Optimal Design of an Experiment

(Solution to Problem 1.2 of the Chapter on Power)

Benjamin Yakir

December 13, 2009

1 Introduction

The goal of this exercise is to find an optimal design for mapping a genetic trait in mice using the backcross design. Mapping a trait is a multistage process that typically initiates with a genome scan of a target population. In this study we concentrate on a population of mice created by backcrossing F_1 mice with one of the founding inbred strains.

The genetic characteristic of the mapping population is such that the genotype at each location is either heterozygote, with one copy originating from each one of the two founding inbred populations, or it is homozygote for the inbred population to which the F_1 was backcrossed. The recombination fraction between loci has rate of $\beta = 0.02$ (for distance is measured in units of cM).

Designing a genome scan involves decisions regarding the type of the target population for mapping, the size of the target population, the measurement of the phenotype(s) and other dependent and independent variable, the density of markers to be typed, the method of genotyping, the technology to be used, etc. Constraining factors may be costs, labor, and caging limitations. In this study we concentrate on the backcross design and assume that the technology of genotyping has been set and it involves the separate measurement of each marker for each of the mice. We also assume the labor and caging are not a limitation. Consequently, factors that need to be selected are the number of mice to be bred and the density of markers to be used.

Selection of the optimal design was based on statistical considerations. The statistical computations used the normal approximation, which may be justified by the Central Limit Theorem and the consideration of local alternatives.

2 Methods

The selection of the optimal design for the backcross involves the identification of the optimal density of markers and size of the mapping population. The constraining factor is the overall cost of the experiment, that is set to be $50,000. We set the overall cost associated with the breeding and phenotyping a mouse to be \$30. The cost of genotyping is a combination of fixed cost of \$70 per marker and a cost of \$2 for every reaction. For simplicity we assumed that the genome of the mouse is divided into 20 chromosomes of length 80 cM each.

Each design is composed of a selection of the size of the population (n.mice) and the number of markers per chromosome (n.mark). The total cost of each design is equal to

$$(\texttt{n.mice} \times 30) + (\texttt{n.mark} \times 70) + (\texttt{n.mice} \times \texttt{n.mark} \times 2)\ .$$

Only designs with a total cost of no more than \$50,000 we allowed.

For each set density of the markers the critical threshold for rejecting the null hypothesis was selected. The computation was carried via simulation of the process of test statistics over the set of markers under the null hypothesis of no genetic effect. The multivariate normal distribution was used in the simulation. The mean of components was zero and the covariance between any two markers on the same chromosome was set to be $\exp\{-0.02|s-t|\}$, where $|s-t|$ is the distance between the markers (in cM). Markers on different chromosomes are independent. The maximal absolute value of the test statistics across all markers and across all 20 chromosomes were computed. The 0.95 percentile of the distribution of that maximal value was set as the critical level. The simulation was based on 10,000 iterations.

Given the density of markers and given threshold a new set of simulations was conducted in order to determine the non-centrality parameter at the QTL that produces a statistical power of 85%. We set the location of the QTL to be midway between the first and the second markers on a given chromosome. The process of test statistics for that chromosome was simulated as a multivariate normal distribution with the same covariance structure as in the null and with a mean that reflects the presence of the QTL, and is equal to the non-centrality parameter at the QTL, times $\exp\{-0.02|q-t|\}$, where $|q-t|$ is the distance between the QTL and the marker (in cM). The maximal absolute value of the test statistics on the chromosome was computed and the power was assessed via the proportion of iterations in which that maximum exceeded the critical threshold. The simulation was based on 10,000 iterations.

Given the density of markers, the population size that produces a total cost of approximately \$50,000 was computed. Given this population size and given the non-centrality parameter we computed the minimal genetic effect that can be mapped with the given marker density using the formula

$$(\alpha + \delta)/\sigma_y = 2\xi/(\texttt{n.mice})^{1/2}\ .$$

The simulation was implemented in R and was carried out for 2, 3, 4, and 5 markers per chromosome. The code can be found on the course web site.

3 Results

The design characteristics of designs with a total cost of \$50,000 are listed in Table 1. In the first column the number of markers per chromosome is listed.

In the second column the number maximal number of mice that will produce the given total cost is recorded. In the third column the minimal genetic effect that will produce a power of 85% is given. Observe that the smallest effect is detectable with the density of 3 markers per chromosome and it increases as the number of markers increases. The effect for four markers is marginally larger.

Markers per Chr.	Number of Mice	Minimal Effect
2	429	0.771
3	305	0.661
4	234	0.684
5	187	0.727

Table 1: Designs with cost of $50,000

4 Discussion

In this study we considered optimal designs for a genome scan with the backcross design. Optimization was with respect to the smallest detectable genetic effect under a constraint on the overall cost of the trial. We found that for the backcross economic efficiency calls for the use of a small number of markers per chromosome, perhaps three, at most four.

The method that was applied here for the backcross design may be easily adapted to other designs, such as the intercross. In such a design different statistics may be considered, the 2 degrees of freedom chi-square statistic for example, and the correlation structure may be different. However, the method of simulation on search for the optimal design are essentially the same.