

**Optimal Two-stage Genome-wide Gene Search Using Affected-sib-pair
Method**

Chi-Hse Teng

Statistical and Mathematical Sciences, Eli Lilly and Company

Lilly Corporate Center, Indianapolis, IN 46285

email: teng_chi-hse@lilly.com

and

Mark C.K. Yang

Department of Statistics, University of Florida

103 Griffin/Floyd Hall, P.O.Box 118545, Gainesville, FL 32611

email: yang@stat.ufl.edu

SUMMARY. Affected-sib-pair (ASP) method, combined with genetic marker technology, can now be used for genome-wide search for disease genes. A two-stage search consisting of a screening search followed by an intensive second-stage search is considered. The spacing between markers, the number of ASP to be used in each stage, and the marker selection criteria to pass from the first stage to the second are determined under resource constraint. Powers for rare autosomal recessive and dominant diseases are obtained using the multinomial distribution checked by simulation. The results show that in many situations a two-stage design has a higher power to locate the disease gene locus than a one-stage design. Optimal designs for resource allocation in the two stages are presented in tables.

KEY WORDS: Genome-wide scan, Genome-wide search, Affected-sib-pair method.

1. Introduction

Penrose is considered to be the first person to propose the sib-pair method (Conneally and Rivas, 1980; Shah and Green, 1994). Since Penrose proposed the sib-pair method in 1935 and the affected-sib-pair method in 1950 and 1953, these methods have been widely used in linkage studies. Combined with contemporary DNA-level genetic markers technology, the affected sib pair (ASP) method can now be used for genome-wide searches for disease genes (Botstein, 1980; Lander and Botstein, 1986, 1989). A two-stage design consisting of a screening search to eliminate nonviable marker loci and then an intensive search to identify gene location would be an useful option. Although this design has been used by researchers in genome-wide searches (Davies *et al.*, 1994; Luo *et al.*, 1995), no study has discussed optimality. Consequently, there was no guideline on how to space the markers or how to allocate the available ASPs at each stage. Elston (1992a, 1992b) studied the “optimal” two-stage design and concluded that two-stage designs were more efficient than one-stage designs in an genome-wide search for linkage, but he did not consider the problem at the multiple-test level. While Brown *et al.* (1994) studied multiple-stage genome search using affected-pedigree-member method by simulation, only one pedigree was considered and it was not obvious how to extend their results to the ASP method. Darvasi and Soller (1994) have studied the optimal spacing of genetic markers for detecting quantitative trait locus (QTL), but did not use the two-stage approach. The goal of this paper is to construct optimal two-stage genome wide gene hunting for rare autosomal recessive and dominant diseases with dichotomous phenotypes (affected or not affected).

1.1 *Two-Stage Procedure*

For the two-stage search method considered here, the first stage uses part of the ASPs with a wide spread of markers in the genome. The rest of the resources will be used

only on those promising markers identified in the first stage. Throughout this paper, “markers” and “marker loci” are synonymous, but we may use the term “marker alleles” for different alleles at the same marker locus.

Suppose there are n ASPs and there are enough resources to type c marker loci. Three numbers need to be determined in a two-stage design: n_1 and m , the number of ASPs and the number of markers to be used in the first stage (Stage I), and r , the number of markers to be used in the second stage (Stage II). The markers chosen for the second stage are based on the statistic (score),

$${}_1S_i = \sum_{j=1}^{n_1} X_{ij}, \quad (1)$$

obtained in the first stage, where X_{ij} is defined as Day and Simons’ (1967) two-alleles statistics, i.e., let $X_{ij}=1$ if i th marker IBD (identical by descent) score = 2 for the j th sib-pair, and $X_{ij}=0$ otherwise. Ideally, the r markers with the highest scores are chosen for Stage II. However, in the event of ties, more than r of them may have to be chosen. Thus, R , the actual number of markers used in Stage II, is a random variable. The formal definition of R is $R \geq r$, but if the marker(s) with the lowest score in the chosen group (markers for stage II study) is (are) taken away, then the total number of remaining markers is smaller than r .

In Stage II, R markers on N_2 ASPs are typed, where N_2 is the largest number subject to the resource constraint. Since R is a random variable, N_2 is also a random variable. Without loss of generality, let the R markers be 0, 1, ..., $R - 1$, and the sib-pairs in the Stage II be $n_1 + 1, n_1 + 2, \dots, n_1 + N_2$. Thus, N_2 is the largest x such that $mn_1 + Rx \leq c$. We define

$${}_2S_i = \sum_{j=n_1+1}^{n_1+N_2} X_{ij}. \quad (2)$$

Thus the marker with the uniquely highest ${}_2S_i$ is claimed to be the locus nearest to the disease gene. If two markers both have the same highest score and are adjacent, then

the gene is claimed to lie between them; otherwise, the gene location is undetermined.

Once the location, I , has been chosen, the next step is to check whether we can claim linkage. This result is included later in §4.3. The following assumptions are made for the ASP model.

1.2 Assumptions

- I. There is one and only one gene that can increase the disease affection rate, but the disease is allowed to have nongenetic causes.
- II. For a rare autosomal recessive disease, the parents' disease genotypes are Dd_1 and Dd_2 ; for a rare dominant disease, the parents' disease genotypes are Dd_1 and d_2d_3 , where D is the disease gene.
- III. Highly polymorphic, equally spaced markers are available. This high polymorphism assumption guarantees that the marker IBD can be determined without ambiguity. When m markers are assigned in the first stage, they are evenly distributed throughout the whole genome. Since this is for human gene hunting, 22 chromosomes are assumed. If a chromosome of length L is assigned μ markers, their positions are at $\frac{L}{2\mu}$, $\frac{3L}{2\mu}$, \dots , $\frac{(2\mu-1)L}{2\mu}$. Since the location of the gene is unknown at the design stage, the least favorable configuration, i.e., the gene lies either at the end of a chromosome or at the center of two adjacent markers, is assumed in the design.
- IV. In the same family, the probability is negligible when the disease of one affected sib is caused by a gene but the same disease of the other sib is not.
- V. The cost (for example, funds and time) of typing alleles is a constant, i.e., the cost of typing k markers from one person is same as typing one marker from

k persons. This cost ratio can be changed to suit practical situations, but the numerical result would need to be re-calculated.

VI. In our analytic approach, the X_{ij} 's are assumed to be independent, except the X_{ij} ('s) of the marker locus (loci) adjacent to the disease gene locus (This assumption is relaxed in the simulation study where the IBD scores on the same chromosome are realized from a simulated Markov chain with transition probabilities derived from recombination fractions.)

VII. The total length of chromosome 1 to 22 is equal to 3500 cM.

2. Analytic Approach.

Let the first-stage markers be numbered from 0 to $m - 1$. Without loss of generality, if the gene is at the end of the chromosome, we let the nearest marker be marker 0; if the gene is between two markers, we let it be located at the center of markers 0 and 1. If we assume that gene location is uniformly distributed along the genome, the probability of the gene being at the end is $2/m$.

Let ${}_1S_{(r)}$ be the r th highest of ${}_1S_i$, omitting the one (end case) or the two markers adjacent to the disease gene. Thus, ${}_1S_{(r)}$ contains $m - 1$ scores when the gene is at the end and $m - 2$ scores when it is at the center of the two markers. The probability of finding the gene by the two-stage method is,

$$\begin{aligned}
 & P\{\text{The marker closest to the gene is found}\} \\
 &= P\{\text{Gene is at the end of the chromosome and marker 0 is chosen in Stage II}\} \quad (3) \\
 &+ P\{\text{Gene is not at the end and marker 0 or 1 is chosen in Stage II}\}. \quad (4)
 \end{aligned}$$

It can be intuitively seen that probabilities (3) and (4) can be broken into terms like

$$\sum_{k=0}^{n_1} P\{{}_1S_0 \geq {}_1S_{(r-1)}, {}_1S_{(r-1)} = k, {}_2S_0 \text{ is the highest in Stage II}\} \quad (5)$$

IBD of the gene	IBD of the marker		
	2	1	0
2	$\frac{\Psi^2}{4}$	$\frac{\Psi(1-\Psi)}{2}$	$\frac{(1-\Psi)^2}{4}$
1	$\frac{\Psi(1-\Psi)}{2}$	$\frac{\Psi^2+(1-\Psi)^2}{2}$	$\frac{\Psi(1-\Psi)}{2}$
0	$\frac{(1-\Psi)^2}{4}$	$\frac{\Psi(1-\Psi)}{2}$	$\frac{\Psi^2}{4}$

Table 1: The joint distribution of IBD score of the disease gene and a marker.

or

$$\sum_{k=1}^{n_1} P\{ {}_1S_0 \geq {}_1S_{(r-1)} = k > {}_1S_1, {}_2S_0 \text{ is the highest in Stage II} \}. \quad (6)$$

The details of the derivation are given in Appendix I. In order to compute (5) and (6), we need to find the marginal and joint distributions of ${}_1S_0$ and ${}_1S_1$.

