.89
	0.60	0.23	0.10	0.05	0.53	0.64	0.68	0.89
0.10	1	0.24	0	0.50	0.86	0.88	0.89	0.91
(410)	0.90	0.224	0.022	0.35	0.80	0.84	0.85	0.89
	0.80	0.211	0.041	0.24	0.74	0.81	0.82	0.89
	0.75	0.204	0.05	0.18	0.70	0.78	0.80	0.89
	0.60	0.183	0.08	0.08	0.56	0.68	0.72	0.91
0.05	1	0.162	0	0.58	0.87	0.89	0.89	0.91
(900)	0.90	0.154	0.015	0.48	0.84	0.88	0.89	0.91
	0.80	0.145	0.028	0.34	0.80	0.85	0.86	0.91
	0.75	0.140	0.034	0.27	0.76	0.72	0.84	0.90
	0.60	0.126	0.052	0.13	0.63	0.73	0.76	0.91

Figure Legends.

Figure 1: Optimal relative genotyping burden of the two stage approach for a hypothetical genome with C = 20 chromosomes, each of length 100 centiMorgans, and inter-marker distance Δ = 1 centiMorgan for a backcross based on a single locus model. The horizontal axis represents the relative genotyping burden (equation 13). The vertical axis is the relative non-centrality parameter (equation 10), equivalently the relative information. The two dashed horizontal lines indicate 90% (ϵ = 0.10) and 95% (ϵ = 0.05) relative informations.

Figure 2: Characteristics of the two stage design with 70% relative genotyping burden. The horizontal axis represents the Stage 1 significance level, α_1. The vertical axis takes value between 0 and 1, and corresponds to the the relative non-centrality parameter i.e., relative information (bold line), the Stage 1 power $1 - \beta_1$ (dotted line), and the sampling fraction θ_1 for Stage 1 (dashed line). We find that, when the relative genotyping burden is 70%, testing all the chromosomes at 10% to 35% significance level in Stage 1 provides approximately 95% relative information. In particular, genotyping all the markers on θ_1 = 60% of the individuals in Stage 1, and conducting Stage 2 genotyping on the chromosomes significant at level α_1 = 20% provides a Stage 1 power of $1 - \beta_1$ = 90% and approximately 95% of the relative information.

Figure 3: Characteristics of the two stage design for an F2 cross with C = 20 chromosomes, each of length 100 centiMorgans, and average marker distance Δ = 1 centiMorgan. The overall significance level and power of a one stage design are α = 0.05 and $1 - \beta$ = 0.80. The heritability of the single QTL is approximately 5%. Shown are the contour plots corresponding to the relative information (left panel) and the relative genotyping burden (right panel) as a function of the sampling fraction θ_1 (horizontal axis) and Stage 1 significance level α_1 (vertical axis). The dotted vertical and horizontal lines indicate that 95% relative information is attained (left panel) with θ_1 = 0.60 and α_1 = 0.20. This corresponds to 70% relative genotyping burden (right panel).

Figure 1:

