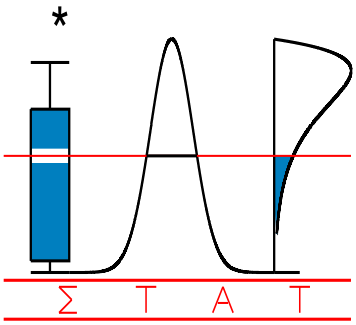
T E C H N I C A L

R E P O R T

0216

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

P. TILQUIN, I. VAN KEILEGOM, W. COPPIETERS, E. LE BOULENGE and P.V. BARET



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Non-parametric interval mapping in half-sib designs : use of midranks to account for ties

P. Tilquin[1], I. Van Keilegom[2], W. Coppieters[3], E. Le Boulengé[4] 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] Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

[3] Department of Genetics, Faculty of Veterinary Medicine, University of Liège, Boulevard de Colonster 20, B-4000 Liège, Belgium

[4] Unité d'environmétrie et géomatique, Faculté d'ingénierie biologique, agronomique et environnementale, Université catholique de Louvain, Croix du Sud 2 bte 16, B-1348 Louvain-la-Neuve, Belgium

Proofs to be sent to : Philippe Baret, Université catholique de Louvain, Faculté d'ingénierie biologique, agronomique et environnementale, Unité de Génétique, Croix du Sud 2 bte 14, B-1348 Louvain-la-Neuve, Belgium. Tel: +32-10-47.37.23, Fax: +32-10-47.37.28, E-mail: baret@gena.ucl.ac.be

Running title : Non-parametric QTL mapping

[*] Corresponding author. e-mail: baret@gena.ucl.ac.be

SUMMARY

In QTL analysis of non-normally distributed phenotypes, non-parametric approaches have been proposed as an alternative to the use of parametric tests on mathematically transformed data. The non-parametric interval mapping test uses random ranking to deal with ties. Another approach is to assign to each tied individual the average of the tied ranks (midranks). This approach is implemented and compared to the random ranking approach in terms of statistical power and accuracy of the QTL position. Non-normal phenotypes such as bacteria counts showing high numbers of zeros are simulated (from 0% to 80% of zeros). We show that for low proportions of zeros, the power estimates are similar, but for high proportions of zeros, it is worth using midranks instead of the random ranking approach : *e.g.* with a QTL accounting for 8% of the total phenotypic variance, a gain from 8 to 11% of power can be obtained. Furthermore, the accuracy of the estimated QTL location is increased when using midranks. Therefore, if non-parametric interval mapping is chosen, the midrank approach should be preferred. This test might be especially relevant for the analysis of disease resistance phenotypes such as those observed when mapping QTL for resistance to infectious diseases. Indeed, the resistance of animals to bacterial and parasitic diseases is often based on counts that can show high number of ties.

1. 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. Examples include counts (bacteria counts [CFU], faecal egg counts [FEC], plaque forming units [PFU], somatic cell counts [SCC], number of tumours, …), survival times and qualitative data (severity grades assigned upon histological examination). Traditional QTL mapping methods cannot be directly applied in such cases. On the one hand, ANOVA and least-squares based methods (Haley & Knott, 1992; Martinez & Curnow, 1992; Weller *et al.*, 1990) assume that residual errors, *i.e.* residuals within QTL genotype classes, are normally distributed. Such methods are commonly said to be robust against non-normality; however, the limits of this robustness against non-normality in the context of QTL mapping methods have, to our knowledge, never been established. On the other hand, maximum-likelihood based methods (Lander & Botstein, 1989) use the Normal density function for the structure of the likelihood itself. Quality of estimations is therefore very dependent on phenotype's normality. One approach to circumvent the assumption of normality is to apply a mathematical transformation that will convert the trait into an approximately Normal variable. An alternative approach is to use non-parametric (or distribution-free) methods. Recently, Kruglyak & Lander (1995*a*) 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.

As mentioned by Coppieters *et al.* (1998), this non-parametric test extends the scope of QTL mapping in outbred half-sib pedigrees to a variety of non-normally distributed traits. However, a disadvantage of this rank-based method is that it doesn't provide convenient estimates of QTL effects. It is therefore suitable for the detection of QTL but has to be combined with other methods, such as least-squares or maximum-likelihood techniques, in order to estimate QTL effects.

In a recent study, Tilquin *et al.* (2001) compared three different QTL mapping methods in the context of resistance to bacterial diseases : 1) least-squares analysis, 2) maximum-likelihood analysis and 3) non-parametric mapping. When searching for QTL implied in the resistance to bacterial diseases, the distribution of the trait under study (bacteria counts) is extremely skewed to the right and can present a high frequency of zero values. From a statistical point of view, those zero values are ties (ex-aequos). Tilquin *et al.* (2001) showed that, when analysing bacteria counts, losses of power in parametric and non-parametric methods can be explained by two distinct causes : respectively the asymmetry of the phenotypic distribution and the existence of ties.

Two methods of treating ties have been discussed in the literature. The first is to order tied observations at random (often referred to as the random ranking approach). The second method is to attribute to each of the tied observations a same value : the average of the tied ranks, *i.e.* the mean of the ranks that the observations would have if they were not tied (often referred to as the midranks method) (see *e.g.* Lehmann, 1975).

In order to respect the null distribution of their interval mapping test statistic which is asymptotically Normal, Kruglyak & Lander (1995*a*) and Coppieters *et al.* (1998) choose to rank tied individuals at random.

The random ranking approach has the benefit of simplicity, because no new theory is necessary : the variance of the test statistic is unchanged in presence of ties. However, when dealing with phenotypes presenting a high number of ties (for example, zero-inflated distributions as encountered in the resistance to bacterial or parasitic infections), the information of tying is ignored, and furthermore, new information is added in the data (*i.e.* individuals are ordered although they were tied). According to Kendall & Stuart (1979), random ranks could lead to a reduced efficiency. As stated by Lehmann (1975) for the classic Wilcoxon rank-sum test, when attributing midranks instead of random ranking, the variance of the test statistic now depends on the number and type of ties observed. Therefore, the asymptotic Normal distribution of the test statistic no longer holds. Consequently, to allow for the use of midranks, it is necessary to modify the test statistic of Coppieters *et al.* (1998) and to reformulate its asymptotic properties.

Therefore, the aim of the present study is to obtain a non-parametric test statistic using midranks, and to compare it to the random ranking approach in the presence of ties (counts data). Comparisons will be carried out in terms of statistical power and in terms of the accuracy of the estimated QTL location.

2. MATERIALS AND METHODS

After recalling briefly the principle of the Wilcoxon rank-sum test, the adaptation of the test statistic for the interval mapping problem in animal genetics is detailed. Afterwards, a modification of the test statistic to account for high number of ties by attributing midranks is presented. Formulae of the variance of the original and of the new test statistic are demonstrated in the Appendix. By means of simulations, the new interval mapping test is eventually compared to the non-parametric interval mapping test using random ranks and to the classical regression interval mapping test.

(i) *Non-parametric interval mapping*

(a) *The Wilcoxon rank-sum test*

Assume that you study a sample of size N and that approximately half (m) of the sample has received a treatment (A), the other half (n=N-m) having received treatment (B). The individuals in the sample can be ranked according to their value for a continuous variable and be assigned the corresponding rank. The test statistic commonly used is the sum of the ranks within treatment. For example for individuals having received treatment (A) :

$$T_A = \sum_{j(A)=1}^{m} rank_{j(A)} \qquad (1)$$

For small values of m and n and in the absence of ties, it is possible to obtain the distribution of T_A under the null hypothesis of no difference between treatments.