When a marker is not linked to the disease gene, then the probabilities of getting the IBD score of the marker equal to 2, 1, and 0 are 0.25, 0.5, and 0.25, respectively. Thus, under Assumption IV, all of the X_{ij} 's that are not next to the disease gene have the Bernoulli distribution with parameter of 0.25 and distribution of ${}_lS_i$ is binomial(n_l , 0.25), for $l=1, 2$.

When a marker and the gene are linked, the joint distribution of the IBD scores $P\{\text{IBD of the gene, IBD of the marker}\}$ is given using Table 1 adapted from Haseman and Elston (1972), where $\Psi = \theta^2 + (1 - \theta)^2$ and θ is the recombination fraction between the two marker and the gene.

The conditional distribution of of X_{ij} , given the gene IBD score derived from Table 1, is shown in Table 2.

For the cases of the gene at the end of a chromosome or only one of two marker loci adjacent to the gene locus being chosen for second stage, we need the distribution of X_{0j} (assume X_{0j} is the one chosen in the second case). To derive the distribution

Gene IBD	X_{ij}	
	1	0
2	Ψ^2	$1 - \Psi^2$
1	$\Psi(1 - \Psi)$	$1 - \Psi + \Psi^2$
0	$(1 - \Psi^2)$	$2\Psi - \Psi^2$

Table 2: The conditional distribution of X_{ij} conditioned on disease gene IBD.

of X_{0j} , let ε be the probability that the disease of a sib-pair is caused by the gene and θ be the recombination fraction between marker 0 and the gene. For $x = 0$ or 1 ,

$$\begin{aligned}
& P\{X_{0j} = x\} \\
&= \sum_{i=0}^2 Pr\{X_{0j} = x \mid \text{gene IBD}=i, \text{disease by gene}\} \cdot \\
&\quad \cdot P\{\text{gene IBD}=i \mid \text{disease by gene}\} P\{\text{disease by gene}\} \\
&\quad + P\{X_{0j} = x \mid \text{disease not by gene}\} P\{\text{disease not by gene}\}, \quad (7)
\end{aligned}$$

under Assumption V that $P\{\text{one sib is caused by gene and the other is not}\} = 0$. For a dominant disease, if an ASP is caused by a gene, then both sibs must at least share the disease gene and there is a 50-50 chance they share the other allele. Therefore,

$$\begin{aligned}
& P\{X_{0j} = 1\} \\
&= P\{X_{0j} = 1 \mid \text{gene IBD}=2\} \frac{1}{2} \varepsilon \\
&\quad + P\{X_{0j} = 1 \mid \text{gene IBD}=1\} \frac{1}{2} \varepsilon \\
&\quad + P\{X_{0j} = 1 \mid \text{disease not by gene}\} (1 - \varepsilon) \\
&= \frac{\varepsilon}{2} \Psi^2 + \frac{\varepsilon}{2} \Psi(1 - \Psi) + (1 - \varepsilon) 0.25 \\
&= \varepsilon \Psi/2 + (1 - \varepsilon) 0.25.
\end{aligned}$$

For a recessive disease, an individual has to have two recessive disease genes in order to be affected. Therefore, $P\{\text{gene IBD} = 2 \mid \text{disease by gene}\} = 1$. Eq. (7) is then equal to,

$$\begin{aligned} & P\{X_{0j} = 1 \mid \text{gene IBD} = 2\}\varepsilon \\ & + P\{X_{0j} = 1 \mid \text{disease not by gene}\}(1 - \varepsilon) \\ & = \varepsilon \Psi^2 + (1 - \varepsilon) 0.25. \end{aligned}$$

Therefore, ${}_l S_0$ is distributed as binomial($n_l, \varepsilon \Psi^2 + (1 - \varepsilon) 0.25$) for a recessive disease, and as binomial($n_l, \varepsilon \Psi/2 + (1 - \varepsilon) 0.25$) for a dominant disease.

For the cases where the gene is between two markers, we need the joint distribution of $({}_l S_0, {}_l S_1)$. To derive the joint distribution, let n_{xyl} be the number of (X_{0j}, X_{1j}) that has value (x, y) in Stage l , $x = 0, 1, y = 0, 1, j=1, \dots, n_l$. Then $(n_{00l}, n_{01l}, n_{10l}, n_{11l})$ has a multinomial distribution $(n_l, p_{00}, p_{01}, p_{10}, p_{11})$. Moreover,

$$\begin{aligned} & P\{{}_l S_0 = a, {}_l S_1 = b\} \\ & = P\{n_{10l} + n_{11l} = a, n_{01l} + n_{11l} = b\} \\ & = \sum_{k = \max(0, a + b - n)}^{\min(a, b)} P\{n_{11l} = k, n_{10l} = a - k, n_{01l} = b - k, \\ & \quad n_{00l} = n_l - a - b + n_{11l}\}, \\ & = \sum_{k = \max(0, a + b - n)}^{\min(a, b)} \frac{n_l!}{k! a - k! b - k! (n_l - a - b + k)!} p_{11}^k p_{10}^{a-k} p_{01}^{b-k} p_{00}^{n_l - a - b + k} \quad (8) \end{aligned}$$

can be computed once p_{00}, p_{01}, p_{10} and p_{11} are formulated.

If an ASP does not carry the disease gene, then from Table 1 we can deduce the joint distribution of X_{0j} and X_{1j} . The distribution is shown in Table 3 with $\Psi_2 = \theta_2^2 + (1 - \theta_2)^2$, θ_2 being the recombination fraction between the two marker loci.

	X_{1j}	
X_{0j}	1	0
1	$\frac{\Psi_2^2}{4}$	$\frac{(1-\Psi_2^2)}{4}$
0	$\frac{(1-\Psi_2^2)}{4}$	$\frac{\Psi_2^2+2}{4}$

Table 3: The joint distribution of X_{0j} and X_{1j} when a ASP does not carry the disease gene.

If an ASP carries the disease gene, we can derive the joint distribution of X_{0j} and X_{1j} as follows. Although in this paper we assume the gene is at the center of two adjacent markers, the formulae are derived for the gene anywhere in between. Let the recombination fraction between the gene and Marker 0 be θ_0 and between the gene and Marker 1 be θ_1 . Let θ_2 be the recombination fraction between two markers. The joint distribution of X_{0j} and X_{1j} , for $x = 0, 1, y = 0, 1$, is,

$$\begin{aligned}
p_{xy} &= P\{X_{0j} = x, X_{1j} = y\} \\
&= \sum_{i=0}^2 P\{X_{0j} = x, X_{1j} = y \mid \text{gene IBD}=i, \text{disease by gene}\} \cdot \\
&\quad P\{\text{gene IBD}=i \mid \text{disease by gene}\} P\{\text{disease by gene}\} \\
&\quad + P\{X_{0j} = x, X_{1j} = y \mid \text{disease not by gene}\} \cdot \\
&\quad P\{\text{disease not by gene}\}
\end{aligned} \tag{9}$$

Let $\Psi_l = \theta_l^2 + (1 - \theta_l)^2$, $l=0, 1, 2$. Applying Table 3 for a recessive disease, we have, under Assumption II,

$$\begin{aligned}
p_{00} &= (1-\Psi_0^2)(1-\Psi_1^2) \varepsilon + \frac{\Psi_2^2 + 2}{4} (1 - \varepsilon), \\
p_{10} &= \Psi_0^2(1-\Psi_1^2) \varepsilon + \frac{1 - \Psi_2^2}{4} (1 - \varepsilon),
\end{aligned}$$

$$p_{01} = (1-\Psi_0^2) \Psi_1^2 \varepsilon + \frac{1 - \Psi_2^2}{4} (1 - \varepsilon),$$

$$p_{11} = \Psi_0^2 \Psi_1^2 \varepsilon + \frac{\Psi_2^2}{4} (1 - \varepsilon).$$

Again applying Table 3 for a dominant disease, we have

$$p_{00} = (1-\Psi_0^2)(1-\Psi_1^2) \frac{\varepsilon}{2} + (1-\Psi_0+\Psi_0^2)(1-\Psi_1+\Psi_1^2) \frac{\varepsilon}{2} + \frac{\Psi_2^2 + 2}{4} (1 - \varepsilon),$$

$$p_{10} = \Psi_0^2(1-\Psi_1^2) \frac{\varepsilon}{2} + \Psi_0(1-\Psi_0)(1-\Psi_1+\Psi_1^2) \frac{\varepsilon}{2} + \frac{1 - \Psi_2^2}{4} (1 - \varepsilon),$$

$$p_{01} = (1-\Psi_0^2) \Psi_1^2 \varepsilon + (1-\Psi_0+\Psi_0^2)\Psi_1(1-\Psi_1) \frac{\varepsilon}{2} + \frac{(1 - \Psi_2^2)}{4} (1 - \varepsilon),$$

$$p_{11} = \Psi_0^2 \Psi_1^2 \frac{\varepsilon}{2} + \Psi_0(1-\Psi_0)\Psi_1(1-\Psi_1) \frac{\varepsilon}{2} + \frac{\Psi_2^2}{4} (1 - \varepsilon).$$

Thus, the probabilities based on Eq. (8) can be calculated. The optimal designs, defined as the designs that gave the highest power for fixed resources, were determined by an “exhaustive search” subject to certain gaps due to the amount of computation. The ranges of the parameters were: $c = 2500, 5000, 10,000,$ and $20,000$; $ASP = 50, 75, 100, 150,$ and 200 ; $\varepsilon = 0.25, 0.5, 0.75,$ and 1 ; $m = 700, 350, 233,$ and 175 , corresponding to 5 cM, 10 cM, 15 cM, and 20 cM maps for both two-stage design and one-stage design; r from 5 to $m - 1$ with an increment 5; and n_1 from 5 to $n - 5$ with increment 5. Due to these discrete increments, our exhaustive search was not truly exhaustive, but it can serve as a guideline in practice.

3. Simulation study.

In a simulation study, some of the assumptions can be relaxed. In analytic derivation, the X_{ij} 's are assumed independent except for the markers adjacent to the gene. In simulation, X_{ij} 's are simulated by Markov chains with transition probabilities derived from real recombination fraction using Haldane map function (Haldane, 1919). The

length of each chromosome is obtained from Schuler *et al.*, (1996). Also, in the analytic model the first-stage data were not used in the second-stage decision to locate the gene, but in simulation, both data were used. However, we also included the using the second-stage data only decision rule to check the analytic results. Similarly, we simulated X_{ij} 's under independence assumption to check the analytic results.

The programs were compiled using GCC version 2.7.2 on two PCs with Intel Pentium CPU (166MHz and 200MHz) in OS/2 environment. The random number generator for the simulation was adapted from *Numerical Recipes in C* (Press *et al.*, 1992).

4. Tables for Optimal Design and Criterion for Claiming Linkage

4.1 Tables for Optimal Designs

Based on analytic computation and simulation, the design that gives the highest probability of finding the marker adjacent to the gene (power) was identified. These optimal designs are given in Table 4 for searching a rare recessive gene and Table 5 for a rare dominant gene. In both tables, the [ASP] columns are the numbers of available affected-sib-pairs; the [epsilon] columns are the probabilities that the disease of an ASP is caused by the gene; the [m] columns are the numbers of marker loci used in the first stage; the [r] columns are the proposed numbers of loci to be chosen in Stage I for Stage II; the [n_1] columns are the numbers of ASPs used in first stage; the [F2] columns are the probabilities of locating the correct marker by the best two-stage design obtained by analytic formula; the [Indep] columns and the [P2M] columns are the probabilities obtained by simulation with independent models and Markov chain models assumption, respectively, without combining first- and second-stage data in the final decision; the [Comb.] columns are the simulated probabilities of the two-stage design with Markov chain models with first- and second-stage data combined. The [F1] columns are the probabilities of the best one-stage design, i.e.,

with optimal m and n subject to $mn \leq c$ obtained by analytic formula, the [P1M] columns are the probabilities obtained by simulation with Markov chain models. The [F2–F1] columns are the power increases of two-stage design over one-stage obtained by analytic formula, and the last columns of each table, [P2M–P1M], show the power increases of two-stage design over one-stage design by simulation. An asterisk is used when the increase is over 0.40.

4.2 Implications from the Tables

The greatest advantage of the two-stage design over the one-stage design occurs when the resources are limited but adequate, i.e., when the power is in the range of 0.7-0.9. Many of the power increases are over 100% (e.g. in the recessive case, $2,500 \leq c \leq 5,000$ for large ε and $c = 10,000$ for small ε , and in the dominant case, $5,000 \leq c \leq 10,000$ for large ε .) These are usually the required powers at the experimental design stage. When the resources are more than adequate with respect to the number of ASPs, the advantage of the two-stage design disappears for the obvious reason that enough information can be gathered by typing all the ASPs. Our two-stage design does not consider one-stage as a special case because the decision is made in the second stage. While this produces negative gains in a few occasions, all the negative gains are negligible from a practical viewpoint. Thus, we may conclude that the two-stage designs are better or at least as good as the one-stage designs.

As shown in Tables 4 and 5, more resources are required to locate a dominant disease gene than a recessive one. For example, a two-stage design with 50 ASPs and $c = 1,000$ can locate a recessive disease gene with power 0.72 when $\varepsilon = 1$, but 75 ASPs $c = 10,000$ are needed to achieve the same power for a dominant disease gene with $\varepsilon = 1$. Another point worth noting is that phenocopy can severely reduce the probability of finding the correct gene location. For example, to locate a recessive gene, with resource $c = 2500$, 50 ASPs, and $\varepsilon = 1$, the chance of finding it is higher

than 0.98. However, when $\varepsilon = 0.5$, even with the resource doubled ($c = 5,000$, and 100 ASPs), the chance of finding correct loci is only about 90%. Although both of them have about 50 ASPs with the disease caused by the same gene, the extra phenocopy ASPs reduce the probability considerably.

The Monte Carlo results indicate that the probabilities calculated from the analytic formulas and those from simulations are close. For example, 93 out of 100 designs were under 5% in relative error in the dominant cases, and 99 out of 100 under 5% in the recessive cases. Therefore, using independent assumption to approximate dependent marker loci is practically acceptable.

The Monte Carlo studies show that combining Stage I data with Stage II data does not confer any significant advantage. For the dominant cases, the probability of allocating the correct marker increase is less than 5% in 99 out of 100 designs, for 91 designs the increase is less than 3%, and for 81 designs the increases is less than 2%. For the recessive cases, for 97 out of 100 designs the increase is less than 6%, for 94 designs the increase is less than 5%, for 91 designs the increase is less than 4%, for 85 designs the increase is less than 3%, and for 80 designs the increase is less than 2%. There is an intuitive explanation for this. Those markers that passed the first stage have similar high scores. Therefore, combining them with second-stage data will not provide much more extra information.

4.3 Threshold for Claiming Linkage

Tables 4 and 5 give the optimal designs and the powers of finding the locus linked to the responsible gene when that gene exists. Thus, if we are sure that the disease is caused by one gene, then the power is the confidence level of claiming the correct gene location. However, if we are not sure whether such a gene exists, then the power does not provide guidance on the probability of making a false alarm. The usual requirement is that the LOD score in Stage II should be greater than a certain threshold in

order to claim linkage. Let $R = r$ markers and $N_2 = n_2$ ASPs be selected for Stage II and let $t_{\alpha, n_2, r}$ be the $100(1-\alpha)$ percentile of the unique maximum of r binomial(n_2 , 0.25) random variables. Then a linkage can be claimed if ${}_2S_I > t_{\alpha, n_2, r}$, where I is one of the markers chosen in Stage II. This is based on the following derivation:

$$\begin{aligned}
& P\{\text{linkage claim is incorrect}\} \\
& \leq P\{{}_2S_I > t_{\alpha, n_2, r} \mid \text{no gene is responsible}\}P\{\text{no gene is responsible}\} \\
& \quad + P\{I \text{ is wrong and } {}_2S_I > t_{\alpha, n_2, r} \mid \text{a gene is responsible}\} \cdot \\
& \quad P\{\text{a gene is responsible}\}.
\end{aligned}$$

Though the prior $P\{\text{No gene is responsible}\}$ is usually unknown, it can be shown (see Appendix II) that $P\{I \text{ is wrong and } {}_2S_I > t_{\alpha, n_2, r} \mid \text{a gene is responsible}\} \leq P\{{}_2S_I > t_{\alpha, n_2, r} \mid \text{no gene is responsible}\}$. Thus,

$$\begin{aligned}
& P\{\text{linkage claim is incorrect}\} \\
& \leq P\{{}_2S_I > t_{\alpha, n_2, r} \mid \text{no gene is responsible}\} \\
& = \sum_{\substack{\text{all possible choices} \\ \text{from Stage I}}} P\{{}_2S_I > t_{\alpha, n_2, r} \mid \text{Stage I resulted in } R = r, N_2 = n_2\} \cdot \\
& \quad P\{R = r, N_2 = n_2 \text{ from Stage I}\} \\
& = \alpha \sum_{\text{all possibilities}} P\{R = r, N_2 = n_2 \text{ from Stage I}\} \\
& \leq \alpha.
\end{aligned}$$

Due to the discreteness of the binomial distribution, the exact level for α may not be obtainable for each (r, n_2) combination. Thus, for fixed α , the threshold is chosen so that $P\{{}_2S_I > t_{\alpha, n_2, r}\} \leq \alpha$. In other words, if we use $t_{\alpha, n_2, r}$ as the threshold in Stage II, then ${}_2S_I > t_{\alpha, n_2, r}$ enables us to claim that there is a linkage at loci I with a false alarm rate $\leq \alpha$.

5. CONCLUDING REMARKS

This paper has demonstrated the advantage of two stage genome-wide search using the ASP method. In most cases, it outperforms the one-stage design. Tables are provided as guidelines on how resources should be located in the two stages. Also, analytic study and simulation were used to cross-check one another. We expect that future study on genome-wide search, such as under the conditions of more than two affected sibling, affected relatives, or extreme quantitative trait locus allocation, will consider the two-stage option.

REFERENCES

- Botstein, D., White, R. L., Skolnick, M. H., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*. **32**:314-331.
- Brown, D. L., Gorin, M. B., and Weeks, D. E. (1994). Efficient strategies for genomic searching using the affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics*. **54**:544-552.
- Conneally, P. M., and Rivas, M. L. (1980). Linkage analysis in man. *Advances in Human Genetic*, New York: Plenum Press. 209-266.
- Darvasi, A and Soller, M. (1994). Optimum spacing of genetic markers for determining linkage between marker loci and quantitative trait loci. *Theoretical and Applied Genetics*. **89**:351-357.
- Davies, J. L., Kawaguchi, Y., Bennett, S. T., Copeman, J. B., Cordell, H. J., Pritchard, L. E., Reed, P. W., Gough, S. C., Jenkins, S. C., Palmer, S. M., Balfour, K. M., Rowe, B. R., Farrall, M., Barnett, A. H., Baln, S. C., Tood, J. A. (1994). A genome-wide search for human type 1 diabetes susceptibility genes. *Nature*. **371**:130-136.

- Day, N. E. and Simons, M. J. (1976). Disease susceptibility genes—their identification by multiple case family studies. *Tissue Antigens*. **8**:109-119.
- Elston, R. (1992a). Designs for the global search of the human genome by linkage analysis. In: *Proceedings of the 16th International Biometric conference*: Hamilton, New Zealand, December 7-11. Ruakura Agricultural Center, Hamilton, New Zealand, pp 39-51.
- Elston, R. (1992b). P-values, power and pitfalls in the linkage analysis of psychiatric disorders. In: *Proceeding of the Annual Meeting of the American Psychological Association*, Gershon, E. S., Cloninger, C. R., and Barrett, J. E.(eds.); published as *Genetics Approaches to Mental Disorders*. Washington, D.C.: American Psychiatric Press.
- Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics*. **8**:299-309.
- Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*. **2**:3-19.
- Lander, E. S. and Botstein, D. (1986a). Strategies for heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*. **83**:7353-7357.
- Lander, E. S. and Botstein D. (1986b). Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map. *Cold Spring Harbor Symposia on Quantitative Biology*. **51**:49-62.
- Lander, E. S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*. **121**:185-199.
- Luo, D. F., Bui, M. M., Muir, A., Maclaren, N. K., Thomson, G., and She, J. X. (1995). Affected-sib-pair mapping of a novel susceptibility gene to insulin-dependent diabetes mellitus (IDDM8) on chromosome 6q25-q27. *American Journal of Human Genetics*. **57**:911-919.

- Penrose, L. S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics*. **6**:133-138.
- Penrose, L. S. (1950). Data for the study of linkage in man: red hair and the ABO locus. *Annals of Eugenics*. **15**:243-247.
- Penrose, L. S. (1953). The general purpose sib-pair linkage test. *Annals of Eugenics*. **18**:120-124.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical recipes in C*, second edition. New York: Cambridge University Press.
- Schuler G.D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B.B., Butler, A., Castle, A.B., Chiannikulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Hudson, T. J., et al. (1996). A gene map of the human genome. *Science*. **274**(5287):540-546
- Shah, S. and Green, J.R. (1994). Disease susceptibility genes and the sib-pair method: a review of recent methodology. *Annals of Human Genetics*. **58**:381-395.

Table 4: Optimal resource allocation in two-stage genome search for a rare recessive gene.

Resource c=1000															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	175	5	5	0.050	0.052	0.052	0.054	175	5	0.015	0.016	0.017	0.034	0.036
50	0.5	175	5	5	0.190	0.195	0.186	0.201	175	5	0.049	0.053	0.051	0.140	0.136
50	0.75	175	5	5	0.438	0.439	0.435	0.474	175	5	0.125	0.122	0.125	0.313	0.311
50	1	175	5	5	0.722	0.718	0.716	0.778	175	5	0.271	0.279	0.263	0.451*	0.453*
75	0.25	175	5	5	0.050	0.050	0.050	0.053	175	5	0.015	0.014	0.016	0.034	0.034
75	0.5	175	5	5	0.190	0.195	0.191	0.211	175	5	0.049	0.053	0.046	0.140	0.145
75	0.75	175	5	5	0.438	0.447	0.437	0.479	175	5	0.125	0.117	0.124	0.313	0.313
75	1	175	5	5	0.722	0.727	0.720	0.776	175	5	0.271	0.270	0.266	0.451*	0.453*
100	0.25	175	5	5	0.050	0.050	0.048	0.050	175	5	0.015	0.017	0.017	0.034	0.030
100	0.5	350	5	5	0.190	0.189	0.190	0.211	175	5	0.049	0.051	0.049	0.140	0.141
100	0.75	175	5	5	0.438	0.434	0.431	0.470	175	5	0.125	0.124	0.122	0.313	0.310
100	1	175	5	5	0.722	0.724	0.717	0.782	175	5	0.271	0.268	0.260	0.451*	0.457*
150	0.25	175	5	5	0.050	0.049	0.050	0.054	175	5	0.015	0.014	0.015	0.034	0.034
150	0.5	175	5	5	0.190	0.187	0.192	0.211	175	5	0.049	0.052	0.048	0.140	0.144
150	0.75	175	5	5	0.438	0.435	0.442	0.484	175	5	0.125	0.125	0.123	0.313	0.319
150	1	175	5	5	0.722	0.725	0.721	0.777	175	5	0.271	0.266	0.263	0.451*	0.458*
200	0.25	175	5	5	0.050	0.047	0.053	0.054	175	5	0.015	0.014	0.017	0.034	0.037
200	0.5	175	5	5	0.190	0.191	0.194	0.208	175	5	0.049	0.048	0.048	0.140	0.147
200	0.75	175	5	5	0.438	0.443	0.429	0.474	175	5	0.125	0.125	0.120	0.313	0.310
200	1	175	5	5	0.722	0.724	0.710	0.771	175	5	0.271	0.269	0.264	0.451*	0.446*

