Science Publications

American Journal of Agricultural and Biological Science 2(3): 195-201, 2007
ISSN 1557-4989

QTL Mapping Using Multiple Markers Simultaneously

D. Kolbehdari and J. A. B. Robinson
Center for Genetic Improvement of Livestock, Department of Animal and Poultry Science,
University of Guelph, Guelph, Ontario N1G 2W1, Canada

Abstract: Methods for detecting and locating a single QTL within a 10 cM region of DNA using 10 equally spaced SNP markers were compared. The QTL was assumed to be bi-allelic and located between markers 5 and 6. Monte Carlo simulation of a granddaughter design with 30 sires and 400 sons was used. Linear regression nested within sire using either two or four marker haplotypes at a time was used. In addition, the scoring of haplotype transmissions from sire to sons were varied in three ways. Another method assumed linkage disequilibrium and estimated haplotype interval effects for all intervals simultaneously. Other variables compared were the ratio of QTL variance to total genetic variance, the number of generations of historical recombination, and frequencies of marker alleles. Empirical power was dependent on the scoring method in the linear regression method. Four-marker haplotypes gave slightly higher empirical power than two-marker haplotypes. Reducing the proportion of QTL variance decreased empirical power. Empirical power was greater for 25 generations of historical recombination over 100 generations. Empirical power was lower when marker allele frequencies were 0.8 compared to 0.5. The linkage disequilibrium model gave results similar to those of the linear regression model.

Keywords: QTL mapping, linear regression, simultaneous haplotypes model

INTRODUCTION

SNP are commonly used for detection and localization of QTL for complex traits. Several algorithms for identifying haplotypes using SNPs have been developed [6, 10]. In practice, SNP genotypes will be available without phase information and the most frequent haplotypes need to be determined [1, 8]. When there are enough offspring genotypes, as in half-sib designs in dairy cattle, then sire haplotypes can be reconstructed with near certainty. Determining the haplotype of the sire that is inherited by the son is difficult if the sire is heterozygous.

Detection and localization of QTL in livestock populations have used single markers, multiple markers, and haplotype-based methods analyzed with least squares or maximum likelihood procedures. Probabilities of QTL alleles being identical by descent (IBD) have been used with variance component estimation approaches [2, 3]. Haplotype and linkage disequilibrium (LD) studies covering small chromosomal regions assume that sufficient generations have passed such that in small regions, the marker and QTL alleles tend to segregate together [5].

Linear regression has been used with genotypes of single and multiple markers[4]. This approach uses linkage information and follows segregation of markers and QTL within sire families, but the use of haplotypes and linkage disequilibrium information could provide better precision in locating QTLs.

The objective of this study was to estimate QTL location within a group of markers using all marker haplotypes simultaneously. The method is assumes linkage disequilibrium and the needs to know the transmission of haplotypes from sire to sons. In a simulated granddaughter design the method was compared to linear regression with single and multiple marker methodology.

MATERIALS AND METHODS

The Simulation Method: A genomic structure of ten bi-allelic SNPs (each 1 cM from the next) with a single bi-allelic QTL located between the 5th and 6th markers was assumed. Initial allele frequencies, p, were either 0.5 or 0.8 for both markers and QTL. The base population was assumed to be in Hardy-Weinberg equilibrium at each locus and animals unrelated. Initial haplotypes for each sire and dam were generated independently. The number of sires and dams per generation were 30 and 400, respectively, giving an

Corresponding Author: Sumpun Chaitep, Mechanic and Agricultural Engineering Department, Faculty of Engineering, Chiang Mai University, Thailand 50200, Tel: (66 53) 942005, Fax: (66 53) 941352

effective population size of 100. Progeny were generated using the Haldane mapping function (no interference) to determine crossover events between markers and QTL in deriving sire and dam haplotypes. Each mating produced 4 progeny. Each generation was discrete with new males and females randomly selected from the group of progeny.

