NON-PARAMETRIC INTERVAL MAPPING IN HALF-SIB DESIGNS : USE OF MIDRANKS TO ACCOUNT FOR TIES

P. Tilquin[1], W. Coppieters[2], I. Van Keilegom[3] and P.V. Baret[1]

[1] Unité de Génétique, Faculté d'ingénierie biologique, agronomique et environnementale,
Université catholique de Louvain, Croix du Sud 2 bte 14, B-1348 Louvain-la-Neuve, Belgium
[2] Department of Genetics, Faculty of Veterinary Medicine, University of Liège, Boulevard de Colonster 20, B-4000 Liège, Belgium
[3] Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

INTRODUCTION

Most quantitative trait loci (QTL) mapping methods share a common assumption : that the phenotype follows a normal distribution. However, many phenotypes of interest are not normally distributed. Phenotypes such as bacteria counts or CFU (for colony-forming units) are used to assess the resistance of animals to bacterial diseases (*e.g.* Berthelot *et al.*, 1998). The distribution of this type of traits is extremely skewed and a high amount of zero values is often observed. From a statistical point of view, those zero values are ties. Similar distributions are obtained when studying the resistance to parasitic diseases by use of faecal egg counts (FEC) (*e.g.* Bouix *et al.*, 1998). A classic solution is to apply a mathematical transformation. An alternative approach is to use non-parametric methods. Kruglyak and Lander (1995) described a non-parametric interval mapping approach based on the Wilcoxon rank-sum test applicable to experimental crosses. Coppieters *et al.* (1998) adapted this method to half-sib pedigrees in outbred populations.

Tilquin *et al.* (2001) compared the power of QTL mapping methods in half-sib designs when bacteria counts are analysed. In this study, the power of the non-parametric method from Coppieters *et al.* (1998) decreased in the presence of ties. This loss of power is inherent to the way ties are ranked. Indeed, in order to respect the null distribution of their test statistic, Kruglyak and Lander (1995) and Coppieters *et al.* (1998) choose to rank tied individuals at random. This approach has the benefit of simplicity, because no new theory is necessary : the variance of the test statistic is unaffected. However, when dealing with phenotypes presenting a high number of ties, the information about tying is ignored, and furthermore, new information is added in the data (*i.e.* individuals are ordered although they were tied). Another approach in dealing with ties is to assign to each tied individual the average of the tied ranks (often referred to as the midrank method) (*e.g.* Lehmann, 1975).

Therefore, our objective is to develop a non-parametric test statistic using midranks and to compare it to the random ranking approach in terms of statistical power.

MATERIAL AND METHODS

The non-parametric interval mapping test statistic. Principles of this approach were extensively presented in Coppieters *et al.* (1998). For a single family, the test statistic is defined as :

$$Y_K(p)=\sum_{j=1}^{N}\left[N+1-2\,rank_j\right]\left(P_j(A)-P_j(B)\right)$$

where N is the number of progeny in the half-sibship; $rank_j$ is the rank by phenotype of progeny j; and $P_j(A)$ (and $P_j(B)$) are the probabilities – conditional to marker information – that offspring j inherits homologue A (or B) from its sire at the position p being considered. Under the null hypothesis, $Y_K(p)$ can be shown to be normally distributed with mean 0 and variance $Var\left(Y_K(p)\right)=\frac{1}{3}\left(N^3-N\right)Var\left(P_j(A)-P_j(B)\right)$. One uses a standard normal variable to test the significance of the QTL effect at map position p : $Z_K(p)=Y_K(p)\Big/\sqrt{\left\langle Var\left(Y_K(p)\right)\right\rangle}$. The analysis is performed across families by squaring and summing the individual $Z_K(p)$ scores over all s families to yield a χ^2 statistic.

Adaptation to use midranks. To use midranks for ties instead of random ranks, the variance of the test must be corrected by introducing in $Var\left(Y_K(p)\right)$ a correction proposed in Lehmann (1975) for the variance of ranks. If N observations take on e distinct values, and d_1 observations are equal to the smallest value, d_2 to the next smallest, …, d_e to the largest, then the corrected variance of $Y_K(p)$ is :

$$Var\left(Y_K^*(p)\right)=\frac{1}{3}\left[\left(N^3-N\right)-\sum_{i=1}^{e}d_i\left(d_i^2-1\right)\right]Var\left(P_j(A)-P_j(B)\right)$$

The non-parametric test using midranks will be referred to as NP-MI (as opposed to the test using random ranking referred to as NP-RA).

Interval mapping by regression. Both non-parametric tests are compared to regression interval mapping (Knott *et al.*, 1996), hereafter referred to as RIM.

Simulated data set. The segregation of a QTL is simulated in a half-sib design (30 families of 40 half-sibs). Eleven markers are evenly spaced on a 100 cM chromosome. The number of alleles at the markers is equal to 16, to mimic a fully informative situation. The QTL is simulated at position 35, and has 2 alleles with equal frequency. Heritability of the trait is 0.25, and the QTL accounts for 8% of total phenotypic variance. Simulation process is based on an algorithm described by Baret *et al.* (1998). A normally distributed phenotype is simulated and used as a reference. Bacteria counts with increasing proportions of zeros (0%, 20%, 50% and 80%) are simulated using an approach referred to as *normal score back transformation* (Tilquin *et al.*, 2001). This approach uses the quantiles of observed distributions to generate non-normal phenotypes from a normally distributed one. A distribution of CFU with 8.5% of zeros was obtained from the study of Frédéric Lantier (pers. com.) who performed an artificial infection in the sheep on 30 sires families of 40 half-sibs with a live vaccine. The different proportions of zeros were generated by suppressing or adding zero values to the observed distribution.