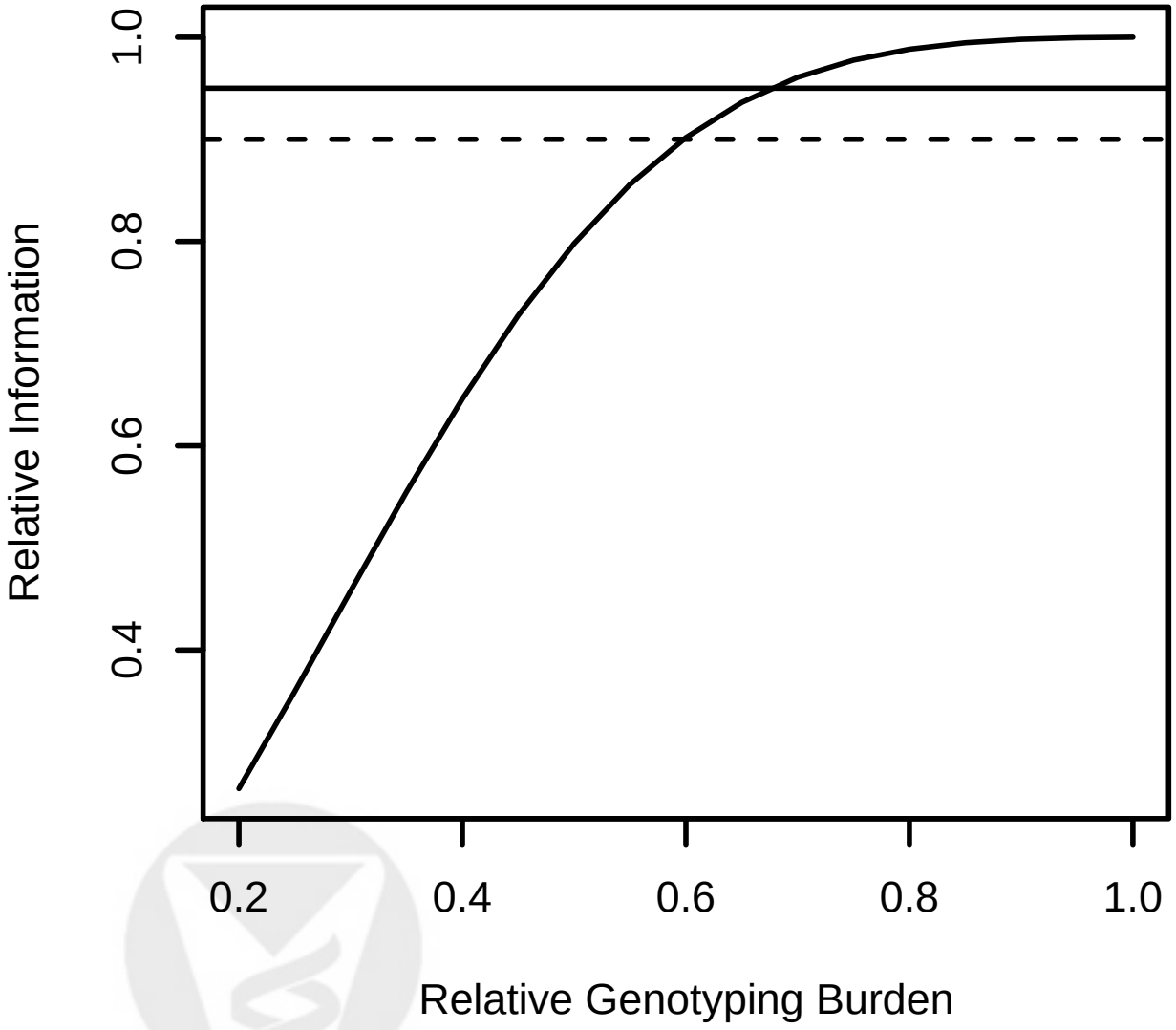

Figure 2:

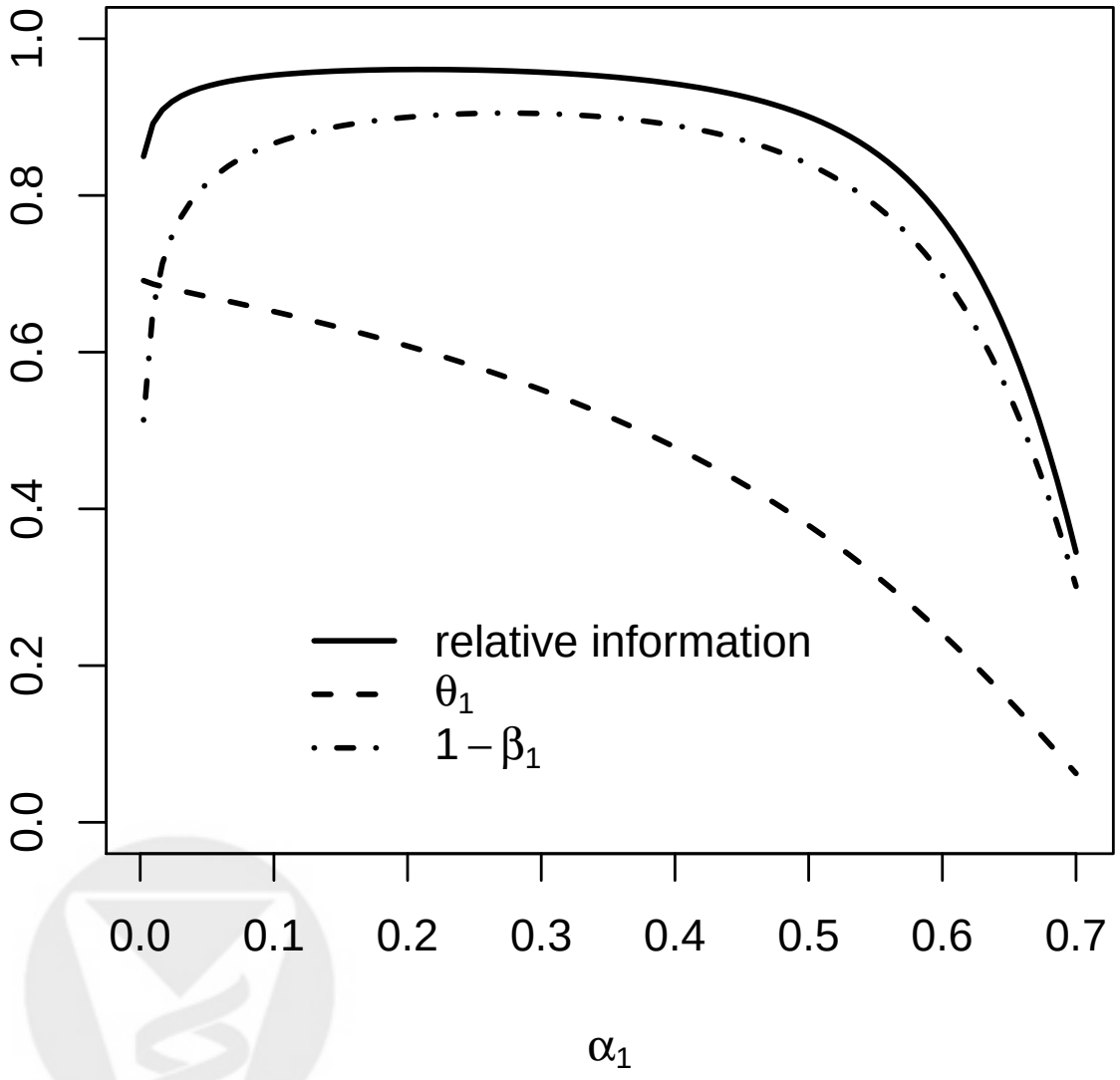

Figure 3: