CONTROLLING THE PROPORTION OF FALSE POSITIVE (PFP) IN A MULTIPLE TEST GENOME SCAN FOR MARKER-QTL LINKAGE

R.L. Fernando[1,2], J.C.M. Dekkers[1] and M. Soller[3]

[1] Department of Animal Science, Iowa State University, Ames, IA 50011, USA
[2] L. H. Baker Center for Bioinformatics and Biol. Stat., Iowa State University, Ames, IA 50011, USA
[3] Department of Genetics, Hebrew University, 91904 Jerusalem, Israel.

INTRODUCTION

When heritability analysis shows that QTL are segregating in the population, the large number of markers employed ensures that an appreciable proportion is in linkage to segregating QTL. The challenge is to identify the markers in linkage, while controlling Type I error in this "multiple test" situation. Previous attempts have been based on controlling the overall genome-wise Type I error (GWER) (Lander and Kruglyak, 1995), or the False Discovery Rate (FDR) (Weller *et al*., 1998). The GWER approach results in a drastic reduction in experimental power, while the FDR approach is theoretically inappropriate (Zaykin *et al*., 2000). Combining previous studies (Southey and Fernando, 1998 ; Mosig *et al*., 2001) we here present an approach based on controlling the proportion of false positives (PFP), which avoids the limitations of both GWER and FDR. The PFP is derived from the posterior Type I error rate (PER) long used in human linkage studies (Morton, 1955 ; Ott, 1991).

THEORY

In linkage experiments between a marker locus and a monogenic disease locus, the PER is the probability that the true status between two loci is one of non-linkage, given a statistical test result interpreted as declaring the presence of linkage. In technical notation, let the true status of linkage between the two loci be represented by a random variable L that can take one of two values : L = 1 if the two loci are linked, and L = 0 if the two loci are not linked ; and let the declared status of linkage between the two loci on the basis of some statistical test, be represented by a random variable R that can also take one of two values : R = 1 if the two loci are declared linked, and R = 0 if the two loci are declared not linked. Then the PER is :

$$Pr(L = 0 \mid R = 1) = \alpha Pr(L=0)/[\alpha Pr(L=0) + \pi(L=1)]$$

where α is the comparison-wise Type I error rate (CWER), and π is the average power of the test for markers for which L = 1.

The PER as defined above is not directly applicable to a genome scan that uses a set of markers in well defined locations. In such a scan, of the total number of k markers in the scan, a set S_N of k_N markers are not linked to a trait gene, and a second set S_L of k_L markers are linked to a trait gene. For any marker in S_N, $Pr(L = 0) = 1$; for any marker in S_L, $Pr(L = 0) = 0$. Thus, regardless of the threshold used for declaring linkage, the PER will be 1 for markers in S_N, and 0 for markers in S_L. Nevertheless, although the concept of PER is not applicable, the allied concept of the proportion of false positives PFP remains applicable, where

$$PFP = k_N\alpha / (k_N\alpha + k_L\pi) \qquad [1]$$

Note that k is not included in the definition of PFP and hence the PFP does not depend on the number of markers used to scan the genome. This exceedingly useful property enables the PFP to be considered for each marker or the entire experiment without taking into account the number of tests for each experiment, or the number of experiments that the future holds.

Estimation of PFP for a given set of experimental results. In estimating PFP on the basis of a given set of experimental results, let k_D be the number of markers declared to be linked to QTL at the comparison-wise error rate (CWER), α. Then,

$$E(k_D)=k_N\alpha+k_L\pi.$$

Substituting in equation [1] and letting an apostrophe (') indicate estimated parameters, gives

$$PFP' = k'_N\alpha/k_D$$

Thus, to estimate PFP for given α, requires substituting the observed value of k_D at that α, and an estimate of k_N (see later) in [1]. k_D for given α is obtained by counting all markers that have a calculated P-value $<\alpha$. To find the critical CWER for the desired PFP, it is convenient to list all markers according to ascending P-values. For each P-value in the list, k_D will simply be the ordinal number of this CWER. The PFP that corresponds to choosing this P-value as the CWER is then calculated. The CWER corresponding to the desired PFP is chosen as the critical CWER.

Estimating k_N. Two methods were employed. Method I is based on somewhat speculative assumption. Method II is theoretically more precise.

Method I. As noted by Mosig *et al*. (2001), for markers in set S_N (unlinked to QTL), P-values generated by the test used to determine linkage, are expected to have a uniform distribution across the range 0 to 1.0, so that $k_N/2$ = the expected number of unlinked markers with P-values > 0.5. Markers in the set S_L (linked to QTL), are expected to show an excess of P-values < 0.5. If we assume all P-values > 0.5 are from S_N, then k_N can be estimated as (pers. comm. D. Nettleton) $k'_N = 2H$ where H is the number of markers with P-values > 0.5.

Method II. In this method, k_L is estimated iteratively as

$$k'_{L(i)} = (1-\gamma'_{(i-1)})k_D + \phi'_{(i-1)}(k-k_D)$$

Where,

γ is the proportion of Type I errors among all markers declared in linkage,

ϕ is the proportion of Type II errors among all markers not declared in linkage,

π is average power,

Subscript i represents the value of a parameter in the i^{th} iteration.

Then, π, γ, ϕ are estimated as

$$\pi'_{(i)} = (1-\gamma_{(i-1)})k'_D/k'_L$$

$$\gamma'_{(i)} = k'_{N(i)}\alpha / (k'_{N(i)}\alpha + k'_{L(i)}\pi_{(i)})$$

$$\phi'_{(i)} = k'_L(1-\pi'_{(i)})/ [k'_{N(i)}(1-\alpha) + k'_{L(i)}(1-\pi_{(i)})]$$

To start the iteration, we take initial values, $k_L = 0.05k$, $k_N = 0.95k$, and $\pi = 0.95$. These are used to compute $\gamma'_{(1)}$ and $\phi'_{(1)}$; iteration is continued until there is no change in the estimate of k_L.

Simulation. To examine the properties of the two methods, a simulation study was carried out, assuming QTL with effects corresponding to tests with power 0.21, 0.58, 0.88, and 0.98.

RESULTS

Simulation study. Table 1 shows the mean and mean squared error of estimates of k_L calculated for Methods I and II, from 500 replications of the simulation. Method I had an upward bias when k_L was small, but otherwise was relatively unbiased and robust. MSE did not vary greatly with k_L or power. Method II was unbiased for small k_L, and for larger k_L at high power; but showed a strong downward bias for larger k_L and moderate to low power, MSE also varied widely. Thus, Method I appears to be optimal when at moderate to low power, when k_L is high ; Method II appears to be optimal when at high power or when k_L is low.

Table 1. Mean of estimates of kL using Method I and Method II as obtained by simulation from 500 replicates of 100 marker tests. In parentheses, mean squared error (MSE) of the estimates. KL = true number of markers in linkage to QTL ; Power = power of the test, according to simulated QTL effect

KL	Power							
	----------Method I----------				----------Method II----------			
	0.21	0.58	0.88	0.98	0.21	0.58	0.88	0.98
1	4 (47)	4 (45)	4 (51)	4 (56)	1 (3)	1 (4)	1 (3)	1(4)
5	6 (54)	6 (62)	7 (62)	6 (54)	3 (9)	5 (6)	5 (7)	5 (6)
20	19 (80)	19 (74)	20 (89)	20 (74)	13 (61)	18 (9)	20 (4)	20 (3)
40	38 (63)	39 (54)	39 (62)	39 (63)	25 (243)	36 (19)	40 (4)	40 (3)
80	76 (42)	80 (20)	80 (18)	80 (22)	50 (942)	73 (59)	79 (2)	80 (1)

Data set study. Methods I and II were applied to the data set of Mosig *et al*. (2001) reporting a complete genome scan for milk protein %, using selective DNA pooling and 138 microsatellite markers. Using Method I, and the data in Table 2 of Mosig *et al*. (2001) we calculated H = 29 giving $k_N = 58$ and $k_L = 80$. Using these values of k_N and k_L, and data in table 5 of Mosig *et al*. (2001) critical CWER for PFP = 0.05 and 0.10 were 0.05 and 0.10, respectively. At these CWER there were 47 and 62 declared linkages, respectively, giving power of 0.59 and 0.78. Using a CWER of 0.001, as recommended by Lander and Kruglyak (1995), would have resulted in only 23 declarations of linkage, with power of 0.29. Using Method II with $\alpha = 0.05$, and the data in Table 5 of Mosig *et al*. (2001), we calculated $k_N = 97$ and $k_L = 41$ and PFP of 0.10 at CWER = 0.05. At this CWER there were 47 declared linkages, and power was 0.98. Thus, in this case, Method II apparently gave downward biased results, consistent with the simulation.

DISCUSSION

The PFP has the advantage of being independent of the number of elements included in the multiple test, and hence can be applied to individual experiments or parts of experiments,

without adjustment for the number of markers or traits included in the experiment, as required by GWER approaches. Furthermore, low levels of PFP (0.05 or 0.01) are obtained with relatively relaxed CWER. This increases the power of the experiment as compared to GWER. As compared to the FDR, the PFP is theoretically correct, and it can be shown that at equivalent levels the PFP will always have greater power than the FDR. Thus, the PFP shares the theoretical advantages of the FDR compared to methods based on controlling the GWER, while providing greater power.

The main requirement for application of the PFP approach is the ability to estimate, k_L, the number of markers in S_L. This can be achieved on a priori grounds as shown by Southey and Fernando (1998), it can also be achieved by analysis of the data themselves, as shown by Mosig *et al*. (2001) and the present study. Two methods for this were presented here and tested by simulation and application to an actual data set. Both need to be evaluated more extensively. Although presented here in terms of single marker mapping, the PFP can readily be adapted to interval analysis.

REFERENCES

Lander, E. and Kruglyak, L. (1995) *Nat. Genet.* **11** : 241-247.

Morton, N. (1955) *Am. J. Hum. Genet.* **7** : 277-318.

Mosig, M., Lipkin, E., Khutoreskaya, G., Tchouryzna, E., Ezra, E., Soller, M. and Friedmann, A. (2001) *Genetics* **157** : 1683-1698.

Ott, J. (1991) "Analysis of Human Genetic Linkage". Johns Hopkins University Press, Baltimore.

Southey, B.R. and Fernando, R.L. (1998) *Proc. 6th WCGALP* **26** : 221-224.

Weller, J.I., Song, J.Z., Heyen, D.W, Lewin, H.A. and Ron, M. (1998) *Genetics* **150** : 1699-1706.

Zaykin, D.V., Young, S. and Westfall, P.H. (2000) *Genetics* **154** : 1917-1918.