Power estimates using permutations. For each phenotype, 1000 replicates were simulated and analysed with both non-parametric methods (NP-RA and NP-MI) and the RIM method. Permutations were used to estimate the significance levels reached for each of these analyses. For each replicate, 1000 permutations were performed and analysed. The proportion of replicates yielding a *p*-value lower than 0.05 was used to measure the corresponding power of both NP and the RIM methods.

RESULTS AND DISCUSSION

Power estimates were obtained for the three methods and for all phenotypes. With the RIM method, bacteria counts were either analysed as such or by making a mathematical transformation prior to analysis (*log*(*X*+1)) (Table 1).

Table 1. Power estimates (%) from 1000 replicates of chromosome scans with a QTL accounting for 8% of total phenotypic variance

	Phenotype				
		Bacteria counts			
Method	Normal	0%	20%	50%	80%
NP-RA	55	55	54	38	13
NP-MI	55	55	54	46	24
RIM	60	27	24	18	11
log+RIM	--	59	56	47	25

For low proportions of zeros in bacteria counts (0% and 20%), both NP tests perform identically. For high proportions of zeros in bacteria counts (50% and 80%), the power is higher using NP-MI : (1) for bacteria counts with 50% of zeros, the *relative* gain of using NP-MI compared to NP-RA is 21% (46 *vs.* 38); (2) for bacteria counts with 80% of zeros, the relative gain reaches the value of 85 % (24 *vs.* 13). The advantage of the NP-MI method compared to the NP-RA method is also observed in the bias of the estimated QTL location (over 1000 replicates). As expected, for all methods, there is a bias towards the centre of the chromosome. For bacteria counts with 50% of zeros, the biases (± s.e.) are respectively 5.0 ± 0.7 and 3.9 ± 0.7 for the NP-RA and NP-MI methods, *i.e.* a difference of 1.1 cM. When the proportion of zeros is 80%, the difference is 4 cM (10.6 ± 0.9 and 6.6 ± 0.8 respectively for NP-RA and NP-MI). The difference in power of both non-parametric methods is reflected in the flatness of the curves of the test statistic along the chromosome (Figure 1). In the conditions of our study, when a *log*-transformation is applied to bacteria counts prior to analysis, the RIM method gives the same results as the NP-MI method.

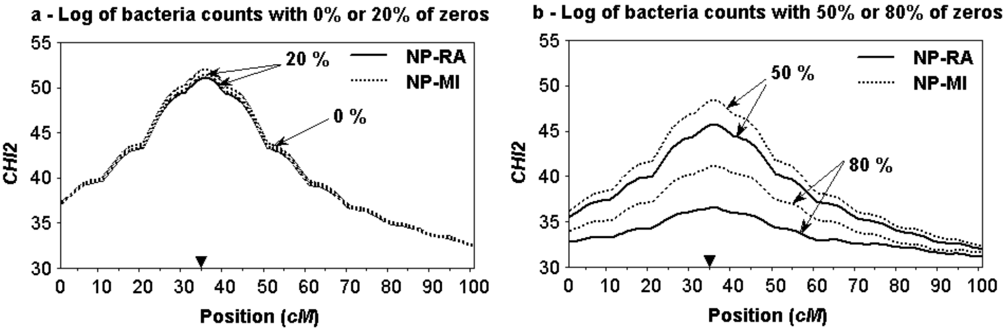

Figure 1. Mean test statistic values (1000 replicates) along the chromosome for both non-parametric methods (arrows indicate the location of the simulated QTL)

CONCLUSION

These results indicate that, it is worth using the midranks approach when non-parametric interval mapping is chosen. Furthermore, the midranks approach is easy to implement. Compared to the classical RIM, the NP-RA and NP-MI methods have the advantage of being distribution-free, which should give higher power when no mathematical transformation is available. The advantage of using midranks for ties instead of random ranking is especially observed with extreme phenotypes which are relatively common when analysing the genetic basis of disease resistance.

ACKNOWLEDGMENTS

We are grateful to Frédéric Lantier for providing us the bacteria counts data, and to Eric Le Boulengé and Jérémie Nsengimana for their comments on earlier versions of the paper.

REFERENCES

Baret, P.V., Knott, S.A. and Visscher, P.M. (1998) *Genet. Res.* **72** : 149-158.

Berthelot, F., Beaumont, C., Mompart, F., Girard-Santosuosso, O., Pardon, P. and Duchet-Suchaux, M. (1998) *Poultry Sci.* **77** : 797-801.

Bouix, J., Krupinski, J., Rzepecki, R., Nowosad, B., Skrzyzala, I., Roborzynski, M., Fudalewicz-Niemczyk, W., Skalska, M., Malczewski, A. and Gruner, L. (1998) *Int. J. Parasitol.* **28** : 1797-1804.

Coppieters, W., Kvasz, A., Farnir, F., Arranz, J. J., Grisart, B., Mackinnon, M., and Georges, M. (1998) *Genetics* **149** : 1547-1555.

Knott, S. A., Elsen, J. M. and Haley, C. S. (1996) *Theor. Appl. Genet.* **93** : 71-80.

Kruglyak, L. and Lander, E. S. (1995) *Genetics* **139** : 1421-1428.

Lehmann, E. L. (1975). « Nonparametrics - Statistical methods based on ranks ». Holden-Day, Inc., San Francisco, USA.

Tilquin, P., Coppieters, W., Elsen, J. M., Lantier, F., Moreno, C. and Baret, P. V. (2001) *Genet. Res.***78** : 303-316.