Resource c=2500															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	233	10	5	0.189	0.192	0.174	0.174	233	14	0.056	0.058	0.053	0.132	0.121
50	0.5	175	10	10	0.639	0.641	0.628	0.639	175	14	0.240	0.241	0.242	0.399	0.385
50	0.75	175	10	10	0.922	0.926	0.914	0.920	175	14	0.570	0.560	0.557	0.352	0.357
50	1	350	15	5	0.996	0.997	0.979	0.987	350	14	0.877	0.872	0.863	0.119	0.116
75	0.25	175	10	5	0.222	0.224	0.212	0.213	175	14	0.056	0.053	0.063	0.166	0.150
75	0.5	175	10	10	0.667	0.670	0.652	0.665	175	14	0.240	0.239	0.242	0.427*	0.410*
75	0.75	175	10	10	0.925	0.922	0.923	0.928	175	14	0.570	0.572	0.550	0.355	0.373
75	1	350	15	5	0.996	0.995	0.979	0.986	350	14	0.877	0.875	0.862	0.119	0.116
100	0.25	175	10	5	0.246	0.251	0.236	0.238	175	14	0.056	0.059	0.054	0.189	0.182
100	0.5	175	10	10	0.667	0.669	0.654	0.668	175	14	0.240	0.240	0.237	0.428*	0.417*
100	0.75	175	10	10	0.925	0.921	0.918	0.924	175	14	0.570	0.577	0.557	0.355	0.361
100	1	350	15	5	0.996	0.995	0.977	0.985	350	14	0.877	0.878	0.861	0.119	0.115
150	0.25	175	10	5	0.251	0.252	0.245	0.246	175	14	0.056	0.055	0.059	0.195	0.186
150	0.5	175	10	10	0.667	0.673	0.661	0.674	175	14	0.240	0.234	0.232	0.428*	0.429*
150	0.75	175	10	10	0.925	0.925	0.922	0.927	175	14	0.570	0.577	0.557	0.355	0.365
150	1	350	15	5	0.996	0.994	0.978	0.987	350	14	0.877	0.872	0.858	0.119	0.120
200	0.25	175	10	5	0.251	0.254	0.238	0.238	175	14	0.056	0.056	0.057	0.195	0.182
200	0.5	175	10	10	0.667	0.664	0.658	0.671	175	14	0.240	0.240	0.237	0.428*	0.421*
200	0.75	175	10	10	0.925	0.924	0.922	0.928	175	14	0.570	0.560	0.552	0.355	0.371
200	1	350	15	5	0.996	0.995	0.979	0.987	350	14	0.877	0.875	0.860	0.119	0.118

Resource c=5000															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	233	30	10	0.248	0.248	0.233	0.253	233	28	0.127	0.128	0.124	0.121	0.109
50	0.5	233	35	10	0.827	0.827	0.798	0.832	233	28	0.539	0.539	0.534	0.288	0.263
50	0.75	233	20	15	0.993	0.993	0.985	0.993	233	28	0.910	0.907	0.898	0.083	0.087
50	1	233	20	15	1.000	1.000	0.999	1.000	233	28	0.996	0.995	0.994	0.004	0.005
75	0.25	175	25	10	0.324	0.320	0.324	0.335	175	28	0.127	0.128	0.126	0.198	0.199
75	0.5	175	15	20	0.881	0.882	0.869	0.892	175	28	0.539	0.538	0.526	0.342	0.344
75	0.75	233	20	15	0.995	0.995	0.990	0.994	233	28	0.910	0.902	0.894	0.085	0.096
75	1	233	20	15	1.000	1.000	1.000	1.000	233	28	0.996	0.996	0.993	0.004	0.007
100	0.25	175	15	15	0.391	0.387	0.376	0.386	175	28	0.127	0.130	0.130	0.264	0.246
100	0.5	175	15	20	0.903	0.899	0.897	0.910	175	28	0.539	0.539	0.525	0.364	0.372
100	0.75	233	20	15	0.995	0.996	0.989	0.993	233	28	0.910	0.908	0.898	0.085	0.091
100	1	233	20	15	1.000	1.000	1.000	1.000	233	28	0.996	0.996	0.992	0.004	0.008
150	0.25	175	15	15	0.420	0.420	0.414	0.423	175	28	0.127	0.121	0.126	0.293	0.288
150	0.5	175	15	20	0.904	0.902	0.894	0.905	175	28	0.539	0.541	0.535	0.365	0.359
150	0.75	233	20	15	0.995	0.993	0.991	0.994	233	28	0.910	0.902	0.892	0.085	0.099
150	1	233	20	15	1.000	1.000	0.999	1.000	233	28	0.996	0.994	0.993	0.004	0.006
200	0.25	175	10	15	0.425	0.422	0.414	0.417	175	28	0.127	0.125	0.126	0.298	0.288
200	0.5	175	15	20	0.904	0.901	0.896	0.907	175	28	0.539	0.539	0.524	0.365	0.372
200	0.75	233	20	15	0.995	0.994	0.991	0.994	233	28	0.910	0.912	0.900	0.085	0.091
200	1	233	20	15	1.000	1.000	1.000	1.000	233	28	0.996	0.996	0.993	0.004	0.007

Resource c=10,000															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	350	75	5	0.281	0.285	0.272	0.286	350	50	0.253	0.256	0.246	0.028	0.026
50	0.5	350	80	10	0.892	0.890	0.850	0.886	350	50	0.835	0.839	0.825	0.057	0.024
50	0.75	350	100	10	0.999	0.998	0.986	0.994	350	42	0.996	0.996	0.992	0.003	-0.006
50	1	350	75	15	1.000	1.000	0.999	1.000	350	42	1.000	1.000	1.000	0.000	-0.001
75	0.25	350	50	10	0.428	0.428	0.408	0.421	350	57	0.294	0.300	0.292	0.134	0.116
75	0.5	350	30	20	0.965	0.964	0.942	0.958	350	57	0.885	0.886	0.876	0.080	0.066
75	0.75	350	15	25	1.000	1.000	0.996	1.000	350	57	0.998	0.998	0.996	0.002	-0.001
75	1	350	15	25	1.000	1.000	1.000	1.000	350	57	1.000	1.000	1.000	0.000	-0.000
100	0.25	350	35	15	0.526	0.529	0.485	0.502	350	57	0.294	0.292	0.287	0.232	0.198
100	0.5	233	25	30	0.984	0.984	0.975	0.987	233	57	0.885	0.883	0.872	0.099	0.104
100	0.75	233	15	35	1.000	1.000	1.000	1.000	233	57	0.998	0.998	0.996	0.002	0.004
100	1	233	5	35	1.000	1.000	1.000	1.000	233	57	1.000	1.000	1.000	0.000	0.000
150	0.25	233	25	25	0.640	0.644	0.614	0.633	233	57	0.294	0.293	0.283	0.346	0.332
150	0.5	175	20	40	0.991	0.990	0.989	0.992	175	57	0.885	0.882	0.874	0.106	0.115
150	0.75	175	15	45	1.000	1.000	1.000	1.000	175	57	0.998	0.998	0.997	0.002	0.003
150	1	175	5	45	1.000	1.000	1.000	1.000	175	57	1.000	1.000	1.000	0.000	0.000
200	0.25	175	20	35	0.683	0.687	0.669	0.686	175	57	0.294	0.297	0.284	0.389	0.385
200	0.5	175	20	40	0.992	0.992	0.989	0.991	175	57	0.885	0.889	0.873	0.107	0.115
200	0.75	175	15	45	1.000	1.000	1.000	1.000	175	57	0.998	0.997	0.995	0.002	0.005
200	1	175	5	45	1.000	1.000	1.000	1.000	175	57	1.000	1.000	1.000	0.000	0.000

Resource c=20,000															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	350	145	5	0.291	0.294	0.270	0.290	350	50	0.323	0.335	0.305	-0.032	-0.035
50	0.5	700	155	10	0.925	0.923	0.800	0.859	700	50	0.944	0.945	0.910	-0.020	-0.110
50	0.75	700	210	10	1.000	1.000	0.964	0.980	700	50	1.000	1.000	0.996	-0.000	-0.032
50	1	700	5	20	1.000	1.000	0.987	0.999	700	50	1.000	1.000	1.000	0.000	-0.013
75	0.25	350	150	5	0.467	0.461	0.447	0.461	350	75	0.461	0.461	0.445	0.006	0.002
75	0.5	350	200	5	0.983	0.984	0.964	0.969	350	75	0.983	0.983	0.973	0.001	-0.009
75	0.75	700	30	25	1.000	1.000	0.981	0.996	700	75	1.000	1.000	1.000	0.000	-0.019
75	1	350	5	30	1.000	1.000	1.000	1.000	350	57	1.000	1.000	1.000	0.000	-0.000
100	0.25	350	115	10	0.615	0.616	0.571	0.592	350	100	0.533	0.529	0.529	0.082	0.043
100	0.5	350	105	20	0.996	0.996	0.986	0.993	350	85	0.992	0.991	0.988	0.005	-0.002
100	0.75	350	5	45	1.000	1.000	0.999	1.000	350	85	1.000	1.000	1.000	0.000	-0.001
100	1	233	5	40	1.000	1.000	1.000	1.000	233	85	1.000	1.000	1.000	0.000	0.000
150	0.25	350	50	30	0.786	0.782	0.744	0.776	350	114	0.600	0.596	0.587	0.186	0.157
150	0.5	350	30	45	0.999	0.999	0.993	0.999	350	114	0.996	0.996	0.993	0.004	0.000
150	0.75	233	5	60	1.000	1.000	1.000	1.000	233	114	1.000	1.000	1.000	0.000	-0.000
150	1	175	5	45	1.000	1.000	1.000	1.000	175	114	1.000	1.000	1.000	0.000	0.000
200	0.25	233	35	55	0.853	0.847	0.841	0.877	233	114	0.600	0.596	0.590	0.253	0.252
200	0.5	233	25	65	1.000	1.000	0.999	1.000	233	114	0.996	0.995	0.993	0.004	0.006
200	0.75	175	5	70	1.000	1.000	1.000	1.000	175	114	1.000	1.000	1.000	0.000	0.000
200	1	175	5	45	1.000	1.000	1.000	1.000	175	114	1.000	1.000	1.000	0.000	0.000

Table 5: Optimal resource allocation in two-stage genome search for a rare dominant gene.

Resource c=1000															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	175	5	5	0.015	0.016	0.015	0.015	175	5	0.006	0.006	0.007	0.008	0.008
50	0.5	175	5	5	0.032	0.030	0.029	0.032	175	5	0.011	0.010	0.012	0.021	0.017
50	0.75	175	5	5	0.061	0.057	0.059	0.061	175	5	0.018	0.016	0.019	0.043	0.040
50	1	175	5	5	0.106	0.105	0.109	0.118	175	5	0.029	0.028	0.031	0.077	0.078
75	0.25	175	5	5	0.015	0.016	0.019	0.019	175	5	0.006	0.008	0.006	0.008	0.014
75	0.5	175	5	5	0.032	0.028	0.031	0.032	175	5	0.011	0.010	0.010	0.021	0.021
75	0.75	175	5	5	0.061	0.061	0.057	0.062	175	5	0.018	0.018	0.021	0.043	0.037
75	1	175	5	5	0.106	0.102	0.104	0.114	175	5	0.029	0.029	0.031	0.077	0.073
100	0.25	175	5	5	0.015	0.016	0.014	0.016	175	5	0.006	0.005	0.006	0.008	0.008
100	0.5	175	5	5	0.032	0.033	0.033	0.036	175	5	0.011	0.011	0.010	0.021	0.022
100	0.75	175	5	5	0.061	0.062	0.062	0.066	175	5	0.018	0.018	0.020	0.043	0.043
100	1	175	5	5	0.106	0.108	0.109	0.117	175	5	0.029	0.030	0.029	0.077	0.080
150	0.25	175	5	5	0.015	0.016	0.014	0.016	175	5	0.006	0.008	0.008	0.008	0.006
150	0.5	175	5	5	0.032	0.032	0.034	0.034	175	5	0.011	0.010	0.011	0.021	0.022
150	0.75	175	5	5	0.061	0.057	0.062	0.067	175	5	0.018	0.019	0.020	0.043	0.042
150	1	175	5	5	0.106	0.110	0.108	0.117	175	5	0.029	0.028	0.031	0.077	0.077
200	0.25	175	5	5	0.015	0.014	0.016	0.018	175	5	0.006	0.005	0.007	0.008	0.009
200	0.5	175	5	5	0.032	0.032	0.036	0.038	175	5	0.011	0.011	0.011	0.021	0.025
200	0.75	175	5	5	0.061	0.062	0.060	0.065	175	5	0.018	0.019	0.018	0.043	0.041
200	1	175	5	5	0.106	0.106	0.109	0.119	175	5	0.029	0.030	0.032	0.077	0.077

Resource c=2500															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	175	5	10	0.035	0.035	0.036	0.036	175	14	0.015	0.014	0.016	0.020	0.021
50	0.5	175	10	10	0.104	0.105	0.096	0.099	175	14	0.035	0.033	0.034	0.069	0.062
50	0.75	175	10	10	0.232	0.230	0.227	0.237	175	14	0.071	0.075	0.073	0.161	0.154
50	1	175	10	10	0.406	0.407	0.389	0.402	175	14	0.128	0.125	0.129	0.278	0.260
75	0.25	175	10	5	0.043	0.042	0.045	0.045	175	14	0.015	0.013	0.014	0.028	0.031
75	0.5	175	10	5	0.133	0.131	0.129	0.132	175	14	0.035	0.037	0.034	0.098	0.096
75	0.75	175	10	5	0.275	0.280	0.261	0.263	175	14	0.071	0.073	0.073	0.204	0.188
75	1	175	10	10	0.445	0.448	0.421	0.434	175	14	0.128	0.135	0.129	0.317	0.292
100	0.25	175	10	5	0.048	0.051	0.048	0.049	175	14	0.015	0.015	0.017	0.033	0.031
100	0.5	175	10	5	0.150	0.148	0.147	0.148	175	14	0.035	0.035	0.036	0.115	0.110
100	0.75	175	10	5	0.300	0.302	0.291	0.293	175	14	0.071	0.070	0.072	0.229	0.219
100	1	175	10	5	0.452	0.450	0.444	0.445	175	14	0.128	0.131	0.126	0.324	0.317
150	0.25	175	5	5	0.050	0.049	0.051	0.051	175	14	0.015	0.016	0.014	0.035	0.037
150	0.5	175	10	5	0.155	0.156	0.147	0.148	175	14	0.035	0.037	0.034	0.120	0.112
150	0.75	175	10	5	0.305	0.317	0.293	0.295	175	14	0.071	0.069	0.072	0.234	0.221
150	1	175	10	5	0.455	0.456	0.436	0.439	175	14	0.128	0.130	0.128	0.327	0.309
200	0.25	175	5	5	0.051	0.051	0.052	0.052	175	14	0.015	0.015	0.014	0.036	0.038
200	0.5	175	10	5	0.155	0.154	0.152	0.152	175	14	0.035	0.038	0.031	0.120	0.121
200	0.75	175	10	5	0.305	0.303	0.296	0.296	175	14	0.071	0.065	0.072	0.235	0.224
200	1	175	10	5	0.455	0.457	0.444	0.446	175	14	0.128	0.132	0.121	0.327	0.323

Resource c=5000															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	175	5	25	0.038	0.040	0.036	0.039	175	28	0.025	0.026	0.026	0.013	0.011
50	0.5	233	30	10	0.125	0.129	0.118	0.128	175	28	0.072	0.074	0.073	0.053	0.045
50	0.75	233	30	10	0.305	0.302	0.294	0.318	175	28	0.164	0.170	0.162	0.141	0.132
50	1	233	35	10	0.541	0.549	0.503	0.538	175	28	0.306	0.300	0.296	0.235	0.207
75	0.25	175	25	10	0.051	0.055	0.049	0.053	175	28	0.025	0.023	0.025	0.026	0.025
75	0.5	175	25	10	0.182	0.188	0.182	0.189	175	28	0.072	0.070	0.071	0.110	0.112
75	0.75	175	25	10	0.409	0.412	0.400	0.412	175	28	0.164	0.160	0.157	0.245	0.242
75	1	175	15	20	0.656	0.652	0.637	0.668	175	28	0.306	0.302	0.301	0.350	0.336
100	0.25	175	15	15	0.062	0.061	0.062	0.064	175	28	0.025	0.024	0.024	0.037	0.037
100	0.5	175	15	15	0.226	0.222	0.211	0.216	175	28	0.072	0.073	0.072	0.154	0.139
100	0.75	175	15	15	0.483	0.478	0.461	0.469	175	28	0.164	0.160	0.160	0.319	0.301
100	1	175	20	15	0.722	0.725	0.689	0.707	175	28	0.306	0.298	0.296	0.416*	0.393
150	0.25	175	10	15	0.073	0.073	0.071	0.073	175	28	0.025	0.026	0.022	0.048	0.048
150	0.5	175	10	15	0.252	0.253	0.243	0.245	175	28	0.072	0.069	0.066	0.180	0.177
150	0.75	175	15	15	0.512	0.512	0.504	0.516	175	28	0.164	0.165	0.153	0.348	0.351
150	1	175	15	15	0.735	0.728	0.708	0.719	175	28	0.306	0.315	0.301	0.429*	0.408*
200	0.25	175	10	10	0.083	0.084	0.080	0.081	175	28	0.025	0.025	0.026	0.058	0.055
200	0.5	175	10	15	0.270	0.269	0.273	0.274	175	28	0.072	0.070	0.070	0.198	0.204
200	0.75	175	15	15	0.513	0.520	0.504	0.512	175	28	0.164	0.165	0.157	0.349	0.347
200	1	175	15	15	0.736	0.735	0.716	0.723	175	28	0.306	0.313	0.291	0.430*	0.425*

Resource c=10,000															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	175	5	25	0.038	0.040	0.039	0.039	175	50	0.040	0.039	0.037	-0.002	0.001
50	0.5	233	95	5	0.131	0.130	0.125	0.133	175	50	0.138	0.140	0.137	-0.007	-0.012
50	0.75	350	75	5	0.328	0.332	0.295	0.310	175	50	0.327	0.325	0.317	0.001	-0.021
50	1	350	75	5	0.585	0.588	0.521	0.540	175	50	0.570	0.571	0.549	0.016	-0.028
75	0.25	175	65	5	0.053	0.053	0.052	0.054	175	57	0.045	0.045	0.044	0.009	0.008
75	0.5	233	65	10	0.209	0.211	0.198	0.210	175	57	0.161	0.160	0.156	0.049	0.042
75	0.75	233	70	10	0.493	0.495	0.461	0.494	175	57	0.378	0.375	0.361	0.114	0.100
75	1	233	75	10	0.769	0.762	0.727	0.761	175	57	0.639	0.640	0.617	0.131	0.110
100	0.25	175	50	10	0.069	0.067	0.070	0.073	175	57	0.045	0.044	0.047	0.025	0.023
100	0.5	233	45	15	0.282	0.281	0.267	0.286	175	57	0.161	0.161	0.159	0.122	0.109
100	0.75	233	50	15	0.615	0.617	0.577	0.598	175	57	0.378	0.375	0.374	0.237	0.203
100	1	233	50	15	0.862	0.862	0.819	0.841	175	57	0.639	0.634	0.610	0.223	0.209
150	0.25	175	25	30	0.098	0.099	0.098	0.105	175	57	0.045	0.043	0.044	0.053	0.054
150	0.5	175	25	30	0.387	0.384	0.376	0.395	175	57	0.161	0.161	0.152	0.227	0.224
150	0.75	175	25	30	0.737	0.735	0.722	0.743	175	57	0.378	0.368	0.368	0.359	0.353
150	1	175	20	40	0.926	0.925	0.909	0.930	175	57	0.639	0.636	0.622	0.287	0.286
200	0.25	175	20	25	0.119	0.119	0.118	0.121	175	57	0.045	0.045	0.048	0.074	0.070
200	0.5	175	15	35	0.446	0.439	0.434	0.444	175	57	0.161	0.169	0.160	0.286	0.274
200	0.75	175	20	35	0.781	0.783	0.760	0.776	175	57	0.378	0.381	0.372	0.403*	0.388
200	1	175	20	35	0.937	0.937	0.925	0.933	175	57	0.639	0.643	0.613	0.298	0.313

Resource c=20,000															
Parameter		Two-stage design & power							One-stage design & power					Improv.	
ASP	epsilon	m	r	n ₁	F2	Indep	P2M	Comb.	m	n ₁	F1	Indep	P1M	F2-F1	P2M-P1M
50	0.25	175	5	25	0.038	0.038	0.038	0.039	175	50	0.040	0.040	0.043	-0.002	-0.005
50	0.5	233	95	5	0.131	0.129	0.125	0.134	233	50	0.144	0.142	0.138	-0.013	-0.013
50	0.75	350	150	5	0.340	0.336	0.311	0.343	350	50	0.377	0.379	0.331	-0.036	-0.020
50	1	350	170	5	0.615	0.618	0.549	0.585	350	50	0.664	0.665	0.585	-0.049	-0.036
75	0.25	175	80	5	0.054	0.052	0.051	0.055	175	75	0.057	0.057	0.055	-0.003	-0.004
75	0.5	233	125	5	0.220	0.219	0.213	0.224	233	75	0.237	0.231	0.217	-0.017	-0.004
75	0.75	350	150	5	0.534	0.540	0.475	0.496	233	75	0.552	0.556	0.521	-0.018	-0.046
75	1	350	165	5	0.822	0.822	0.733	0.755	233	75	0.831	0.830	0.787	-0.009	-0.054
100	0.25	175	90	5	0.072	0.072	0.073	0.072	175	100	0.075	0.072	0.072	-0.004	0.001
100	0.5	233	135	5	0.312	0.309	0.289	0.302	175	100	0.306	0.302	0.292	0.006	-0.004
100	0.75	350	120	10	0.683	0.676	0.611	0.633	175	100	0.650	0.643	0.632	0.033	-0.021
100	1	350	100	20	0.915	0.919	0.823	0.866	175	100	0.894	0.895	0.877	0.021	-0.054
150	0.25	175	80	5	0.105	0.104	0.105	0.106	175	114	0.086	0.085	0.087	0.020	0.019
150	0.5	350	50	30	0.464	0.461	0.417	0.443	175	114	0.353	0.351	0.340	0.111	0.078
150	0.75	350	50	30	0.840	0.839	0.761	0.790	175	114	0.717	0.716	0.697	0.123	0.064
150	1	233	80	25	0.974	0.973	0.946	0.959	175	114	0.932	0.930	0.913	0.042	0.033
200	0.25	233	45	35	0.139	0.140	0.132	0.144	175	114	0.086	0.083	0.083	0.053	0.049
200	0.5	233	40	45	0.579	0.583	0.546	0.581	175	114	0.353	0.354	0.343	0.226	0.203
200	0.75	233	35	55	0.912	0.907	0.867	0.905	175	114	0.717	0.717	0.705	0.195	0.162
200	1	233	35	55	0.991	0.990	0.972	0.985	175	114	0.932	0.931	0.914	0.059	0.058

APPENDIX I

Derivation of the Probability that the Disease Gene Is Found

The derivation starts with (3) and (4). Eq. (3) can be written as

$$\begin{aligned}
& \frac{22}{m} P\{\text{Marker 0 is chosen at the end of Stage II} \mid \text{gene is at the end}\} \\
&= \frac{22}{m} P\{{}_1S_0 \text{ passes Stage I and } {}_2S_0 \text{ is the highest in Stage II}\} \\
&= \frac{22}{m} \sum_{k=0}^{n_1} P\{{}_1S_0 \text{ passes Stage I, } {}_2S_0 \text{ is the highest in Stage II, } {}_1S_{(r-1)} = k\} \\
&= \frac{22}{m} \sum_{k=0}^{n_1} \{P\{{}_1S_0 \geq {}_1S_{(r-1)}, {}_1S_{(r-1)} = k, {}_2S_0 \text{ is the highest in Stage II}\} \quad (10)
\end{aligned}$$

$$+ P\{k = {}_1S_{(r-1)} > {}_1S_0 \geq {}_1S_{(r)}, {}_2S_0 \text{ is the highest in Stage II}\}. \quad (11)$$

Eq. (4) can be written as

$$\begin{aligned}
& \frac{m-22}{m} P\{\text{Marker 0 or 1 is chosen at the end of Stage II} \mid \text{gene is not at the end}\} \\
&= \frac{m-22}{m} \{ \\
& \quad P\{{}_1S_0 \text{ passes Stage I and } {}_1S_1 \text{ does not, } {}_2S_0 \text{ is the highest in Stage II}\} \quad (12)
\end{aligned}$$

$$+ P\{{}_1S_1 \text{ passes Stage I and } {}_1S_0 \text{ does not, } {}_2S_1 \text{ is the highest in Stage II}\} \quad (13)$$

$$+ P\{{}_1S_0 \text{ and } {}_1S_1 \text{ pass Stage I and } {}_2S_0 \text{ or } {}_2S_1 \text{ is the highest in Stage II}\}. \quad (14)$$

Since the gene is assumed to be at the center of two markers, Eq. (12) is equal to Eq. (13) and they are

$$\begin{aligned}
&= P\{{}_1S_0 \text{ passes Stage I and } {}_1S_1 \text{ does not, } {}_2S_0 \text{ is the highest in Stage II}\} \\
&= \sum_{k=1}^{n_1} P\{{}_1S_0 \text{ passes Stage I and } {}_1S_1 \text{ does not, } {}_2S_0 \text{ is the highest in Stage II, } {}_1S_{(r-1)} = k\} \\
&= \sum_{k=1}^{n_1} \{P\{{}_1S_0 \geq {}_1S_{(r-1)} = k > {}_1S_1, {}_2S_0 \text{ is the highest in Stage II}\} \quad (15)
\end{aligned}$$

$$+ P\{{}_1S_{(r-1)} = k > {}_1S_0 \geq {}_1S_{(r)}, {}_1S_0 > {}_1S_1, {}_2S_0 \text{ is the highest in Stage II}\}. \quad (16)$$

Thus, Eq. (14) is equivalent to

$$P\{ {}_1S_0 \text{ and } {}_1S_1 \text{ pass Stage I, gene is found in Stage II} \}$$

$$= \sum_{k=1}^{n_1} \{ P\{ {}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, \text{ gene is found in Stage II} \} \} \quad (17)$$

$$+ P\{ {}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 \geq {}_1S_{(r-1)}, \text{ gene is found in Stage II} \}. \quad (18)$$

$$+ P\{ {}_1S_1 \geq {}_1S_{(r-2)} = k > {}_1S_0 \geq {}_1S_{(r-1)}, \text{ gene is found in Stage II} \} \quad (19)$$

$$+ P\{ {}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 \geq {}_1S_{(r-1)}, \text{ gene is found in Stage II} \} \quad (20)$$

$$+ P\{ {}_1S_{(r-2)} = k > {}_1S_1 > {}_1S_0 \geq {}_1S_{(r-1)}, \text{ gene is found in Stage II} \} \quad (21)$$

$$+ P\{ {}_1S_{(r-2)} = k > {}_1S_1 = {}_1S_0 \geq {}_1S_{(r-1)}, \text{ gene is found in Stage II} \} \quad (22)$$

$$+ P\{ {}_1S_{(r-1)} = k > {}_1S_1 = {}_1S_0 \geq {}_1S_{(r)}, \text{ gene is found in Stage II} \}. \quad (23)$$

Again, because the gene is at the center of two markers, Eq. (18) is equal to (19), and (20) is equal to (21). We use (17) as an example to demonstrate how to calculate the probabilities (17) – (23):

$$P\{ {}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, \text{ gene is found in Stage II} \}$$

$$= \underbrace{P\{ {}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, {}_2S_0 \text{ or } {}_2S_1 \text{ is uniquely highest in Stage II} \}}_{(A)}$$

$$+ \underbrace{P\{ {}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, {}_2S_0 = {}_2S_1 \text{ is the highest} \}}_{(B)}, \text{ where,}$$

$$(A)$$

$$= \sum_{s+t \geq (r-2), s < (r-2)} \sum_t P\{ {}_1S_0 \geq k, {}_1S_1 \geq k, {}_1S_{(r-2)} = k, s \text{ } {}_1S_i' s > k, \\ t \text{ } {}_1S_i' s = k, (m-2-s-t) \text{ } {}_1S_i' s < k, {}_2S_0 \text{ or } {}_2S_1 \text{ is uniquely highest in Stage II} \}$$

$$= \sum_{s=0}^{r-3} \sum_{t=r-2-s}^{m-2-s} \left\{ P\{ {}_1S_0 \geq k, {}_1S_1 \geq k \} \frac{(m-2)!}{s! t! ((m-2)-s-t)!} \right.$$

$$P\{ {}_1S_i > k \}^s P\{ {}_1S_i = k \}^t P\{ {}_1S_i < k \}^{m-2-s-t}$$

$$\begin{aligned}
& \left[P\{ {}_2S_0 > \max_{i \neq 0,1} {}_2S_i, {}_2S_0 > {}_2S_1 \} + P\{ {}_2S_1 > \max_{i \neq 0,1} {}_2S_i, {}_2S_1 > {}_2S_0 \} \right] \Bigg\} \\
& \text{(assuming the gene is at the center of two markers)} \\
= & \sum_{s=0}^{r-3} \sum_{t=r-2-s}^{m-2-s} \left\{ P\{ {}_1S_0 \geq k, {}_1S_1 \geq k \} \frac{(m-2)!}{s! t! ((m-2) - s - t)!} \right. \\
& P\{ {}_1S_i > k \}^s P\{ {}_1S_i = k \}^t P\{ {}_1S_i < k \}^{m-2-s-t} \\
& \left. 2 \left[\sum_{k'=1}^{N_2} P\{ {}_2S_0 = k', {}_2S_1 < k' \} P\{ {}_2S_i < k' \}^{s+t} \right] \right\}, \\
& \text{where } N_2 = \left[\frac{c - n_1 m}{r + t + 2} \right]_{Int}. \tag{24}
\end{aligned}$$

Since $P\{ {}_2S_0 = {}_2S_1 \}$ is uniquely highest in Stage II

$$= \sum_{k'=1}^{N_2} P\{ {}_2S_0 = k', {}_2S_1 = k' \} P\{ {}_2S_i < k' \}^{r+t},$$

(A) + (B)

$$\begin{aligned}
= & \sum_{s=0}^{r-3} \sum_{t=r-2-s}^{m-2-s} \left\{ P\{ {}_1S_0 \geq k, {}_1S_1 \geq k \} \frac{(m-2)!}{s! t! ((m-2) - s - t)!} \right. \\
& P\{ {}_1S_i > k \}^s P\{ {}_1S_i = k \}^t P\{ {}_1S_i < k \}^{m-2-s-t} \\
& \left. \left[\sum_{k'=1}^{N_2} [2P\{ {}_2S_0 = k', {}_2S_1 < k' \} + P\{ {}_2S_0 = k', {}_2S_1 = k' \}] P\{ {}_2S_i < k' \}^{r+t} \right] \right\},
\end{aligned}$$

where N_2 is defined by (24).

In order to calculate the probability, we need the distribution of ${}_2S_i$ and the joint distribution of $({}_lS_0=a, {}_lS_1=b)$, $a=0, 1, \dots, n_l$, $b=0, 1, \dots, n_l$, $l=1, 2$. Its computation has been laid out in (8).

APPENDIX II

Derivation of the Relation between Gene Detectabilities When There Is a Gene
That Is Responsible and When There Is No Gene That Is Responsible

$$P\{I \text{ is wrong and } {}_2S_I > t \mid \text{A gene is responsible}\} \quad (25)$$

$$= P\{{}_2S_I > {}_2S_j, \forall j \neq I, {}_2S_I > t \text{ and } I \neq i_0 \mid \text{the gene is at } i_0\}$$

$$= \sum_{k=t+1}^{n_2} P\{k = {}_2S_I > {}_2S_j, \forall j \neq I, \text{ and } I \neq i_0 \mid \text{the gene is at } i_0\}$$

$$= \sum_{k=t+1}^{n_2} P\{k > {}_2S_j, \forall j \neq I, i_0; k > {}_2S_{i_0}; {}_2S_I = k$$

$$\text{and } I \neq i_0 \mid \text{the gene is at } i_0\}$$

$$= \sum_{k=t+1}^{n_2} \{P\{k > {}_2S_j, \forall j \neq I, i_0, I \neq i_0 \mid \text{the gene is at } i_0\} \cdot$$

$$P\{k > {}_2S_{i_0} \mid \text{the gene is at } i_0\} \cdot \quad (26)$$

$$P\{{}_2S_I = k, I \neq i_0 \mid \text{the gene is at } i_0\}\},$$

and

$$P\{{}_2S_I > t \mid \text{no gene is responsible}\}$$

$$= \sum_{k=t+1}^{n_2} \{P\{k > {}_2S_j, j \neq I, i_0 \text{ and } I \neq i_0 \mid \text{no gene is responsible}\} \cdot$$

$$P\{k > {}_2S_{i_0} \mid \text{no gene is responsible}\} \cdot \quad (27)$$

$$P\{{}_2S_I = k \text{ and } I \neq i_0 \mid \text{no gene is responsible}\}\}.$$

Since Probability (26) is less than Probability (27) and the other corresponding terms are equal, $P\{I \text{ is wrong and } {}_2S_I > t \mid \text{a gene is responsible}\} \leq P\{{}_2S_I > t \mid \text{no gene is responsible}\}$.