Tables were set up when both m and n are less than or equal to 10 (see e.g. Lehmann, 1975). For larger values of m and n, one can use an asymptotic approximation, based on the central limit theorem for dependent random variables :

$$\frac{T_A - E(T_A)}{\sqrt{Var(T_A)}} \xrightarrow[D]{} N(0,1), \quad N \to \infty \tag{2}$$

$$T_A \sim N\left(\frac{m(N+1)}{2}, \frac{mn(N+1)}{12}\right) \tag{3}$$

meaning that T_A is asymptotically distributed as a Normal variable with the specified mean and variance. Using this asymptotic approximation, critical values of the two-sided test at the α level can be obtained.

As detailed in Georges & Coppieters (2000), instead of using the classic sum of ranks T_A, one can use the deviation between the expected and observed sum of ranks for group A and group B, as well as the average of the two deviations as :

$$Y_W = \frac{1}{2}\left[\left(\frac{m(N+1)}{2} - \sum_{j(A)=1}^{m} rank_{j(A)}\right) - \left(\frac{n(N+1)}{2} - \sum_{j(B)=1}^{n} rank_{j(B)}\right)\right] \tag{4}$$

and rewrite Y_W as follows :

$$Y_W = \frac{1}{2}\sum_{j=1}^{N}\left[\left(\frac{(N+1)}{2} - rank_j\right)\left(P_j(A) - P_j(B)\right)\right] \tag{5}$$

where $P_j(A)$ and $P_j(B)$ take values of 1 and 0 for samples having received treatment A, and 0 and 1 for samples having received treatment B. Such a rearrangement was

performed in order to set up a test statistic regrouping the sum of ranks of both treatments. This test statistic has the following properties :

$$E(Y_W) = 0 \tag{6}$$

$$Var(Y_W) = \frac{mn(N+1)}{12} \tag{7}$$

For large values of *m* and *n* and under the null hypothesis of no effect of the treatment on the continuous variable, one can use the classic asymptotic approximation :

$$\frac{Y_W}{\sqrt{Var(Y_W)}} \xrightarrow[D]{} N(0,1), \quad N \to \infty \tag{8}$$

$$Y_W \sim N\left(0, \frac{mn(N+1)}{12}\right) \tag{9}$$

which allows for a simple test of the significance of the treatment.

(b) *Interval mapping analysis*

The Wilcoxon rank-sum test was adapted to the interval mapping problem by Kruglyak & Lander (1995*a*). Principles of this approach were extensively presented in Georges & Coppieters (2000) and Coppieters *et al.* (1998).

The slightly modified version of the two-sided Wilcoxon rank-sum test presented here above (equation (5)) can easily be applied for QTL mapping in a single half-sib pedigree by considering that the treatments are either to receive homologue A

from the sire at map position p versus receiving homologue B from the sire. Kruglyak & Lander (1995*a*) for experimental crosses, and Coppieters *et al.* (1998) for outbred half-sib pedigrees, define the statistic $Y_K(p)$, which is equivalent to $4*Y_W$; for the sake of simplicity, only the statistic adapted to outbred half-sib pedigrees will be detailed :

$$Y_K(p) = \sum_{j=1}^{N} \left[N + 1 - 2\, rank_j \right] \left(P_j(A) - P_j(B) \right) \tag{10}$$

In (10), 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 (equivalent to the notation $P[g_{i,A}(p)|g_{i,L},g_{i,R}]$ in Coppieters *et al.*, 1998). At marker location, $P_j(A)$ (and $P_j(B)$) takes values 0 or 1.

The major difference with the conventional Wilcoxon test defined above (equation (5)) is therefore the fact that $P_j(A)$ and $P_j(B)$ can take on any value between 0 and 1 rather than only the values of 0 or 1. Under the null hypothesis of no QTL (*i.e.* no effect of the homologue received from the sire on phenotype), and for large values of N, $Y_K(p)$ can be shown to be normally distributed with mean 0 and variance :

$$Var\left(Y_K(p)\right) = \left(\frac{N^3 - N}{3} \right) Var\left(P_j(A) - P_j(B) \right) \tag{11}$$

where $Var\left(P_j(A)-P_j(B)\right)$ is the expected variance of $P_j(A)-P_j(B)$ over all possible marker genotypes (see Appendix A for the demonstration to obtain $Var\left(Y_K(p)\right)$). Therefore, one can define the statistic :

$$Z_K(p)=\frac{Y_K(p)}{\sqrt{Var\left(Y_K(p)\right)}} \tag{12}$$

which is distributed as a standard Normal variable and allows to test the significance of the QTL effect at map position *p*. This test statistic reduces to a classic Wilcoxon rank-sum test at the marker positions.

In outbred designs (*i.e.* half-sib designs), the expected variance of $P_j(A)-P_j(B)$ depends on marker allele frequencies and genotype of the founder sire. Coppieters *et al.* (1998) calculate the value of $Var\left(P_j(A)-P_j(B)\right)$ for each half-sib pedigree by simulating all possible offspring and by calculating a frequency weighted variance of $P_j(A)-P_j(B)$. Instead, for the sake of simplicity, we used the observed variance of the differences between observed values of $P_j(A)$ and $P_j(B)$.

When doing across family analysis, the different sibships cannot be assumed to segregate for the same QTL or even for the same QTL alleles; *i.e.*, one cannot assume locus and allelic homogeneity across families. Instead of analysing every half-sib family separately, Coppieters *et al.* (1998) proposed to square and sum the

individual $Z_K(p)$ scores over all s families to yield a χ^2 statistic with s degrees of freedom :

$$\chi^2_{(s)} = \sum_{i=1}^{s} Z_K^2(p)_i \qquad (13)$$

(ii) ***Modification to use midranks***

In presence of ties, the null distribution of $rank_j$ (and consequently of the sum of ranks T_A) no longer holds. The mean is unchanged but the variance of the ranks decreases as the number of ties in the data increases (Lehmann, 1975). To respect the null distribution of $rank_j$, Kruglyak & Lander (1995*a*) and Coppieters *et al.* (1998) ranked tied individuals at random. This approach has the benefit of simplicity, because no new theory is necessary and all the results above apply directly. However, it does ignore some information contained in the data. Furthermore, it adds new information in the data (especially when groups of ties are large) : indeed, it ranks differently individuals that should have the same rank.

Instead of ranking tied individuals at random, one could attribute the same rank to all individuals belonging to a same group of ties (*i.e.* the average of the ranks that the observations would have if they were not tied) (see *e.g.* Lehmann, 1975). This approach is somewhat more correct since no new information is added in the data : individuals with the same phenotypic value get the same rank. However, it has the drawback that hypotheses of the test no longer hold (the variance of the test statistic is changed). Hence, the null distribution of the Wilcoxon rank-sum statistic

depends on the number and pattern of tied values within and between the groups. When there are tied values, the variance of ranks is reduced.

The null distribution of the interval mapping statistic ($Z_K(p)$) depends also on the number and pattern of ties. Indeed, the test statistic within a single family is composed of $Y_K(p)$ divided by its expected standard deviation $\sqrt{Var\left(Y_K(p)\right)}$, which uses the variance of ranks in its formula (see equation [A 10 *c*] in the Appendix A).

In their study, Tilquin *et al.* (2001) showed that when the proportion of ties (zeros) is high in the phenotype (*e.g.* in bacteria counts), the power of the non-parametric test is decreased. It is therefore of interest to modify the computer program from Coppieters *et al.* (1998) in a way that tied individuals are given midranks, and to obtain a new formula for the variance of the test statistic taking into account that midranks are used rather than random ranking.

A correction for the variance $Y_K(p)$ is easily found by introducing in (11) a correction for the variance of ranks proposed by Lehmann (1975) (see Appendix B for a demonstration). Suppose that the N observations take on e distinct values, and that d_1 of the N observations are equal to the smallest value, d_2 to the next smallest, …, d_e to the largest. With this notation :

$$Var\left(Y_K^*(p)\right) = \frac{1}{3}\left[\left(N^3 - N\right) - \sum_{i=1}^{e} d_i\left(d_i^2 - 1\right)\right] Var\left(P_j(A) - P_j(B)\right) \qquad (14)$$

which reduces to the previous variance formula in the absence of ties. Therefore, one can define the following new statistic corrected for the presence of ties :

$$Z_K^*(p) = \frac{Y_K^*(p)}{\sqrt{Var\left(Y_K^*(p)\right)}} \qquad (15)$$

which is distributed as a standard Normal variable and allows to test the significance of the QTL effect at map position *p*. As the previous $Z_K(p)$, this test statistic reduces to a classic Wilcoxon rank-sum test at the marker positions and, if summed over all *s* families, yields a χ^2 statistic with *s* degrees of freedom.

(iii) *Interval mapping by regression*

Both non-parametric tests (using random ranking of ties or using midranks, hereafter referred to as NP-RA and NP-MI respectively) were compared to regression interval mapping (RIM). The RIM approach is based on a regression of the value of the trait on the probabilities of inheriting a given gamete from the sire (see Knott *et al.*, 1996). The evidence in favour of a segregating QTL, at a given chromosome position, is assessed by an *F* test : ratio of the between-marker alleles within-sire mean square to the residual mean square.

(iv) ***Comparison of methods***

(a) *Simulated data set*

To test the efficacy of the non-parametric test using midranks (NP-MI), we simulated the segregation of a QTL in a half-sib design : 30 families of 40 half-

sibs. Each sire is randomly mated to 40 unrelated dams and the trait is measured on a single offspring per mating, which amounts to 1,200 measured individuals. Heritability of the trait is set to 0.25 in all simulations. Simulations are carried out under the hypothesis of one QTL segregating (H_1) and are repeated 1,000 times. Eleven markers are evenly spaced (interval 10 cM) on a 100 cM segment. The QTL is positioned in the centre of the fourth interval (position 35 cM on a 100cM chromosome). In all simulations, the number of alleles at the markers is equal to 16, occurring with equal frequency. Such a number of alleles was chosen to mimic a fully informative situation. Furthermore, the dam allele is specifically coded to be always identifiable. The number of alleles at the QTL is equal to 2 with equal frequency (0.5). Seven levels of heritability due to the QTL are simulated : 0.5%, 2%, 4.5%, 8%, 12.5%, 18% and 24.5% of the total phenotypic variance (*i.e.* 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7 in terms of Falconer & Mackay's substitution effect).

The simulation process is based on an algorithm already described by Baret *et al.* (1998). A normally distributed phenotype is simulated and used as a reference. Non-normally distributed phenotypes presenting various proportions of ties are simulated according to an approach used in geostatistics and referred to as *normal score back transformation*; see Tilquin *et al.* (2001) for the rationale of this transformation. It can be seen as a correspondence table between the quantiles of two different distributions, a simulated normally distributed phenotype and an observed non-normal phenotype.

In order to compare both non-parametric statistics in the presence of ties, zero-inflated continuous phenotypes will be simulated with increasing proportions of

zeros (ties), *i.e.* bacteria counts presenting 0%, 8.5%, 20%, 35%, 50%, 65% and 80% of zeros. The distribution of bacteria counts with 8.5% of zeros is simulated using an observed distribution from the study of Frédéric Lantier (personal communication) who performed an artificial infection in the sheep on 30 sires families of 40 half-sibs with a live vaccine (vaccinal *Salmonella abortusovis* strain Rv6). In this experiment, blood, spleen and lymph nodes of sheep were sampled at slaughtering (10 days later) for the assessment of bacterial colonisation and antibody responses, and for DNA typing. Bacteria counts (per gram of the organ dry weight), *i.e.* CFU were measured in the spleen and in left and right lymph nodes. Counts from both lymph nodes were pooled and the resulting distribution had 8.5% of zeros (n=1170). This distribution was used as a reference for our simulations. Other distributions of bacteria counts used for the simulations (with 0%, 20%, 35%, 50%, 65% and 80%) will be artificially generated by suppressing zero values (yielding 0% of zeros) or adding zero values to the original distribution of bacteria counts with 8.5% of zeros (yielding distributions with the desired proportion of zeros). The simulation of all bacteria counts is made by applying the *normal score back transformation* to all values of the simulated normal phenotype.

As the RIM method requires normally distributed residuals, for this method, a logarithmic transformation was applied to bacteria counts prior to analysis. A constant 1 was added to the phenotypic values prior to transformation to avoid negative infinity problems due to the log of zero.

(b) *Significance thresholds and power estimates*

For both non-parametric methods (NP-RA and NP-MI) and for the RIM method, significance thresholds were determined using permutations of the phenotypes (or ranks) as suggested by Churchill & Doerge (1994). For each simulation, 1,000 permutations were carried out. Phenotypes were permuted within family. For each permutation, the highest value of the test statistic over the entire chromosome was retained to yield chromosome-wise distributions of the maximum test statistics under the null hypothesis and under the specific conditions of each simulation. A simulation was declared as significant when its test statistic value was higher than the 5% chromosome-wise threshold obtained by permutations. Among the 1,000 simulations, the proportion of significant simulations was used as a power estimate. Power estimates are considered as different if difference is higher than $z_{1-\alpha/2}\sqrt{u(1-u)\ (1/n_1+1/n_2)}$, where $z_{1-\alpha/2}$ is the 1-α/2 quantile of the standard $N(0,1)$, u is the proportion of significant runs pooled across methods; and n_1 and n_2 are the number of runs for each method. The significance level of this test was adjusted for multiple comparisons using the Bonferroni method. As power is calculated in parallel on the same simulated data, the results of the methods of analysis are positively correlated and this test is conservative (Baret *et al.*, 1998).

To compare the accuracy of the three QTL mapping methods, the mean bias (± s.e.) of the estimated QTL location was computed over the 1,000 simulations. For each simulation, the estimated QTL location was given by the position maximising the test statistic along the chromosome. The bias of the estimated QTL location was computed by subtracting 35, the simulated location of the QTL, from the estimated position. Mean bias values were compared using Bonferroni *t* tests.

Furthermore, 95% confidence intervals of the mean estimated QTL location were computed.

3. RESULTS

(i) *Simulated data*

Bacteria counts with various proportions of zeros were simulated. Average percentages of zeros (mean ± s.e.) in simulated bacteria counts over the 1,000 replicates and with a QTL substitution effect of 0.4 were 8.2 ± 0.03%, 19.6 ± 0.4%, 34.7 ± 0.8%, 49.9 ± 1.3%, 65.1 ± 1.8% and 80.3 ± 2.3% respectively for bacteria counts with 8.5%, 20%, 35%, 50%, 65% and 80% of zeros. When other values of QTL substitution effect were simulated, proportions of zeros in simulated bacteria counts were very similar to the proportions observed with a QTL effect of 0.4 (data not shown).

(ii) *Power estimates*

For the three QTL mapping methods (NP-RA, NP-MI and RIM), the power was estimated at various levels of QTL substitution effect (a) and for increasing proportions of zeros (ties) in the phenotype. Full results are presented for a QTL substitution effect of 0.4 (*i.e.* 8% of total phenotypic variance) (Table 1).

[Table 1]

The evolution of the power according to the QTL effect and for increasing proportion of zeros is depicted by means of power curves (Figure 1).

[Figure 1]

Up to 20.0% of zeros, there were no differences in power between the random ranking and the midranks approaches. From 35.0% to 80.0% of zeros, the advantage of using the midranks approach increased from 1% to 43% according to the QTL substitution effect. For a QTL effect of 0.4 (8% of total phenotypic variance), the gains of using the midranks approach for phenotypes with 35%, 50%, 65% and 80% of zeros were 2%, 8%, 11% and 11% of power respectively (gains were significant for 50%, 65% and 80% of zeros). The power of NP-MI was always significantly higher than the power of NP-RA when searching for QTL with effects higher or equal to 0.4 (8% of total phenotypic variance) and for high proportion of zeros (50%, 65% and 80% of zeros). The gain in power using NP-MI compared to NP-RA is therefore dependent on the levels of the QTL substitution effect (*a*) and of the proportions of zeros (ties) in the phenotype (Figure 2).

[Figure 2]

For phenotypes presenting 80% of zeros, a little gain of power of 2-3% was obtained using NP-MI compared to RIM. However, power estimates of NP-MI and RIM were not significantly different. The power of RIM was significantly higher (4%) than the power of NP-MI only when analysing the normal phenotype for a QTL substitution effect of 0.5 (12.5% of total phenotypic variance).

(iii) ***Position estimates***

The mean bias (± s.e) of the estimated QTL location was computed over the 1,000 replicates and for each alternative of simulations. Full results are presented for a QTL substitution effect of 0.4 (*i.e.* 8% of total phenotypic variance) (Table 2).

[Table 2]

For all QTL mapping methods, and for all phenotypes and QTL effects, the bias was positive, meaning that the estimated position tends to be biased towards the centre of the chromosome, as it was already observed by other authors (e.g. Walling *et al.*, 2002). As shown in Table 2, for all three methods and for a QTL effect of 0.4, the bias is smaller for small proportions of ties (zeros) in the phenotype. The mean bias increases regularly as the proportion of zeros in the phenotypic distribution increases. This increase is higher for the non-parametric test using random ranking of ex-aequos (NP-RA) than for the test using the midranks approach (NP-MI) : the NP-MI method is 4 cM closer to the simulated position of the QTL than the NP-RA method. Furthermore, for high proportion of zeros, the mean bias of the NP-MI method is also lower than the mean bias of the RIM method (0.2, 0.4 and 0.4 cM for bacteria counts with 50%, 65% and 80% of zeros respectively).

The variation of the estimated QTL position according to the QTL effect and for the three QTL mapping methods is depicted by use of 95% confidence intervals of the mean position estimate over the 1,000 simulations, and for three proportions of zeros in the phenotype : 20%, 50% and 80% (Figure 3).

[Figure 3]

It has not escaped our notice that, for high proportion of zeros, the 95% confidence intervals do not contain the true value of the QTL position (35 cM), but this is due to the bias towards the centre of the chromosome when the QTL effect is small. Indeed, when the QTL effect is small (a = 0.1 or 0.2), the mean estimated QTL position is shifted to the centre of the chromosome (50 cM). Furthermore, the mean bias decreases until zero as the effect of the simulated QTL increases. The size of the confidence interval is also influenced by the QTL effect : the larger the effect, the smaller the interval.

For QTL effects larger than 0.2, the mean estimated position of the QTL using the NP-MI method was closer to the simulated position than using the NP-RA method. This results was especially observed for 80% of zeros (Figure 3), for which the mean location estimate of the NP-RA method was significantly (Bonferroni t tests) more biased towards the centre of the chromosome than the NP-MI and RIM methods (for QTL effects higher than 0.3).

(iv) ***Mean value of test statistics along the chromosome***

All three interval mapping methods clearly yield very similar curves when bacteria counts with small proportions of zeros are analysed (Figure 4).

[Figure 4]

However, when the proportion of zeros is higher than 20%, and as a result of the decreased power, the NP-RA method gives lower peaks than the NP-MI method.

4. DISCUSSION

In this study, we have adapted the non-parametric QTL mapping method based on the sum of ranks that was described previously by Kruglyak & Lander (1995*a*) and Coppieters *et al.* (1998) to the analysis of aggregated or discontinuous phenotypes (presenting high number of ties). This is particularly relevant for the analysis of disease resistance phenotypes. Indeed, the assessment of the disease resistance status of animals to bacterial, parasitic or viral diseases is often based on a counting process : *e.g.* colony-forming units (CFU) in infected organs (Berthelot *et al.*, 1998), nematode eggs in faeces (FEC) (Bouix *et al.*, 1998), or plaque-forming units (PFU) in peripheral blood cells (Vallejo *et al.*, 1998) are counted to measure the infection degree in animals.

The distribution of this type of traits is right-skewed with a long tail, and a high frequency of zero values is often observed, e.g. Bishop & Stear (2001). Such a feature comes from the difficulty of choosing appropriately the moment for sampling individuals, in order to have the most informative data set. Because of the high costs of such experiments, data cannot be thrown away, and need to be analysed, even if there is a high frequency of zeros in the phenotypic distribution.

Such distributions are neither continuous nor categorical. Furthermore, because of the excess of zeros and because of the high values taken by those counts, a Poisson structure cannot generally be assumed. In the context of parasitology, Wilson & Grenfell (1997) have proposed the use of generalised linear modelling with a negative binomial structure of errors for this type of distribution. However, up to now, to our knowledge, no QTL mapping methods were developed with such a

feature. Since the power of parametric and non-parametric QTL mapping methods is relatively low when analysing this type of trait, and since the non-parametric methods do not provide estimates of QTL effects, further research is still needed to set up efficient QTL mapping tools designed to deal with this type of trait.

By simulating increasing proportions of ties, we have shown that it is worth using midranks instead of random ranks when non-parametric interval mapping is chosen : power and precision were both increased when using midranks. For a QTL substitution effect of 0.4 (8% of phenotypic variance), a gain from 8 to 11% of power can be obtained. For higher levels of QTL effects and when there are 80% of zeros in the phenotype, this gain can reach the values of 27%, 37% and 43% respectively for QTL substitution effects of 0.5, 0.6 and 0.7 (Figure 1).

This gain can easily be explained by the fact that, when there is a large bunch of ties in the phenotype, new information is added in the data if ties are ranked at random : individuals with the same value get different ranks although they should have the same rank. As stated by Kendall & Stuart (1979), random ranking of ties has the merit of simplicity and needs no new theory, but obviously sacrifices information contained in the observations and may be expected to lead to loss of efficiency compared with the midranks approach. Other studies on the Wilcoxon rank-sum statistic already showed that when random ranks are used, the efficiency of the test is reduced (Putter, 1955; McNeil, 1967).

However, if the chance of having a group of ties in the phenotype is very low (*e.g.* true continuous distributions in which ties would occur only by chance, for example by the rounding of values), the gain of using midranks would be very

small. Indeed, the magnitude of the correction factor in the variance of the test statistic depends on the number of tied values and on the distribution of these values in different groups. For a given number of tied values, the contribution to the correction factor is bigger if they are all in the same group. This was already stated by Siegel & Castellan (1988) for the classic Wilcoxon rank-sum statistic.

For high proportions of zeros in the phenotype, results of the non-parametric interval mapping method using midranks were identical to the results of regression interval mapping (Knott *et al.*, 1996), while results of the non-parametric interval mapping using random ranks were significantly lower. Even though the RIM method perform equivalently to the newly modified NP-MI, there are situations in which the use of a non-parametric method should be preferred. Indeed, one can sometimes fail to find the appropriate mathematical transformation to normalise the data. Furthermore, other assumptions such as homogeneity of variances are not always fulfilled when a mathematical transformation is applied. Finally, parametric tests are more sensitive than non-parametric methods to the presence of outliers.

Instead of simulating increasing proportions of zeros, one could have simulated one group of ties at the end of the distribution by giving the same value to individuals with a phenotypic value higher than a specified threshold. Since it is the size of the group of ties which determines the correction factor of the variance of the test statistic, this would have had the same effect as one group of ties at the beginning of the distribution. Another alternative would have been to simulate one group of ties randomly positioned in the distribution, by giving the same value to individuals having a phenotypic value in a specified window. However, such a simulation of a

random bunch of ties would have had no obvious biological basis. The simulation of some small groups of ties randomly distributed in the phenotype, as observed when observations are rounded, was not considered since their contribution to the correction factor of the variance of the test statistic would have been too small. Finally, no categorical nor ordinal traits were simulated since specific methods exist to deal with this type of traits (Hackett & Weller, 1995; Visscher *et al.*, 1996).

In conclusion, it has to be mentioned that the midranks approach is easy to implement (Fortran code is available upon request). Therefore, in practice, if non-parametric interval mapping is chosen, one should favour the use of midranks to deal with tied values.

APPENDIX

A. *Variance of the non-parametric test statistic using random ranks for ties*

The aim is to obtain the variance formula of $Y_K(p)$, *i.e.* :

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

If $\left[N+1-2\,rank_j\right]$ is replaced by R_j and $\left(P_j(A)-P_j(B)\right)$ by X_j, in (A 1), one can simplify the notation and develop $Var\left(Y_K(p)\right)$ according to the property of the variance of a sum, as follows :

$$Var\left(Y_K(p)\right) = \sum_{j=1}^{N} Var\left(R_j X_j\right) + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ i\neq j}}^{N} Cov\left(R_i X_i, R_j X_j\right) \qquad \text{(A 2)}$$

which amounts to :

$$Var\left(Y_K(p)\right) = \sum_{j=1}^{N} Var\left(R_j X_j\right) \qquad \text{(A 3)}$$

This can be demonstrated by showing that $Cov\left(R_i X_i, R_j X_j\right)$ in (A 2) is equal to zero. Indeed, the covariance term is first developed as follows :

$$Cov\left(R_i X_i, R_j X_j\right) = E\left[R_i R_j X_i X_j\right] - E\left[R_i X_i\right] E\left[R_j X_j\right] \qquad \text{(A 4)}$$

Since R_i is independent of X_i, $E\left[R_i X_i\right] = E\left[R_i\right] E\left[X_i\right]$. Furthermore, $E\left[R_i X_i\right]$ amounts to zero since $E\left[X_j\right] = 0$; indeed :

$$E\left[X_j\right] = E\left[P_j(A) - P_j(B)\right] = E\left[P_j(A)\right] - E\left[P_j(B)\right] = 0.5 - 0.5 = 0 \qquad \text{(A 5)}$$

As, at marker location, an half-sib has equal chance of getting haplotype A or haplotype B from its sire, $P_j(A)$ and $P_j(B)$ take values 0 or 1. This means that half of the half-sibs would have $P_j(A)$=1 (*i.e.* $P_j(B)$=0) and the other half $P_j(A)$=0 (*i.e.* $P_j(B)$=1). Therefore, the expected value for $E\left[P_j(A)\right]$ and $E\left[P_j(B)\right]$ is 0.5. For more information, see similar applications by Kruglyak & Lander (1995*b*) and Spelman *et al.* (1996). Consequently, the covariance reduces to :

$$Cov\left(R_i X_i, R_j X_j\right) = E\left[R_i R_j X_i X_j\right] \qquad \text{(A 6)}$$

Since under the null hypothesis $R_i R_j$ is independent of $X_i X_j$:

$$Cov\left(R_i X_i, R_j X_j\right) = E\left[R_i R_j\right] E\left[X_i X_j\right] \qquad \text{(A 7)}$$

Since X_i is independent of X_j :

$$Cov\left(R_i X_i, R_j X_j\right) = E\left[R_i R_j\right] \underbrace{E\left[X_i\right]}_{0} \underbrace{E\left[X_j\right]}_{0} = 0 \qquad \text{(A 8)}$$

This independence implies that the probabilities (elements of X_j) are independent between half-sibs in the family under the null hypothesis, which can be assumed since genotypes of half-sibs at the markers are determined from successive binomial draws, taking into account recombination events. It is worth noting that the ranks R_i and R_j are not independent random variables, since ranks are drawn without replacement. However, $E\left[R_i R_j\right]$ is suppressed since $E\left[X_i\right] = 0$.

The variance of the product of R_j (*i.e.* $\left[N + 1 - 2\, rank_j\right]$) by X_j (*i.e.* $\left(P_j(A) - P_j(B)\right)$) is equal to :

$$Var\left(R_j X_j\right) = Var\left(R_j\right) Var\left(X_j\right) + E^2\left[R_j\right] Var\left(X_j\right) + E^2\left[X_j\right] Var\left(R_j\right) \qquad \text{(A 9)}$$

Ranks and probabilities of inheriting one of the sire haplotypes are independents under the null hypothesis. Based on the fact that $E\left[X_j\right] = E\left[P_j(A) - P_j(B)\right] = 0$ and that $E\left[R_j\right] = E\left[N + 1 - 2\, rank_j\right] = 0$, formula (A 9) simplifies to the product of both variances. Variance of $Y_K(p)$ is then :

$$Var\left(Y_K(p)\right) = \sum_{j=1}^{N} \left(Var\left(N + 1 - 2\, rank_j\right) Var\left(P_j(A) - P_j(B)\right)\right) \qquad \text{(A 10 } a\text{)}$$

$$= N\,Var\left(N+1-2\,rank_j\right)Var\left(P_j(A)-P_j(B)\right) \qquad \text{(A 10 } b\text{)}$$

$$= 4\,N\,Var\left(rank_j\right)Var\left(P_j(A)-P_j(B)\right) \qquad \text{(A 10 } c\text{)}$$

$$= 4\,N\,\frac{N^2-1}{12}Var\left(P_j(A)-P_j(B)\right) \qquad \text{(A 10 } d\text{)}$$

$$= \left(\frac{N^3-N}{3}\right)Var\left(P_j(A)-P_j(B)\right) \qquad \text{(A 10 } e\text{)}$$

Note that in (A 10 *c*), $Var\left(Y_K(p)\right)$ depends on $Var\left(rank_j\right)$, the variance of the ranks. This property will be used in Appendix B.

B. *Variance of the non-parametric test statistic using midranks for ties*

A correction for the variance of the ranks is proposed in the Appendix of Lehman's book, p. 329 Lehmann (1975). Suppose that the *N* observations take on *e* distinct values, and that d_1 of the *N* observations are equal to the smallest value, d_2 to the next smallest, …, d_e to the largest. With this notation :

$$Var\left(rank_j^*\right) = \frac{N^2-1}{12} - \frac{\sum_{i=1}^{e} d_i\left(d_i^2-1\right)}{12N} \qquad \text{(A 11)}$$

The first term is the variance of the ranks, the second gives the correction for ties. When no ties are present, all the d_i are equal to 1, and the correction term is zero as it should be. Consequently, the corrected variance formula of $Y_K(p)$ is easily found by introducing (A 11) in (A 10 *c*) as follows :

$$Var\left(Y_K^*(p)\right) = N\,4\left(\frac{N^2-1}{12} - \frac{\sum_{i=1}^{e} d_i\left(d_i^2-1\right)}{12N}\right) Var\left(P_j(A) - P_j(B)\right) \qquad \text{(A 12)}$$

$$Var\left(Y_K^*(p)\right) = \frac{1}{3}\left[\left(N^3-N\right) - \sum_{i=1}^{e} d_i\left(d_i^2-1\right)\right] Var\left(P_j(A) - P_j(B)\right) \qquad \text{(A 13)}$$

which amounts to the previous variance formula in the absence of ties.

ACKNOWLEDGMENTS

We are grateful to Frédéric Lantier for providing us the bacteria counts data, to Frédéric Farnir for helpful suggestions, and to Jérémie Nsengimana for his comments on an earlier version of the paper. Financial support is acknowledged from the Université catholique de Louvain, from the European Union (FAIR CT-98-4311), from the Belgian National Fund for Scientific Research (FNRS), from the contract `Projet d'Actions de Recherche Concertées' nr 98/03--217 from the Belgian government, and from the IAP research network nr P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs).

REFERENCES

Baret, P. V., Knott, S. A. & Visscher, P. M. (1998). On the use of linear regression and maximum likelihood for QTL mapping in half-sib designs. *Genetical Research* **72**, 149-158.

Berthelot, F., Beaumont, C., Mompart, F., Girard-Santosuosso, O., Pardon, P. & Duchet-Suchaux, M. (1998). Estimated heritability of the resistance to cecal carrier state of Salmonella enteritidis in chickens. *Poultry Science* **77**, 797-801.

Bishop, S. C. & Stear, M. J. (2001). Inheritance of faecal egg counts during early lactation in Scottish Blackface ewes facing mixed, natural nematode infections. *Animal Science* **73**, 389-395.

Bouix, J., Krupinski, J., Rzepecki, R., Nowosad, B., Skrzyzala, I., Roborzynski, M., Fudalewicz-Niemczyk, W., Skalska, M., Malczewski, A. & Gruner, L. (1998). Genetic resistance to gastrointestinal nematode parasites in Polish long-wool sheep. *International Journal for Parasitology* **28**, 1797-1804.

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

Coppieters, W., Kvasz, A., Farnir, F., Arranz, J. J., Grisart, B., Mackinnon, M. & Georges, M. (1998). A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: Application to milk production in a granddaughter design. *Genetics* **149**, 1547-1555.

Georges, M. & Coppieters, W. (2000). Introductory course to QTL mapping in outbred half-sib pedigrees. *Lecture notes of the CGIL Summer Course 1* , June 5-9 (http://apsit.aps.uoguelph.ca/QTL/TitlePage.htm).

Hackett, C. A. & Weller, J. I. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1252-1263.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315-324.

Kendall, M. & Stuart, A. (1979). *The advanced theory of statistics. Vol 2. Inference and relationship*, 4th edn. London: Charles Griffin & Company Limited.

Knott, S. A., Elsen, J. M. & Haley, C. S. (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93**, 71-80.

Kruglyak, L. & Lander, E. S. (1995*a*). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421-1428.

Kruglyak, L. & Lander, E. S. (1995*b*). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439-454.

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

Lehmann, E. L. (1975). *Nonparametrics - Statistical methods based on ranks*. Holden-Day Series in Probability and Statistics, ed. E. L. Lehmann. San Francisco: Holden-Day, Inc.

Martinez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480-488.

McNeil, D. R. (1967). Efficiency loss due to grouping in distribution-free tests. *Journal of the American Statistical Association* **62**, 954-965.

Putter, J. (1955). The treatment of ties in non-parametric tests. *Annals of Mathematical Statistics* **26**, 368-386.

Siegel, S. & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. Singapore: McGraw-Hill.

Spelman, R. J., Coppieters, W., Karim, L., van Arendonk, J. A. M. & Bovenhuis, H. (1996). Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. *Genetics* **144**, 1799-1807.

Tilquin, P., Coppieters, W., Elsen, J. M., Lantier, F., Moreno, C. & Baret, P. V. (2001). Statistical power of QTL mapping methods applied to bacteria counts. *Genetical Research* **78**, 303-316.

Vallejo, R. L., Bacon, L. D., Liu, H. C., Witter, R. L., Groenen, M. A. M., Hillel, J. & Cheng, H. H. (1998). Genetic mapping of quantitative trait loci affecting susceptibility to Marek's disease virus induced tumors in F2 intercross chickens. *Genetics* **148**, 349-360.

Visscher, P. M., Haley, C. S. & Knott, S. A. (1996). Mapping QTLs for binary traits in backcross and F2 populations. *Genetical Research* **68**, 55-63.

Walling, G. A., Haley, C. S., Perez-Enciso, M., Thompson, R. & Visscher, P. M. (2002). On the mapping of quantitative trait loci at marker and non-marker locations. *Genetical Research* **79**, 97-106.

Weller, J. I., Kashi, Y. & Soller, M. (1990). Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of Dairy Science* **73**, 2525-2537.

Wilson, K. & Grenfell, B. T. (1997). Generalized linear modelling for parasitologists. *Parasitology Today* **13**, 33-38.

Table 1. *Power (%) at the 5% significance level for a range of proportions of zeros in bacteria counts and a QTL substitution effect of 0.4 (h^2_{QTL}=0.08)*

Phenotype	Proportion of zeros	NP-RA	NP-MI	*log*+ RIM
Normal		55 [a]	55 [a]	60 [a]
Bacteria counts	0.0%	55 [a]	55 [a]	59 [a]
	8.5%	55 [a]	55 [a]	57 [a]
	20.0%	54 [a]	54 [a]	56 [a]
	35.0%	49 [a]	51 [a]	53 [a]
	50.0%	38 [a]	46 [b]	47 [b]
	65.0%	26 [a]	37 [b]	37 [b]
	80.0%	13 [a]	24 [b]	25 [b]

Power estimates based on 1,000 replicates of each alternative (phenotype and method), using 1,000 permutations to obtain the significance level of each replicate. Heritability of the trait was 0.25.

[a] Comparison of power estimates between methods within phenotype (tests for the comparison of two proportions based on a Normal approximation); values with the same superscript are not significantly different, $P > 0.0167$ (Bonferroni correction of $P > 0.05$).

Table 2. *Mean bias (± s.e.) of the estimated QTL location (cM), i.e. the difference between the mean location estimate over the 1,000 replicates and the true position (35 cM), for a QTL substitution effect of 0.4 (h^2_{QTL} = 0.08) and for a range of proportions of zeros in bacteria counts*

		Method		
Phenotype	Prop. of zeros	NP-RA	NP-MI	*log*+RIM
Normal		3.4 ± 0.6 [a]	3.5 ± 0.6 [a]	3.6 ± 0.6 [a]
Bacteria counts	0.0%	3.5 ± 0.6 [a]	3.5 ± 0.6 [a]	3.3 ± 0.6 [a]
	8.5%	3.5 ± 0.6 [a]	3.4 ± 0.6 [a]	3.1 ± 0.6 [a]
	20.0%	3.8 ± 0.6 [a]	3.5 ± 0.6 [a]	3.5 ± 0.6 [a]
	35.0%	4.8 ± 0.7 [a]	4.1 ± 0.6 [a]	3.9 ± 0.6 [a]
	50.0%	5.0 ± 0.7 [a]	3.9 ± 0.7 [a]	4.1 ± 0.6 [a]
	65.0%	6.9 ± 0.8 [a]	5.3 ± 0.8 [a]	5.7 ± 0.7 [a]
	80.0%	10.6 ± 0.9 [a]	6.6 ± 0.8 [b]	7.0 ± 0.8 [b]

Heritability of the trait was 0.25; one QTL simulated in position 35 cM on a 100 cM chromosome.

[a] Comparison of mean bias of estimated QTL location (Bonferroni *t* tests) between methods (NP-RA, NP-MI or *log*+RIM) within phenotype (lines); values with the same superscript are not significantly different, $P > 0.0167$ (Bonferroni correction of $P > 0.05$).

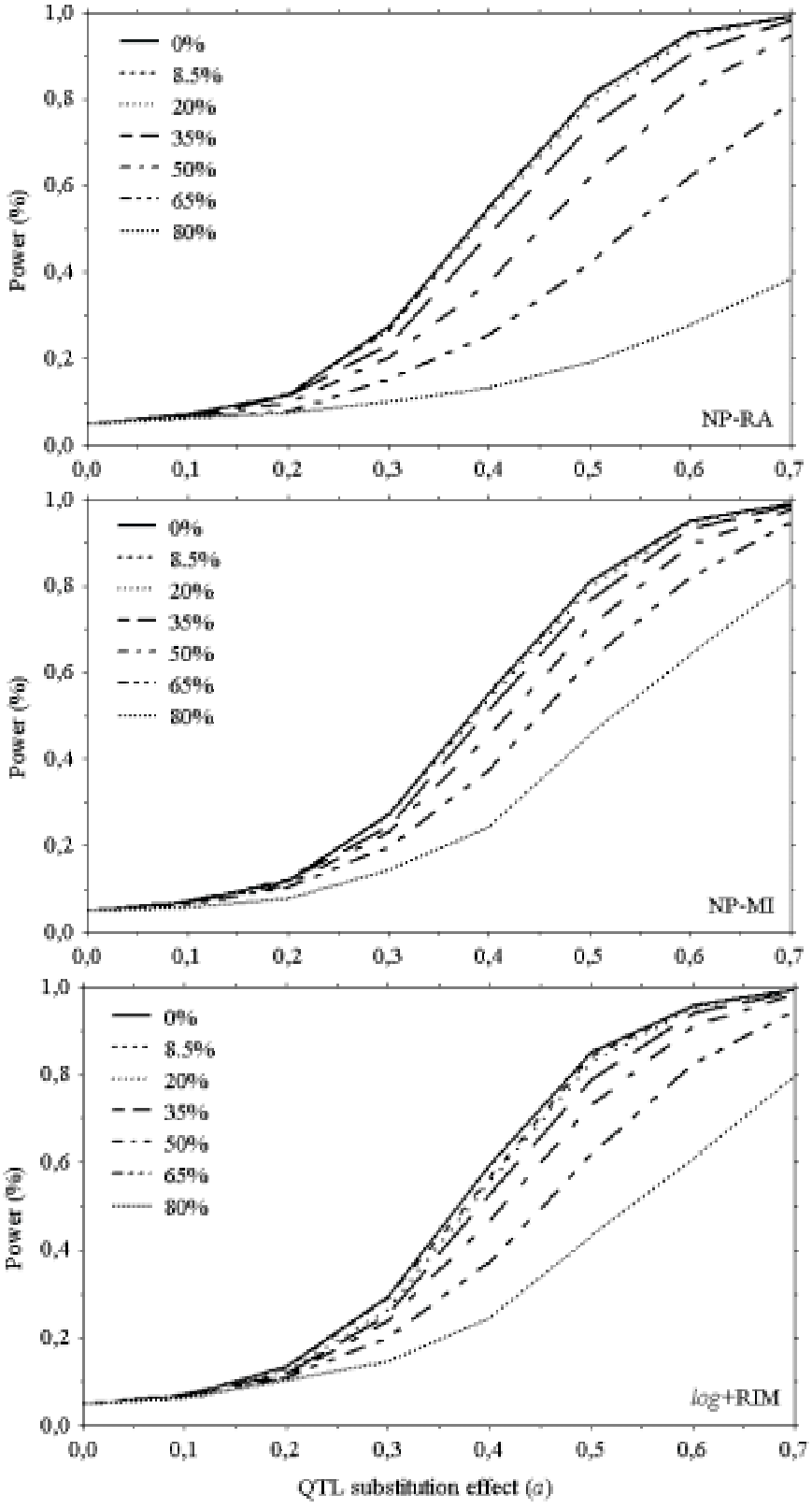
CAPTIONS OF FIGURES

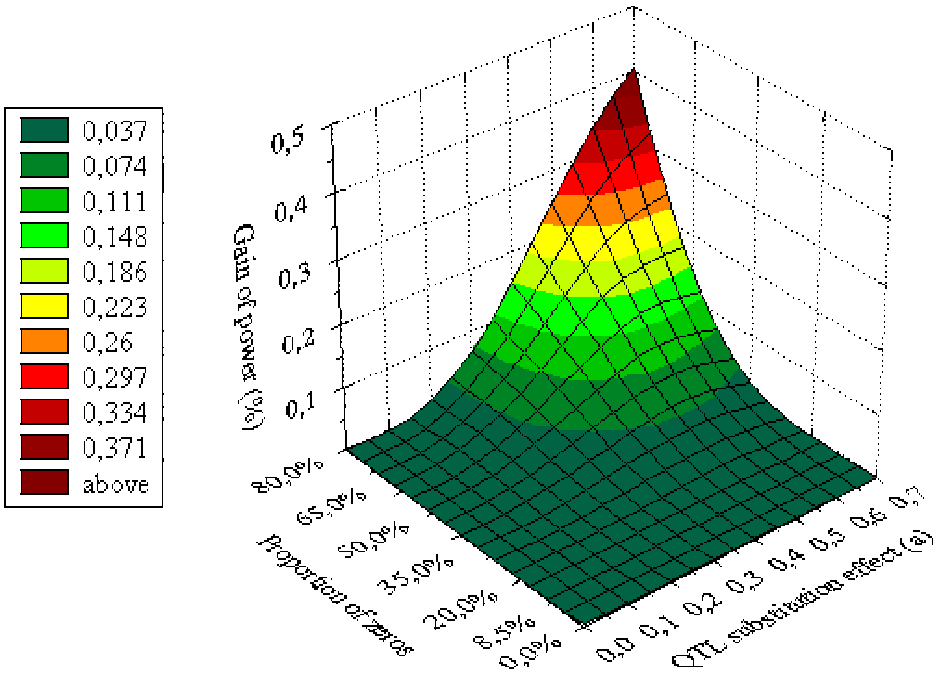
Figure 1. Power (%) of NP-RA, NP-MI and *log*+RIM methods as a function of the QTL substitution effect (*a*) and of the proportion of zeros in bacteria counts.

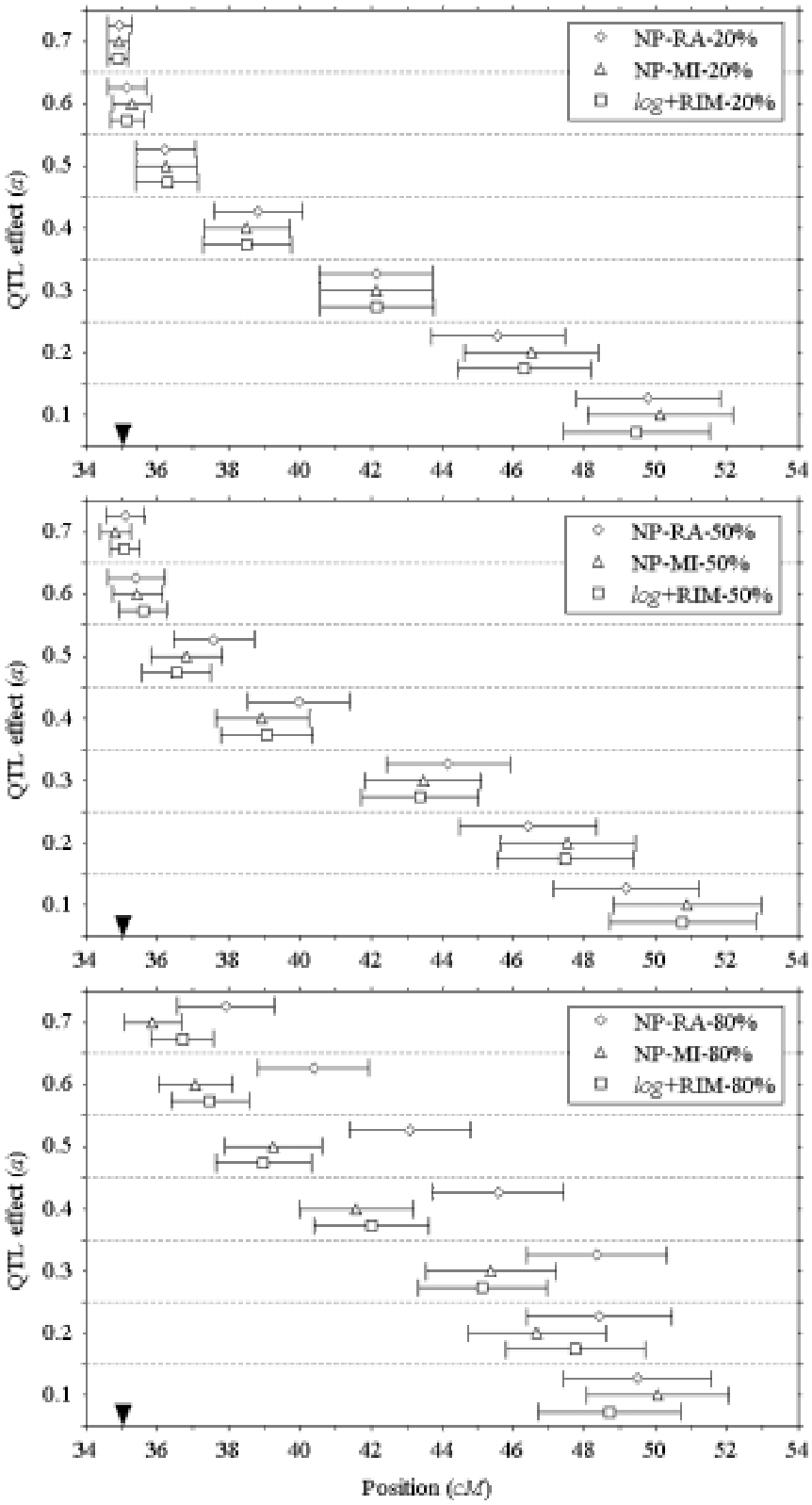
Figure 2. Absolute gain of power (%) using the midrank approach (NP-MI) instead of ranking ties at random (NP-RA) as a function of the QTL substitution effect (*a*) and of the proportion of zeros in bacteria counts.

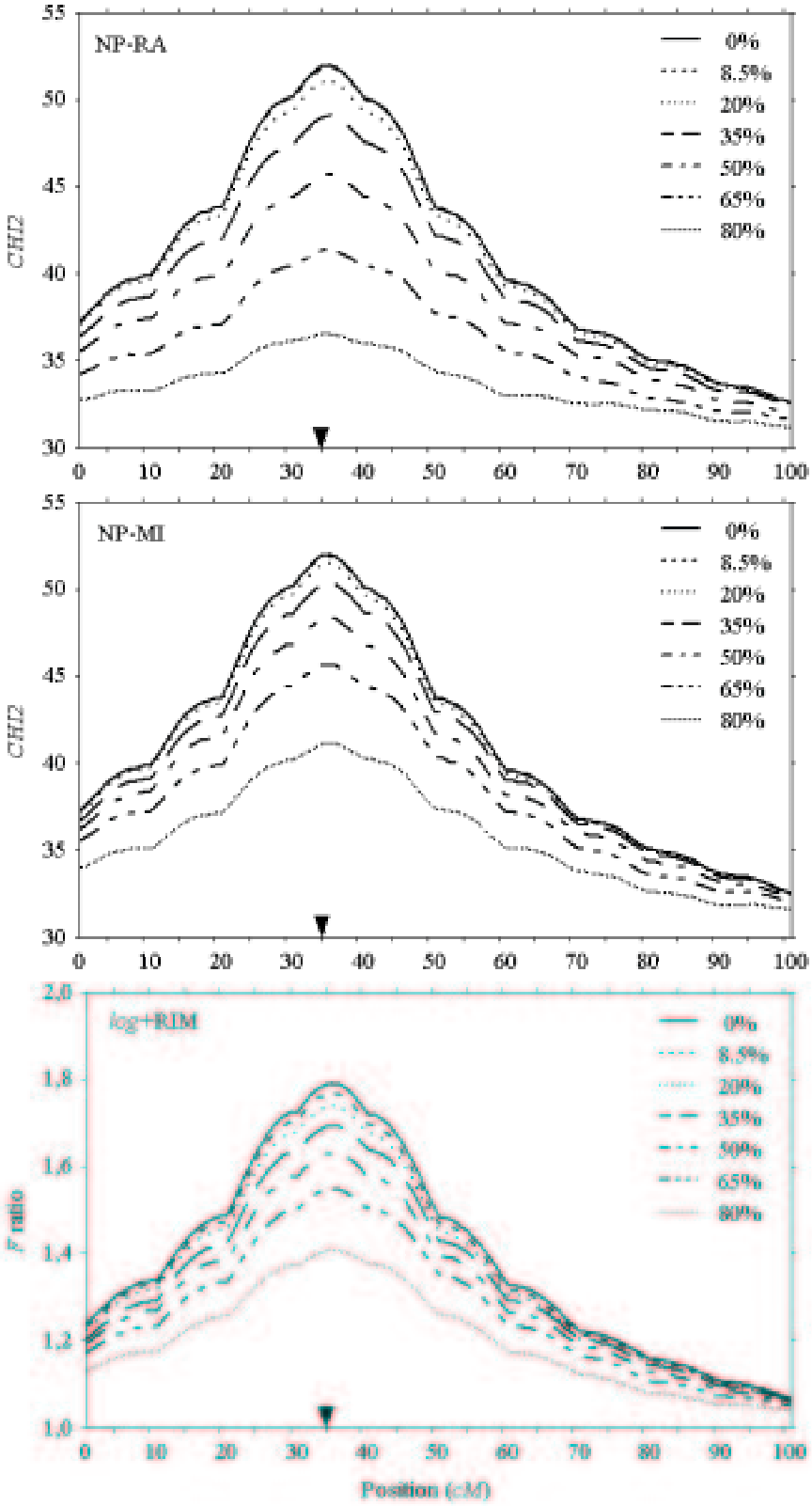
Figure 3. Confidence intervals of the mean estimated QTL position as a function of the QTL substitution effect (a) and when bacteria counts with 20%, 50% or 80 % of zeros are analysed. Arrowheads indicate the location of the simulated QTL.

Figure 4. Mean test statistic value (1,000 replicates) along the chromosome for the three QTL mapping methods when bacteria counts are analysed (for NP-RA and NP-MI) and when *log*-transformed bacteria counts are analysed (for RIM) and for a QTL substitution effect of 0.4 (h^2_{QTL} = 0.08). Arrowheads indicate the location of the simulated QTL.


0%
8.5%
20%
35%
50%
65%
80%
Power (%)
NP-RA
NP-MI
log+RIM
QTL substitution effect (a)
0,0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
1,0


0,037
0,074
0,111
0,148
0,186
0,223
0,26
0,297
0,334
0,371
above
Gain of power (%)
0,5
0,4
0,3
0,2
0,1
80,0%
65,0%
50,0%
35,0%
20,0%
5,0%
0,0%
proportion of zeros
0,0 0,1 0,2 0,3 0,4 0,5 0,6 0,7
QTL substitution effect (a)


NP-RA-20%
NP-MI-20%
log+RIM-20%
NP-RA-50%
NP-MI-50%
log+RIM-50%
NP-RA-80%
NP-MI-80%
log+RIM-80%
QTL effect (a)
Position (cM)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
34
36
38
40
42
44
46
48
50
52
54


NP-RA
0%
8.5%
20%
35%
50%
65%
80%
CHI2
NP-MI
log+RIM
F ratio
Position (cM)