Linkage disequilibrium was created over either 25 or 100 generations of random matings. In the last generation, 10 males were randomly selected and mated to 1000 females to produce 100 sons per sire. Each son was assumed to have 100 progeny. Progeny Yield Deviations (DYD) were created for each son. A DYD is the average phenotype of the progeny adjusted for systematic environmental effects and the genetic values of the dams [7].

$$DYD = 0.5BV_{son} + e \quad (1)$$

where BV_{son} is the breeding value of the son, and e is a residual effect with mean zero and variance equal to residual variance divided by number of offspring (100 in this study). The trait was assumed to have a heritability of 0.3 and σ_P =100.

Parameters allowed to vary were the ratio of QTL variance to total additive genetic variance (k=0.1 or 0.05), the allele frequencies of the SNPs and QTL (p=0.5 or 0.8), and the number of generations of historical recombination (g=100 or 25).

Linear Regression Method: The model was

$$y_{ij} = x_{ij}\beta_i + s_i + e_{ij}, \quad (2)$$

where y_{ij} is the DYD of son j of sire i, x_{ij} indicates the transmission score of the haplotype from sire i to son j, β_i is the within sire family regression coefficient, s_i is the polygenic effect of the sire, and e_{ij} is the residual effect. The assumption was made that only sires and sons were genotyped for the markers. The scores for x_{ij} normally used by Knott *et al.* (1996) are shown in Table 1. Note that progeny of homozygous sires are not used. This method uses linkage information and segregation of marker alleles within heterozygous families. In this study, two different sets of transmission scores were used and included progeny of both heterozygous and homozygous sires. The two sets of scores differed from each other by two numbers as shown in Table 1. When a sire is homozygous, then his progeny have the same haplotype, and so the transmission score in Set 1 was set to 1. With Set 2, the transmission score was set to 2 for homozygous sires with homozygous sons.

Haplotypes were created using either two contiguous markers at a time (9 sets of two) denoted as method LR-2, or four contiguous markers at a time (7 sets of four) denoted as method LR-4. The analyses proceeded from one marker set to the next, and F-ratios were used to determine the location of the QTL. Transmission scores were applied to the haplotype of the set of markers.

Simultaneous Haplotype Model: The simultaneous haplotype model,

$$y_{ijkl} = \mu + h_{ij} + s_k + e_{ijkl} \quad (3)$$

where y_{ijkl} is the DYD of the l^{th} son of sire k, μ is the overall mean, h_{ij} is the j^{th} haplotype of i^{th} contiguous pair (i=1 to 9 pairs, j=1 to 4 possible haplotypes within pairs, 36 effects), s_k is the polygenic effect of the sire, e_{ijkl} is a residual effect, σ_e^2 and σ_h^2 were the same for all marker set. Haplotypes were constructed based on each pair of contiguous markers, giving 9 non-overlapping haplotype intervals. For each interval there were 4 possible haplotypes that could be inherited from sire to son, and therefore, a total of 36 haplotype effects. If the haplotype coming from the sire could not be determined, the design matrix contained all zeros for that interval for that son. Otherwise there was a one corresponding to the haplotype that was inherited from the sire in the design matrix.

In matrix notation, the model is

$$\mathbf{y} = \mu + \mathbf{h} + \mathbf{s} + \mathbf{e} \quad (4)$$

and the mixed model equations are

$$\begin{pmatrix} \mathbf{Z'Z} + \mathbf{I}\alpha & \mathbf{Z'X} \\ \mathbf{X'Z} & \mathbf{X'X} \end{pmatrix} \begin{pmatrix} \mathbf{h} \\ \mu + \mathbf{s} \end{pmatrix} = \begin{pmatrix} \mathbf{Z'y} \\ \mathbf{X'y} \end{pmatrix} \quad (5)$$

$\alpha = 1.0$

To locate the QTL, the sum of absolute values of estimated haplotype effects within each haplotype pair was calculated,

$$H_i = \sum_{j=1}^{4} \left| \hat{h}_{ij} \right| \quad (6)$$

The middle of the haplotype interval with the largest H_i was taken as the location of the QTL, and the $\hat{h}_{ij}$ within the largest H_i indicated the magnitude of the QTL effects.

The ratio of residual to haplotype variances was fixed at 1.0 for all pairs of haplotypes. However, the ratios could differ between pairs, such that the ratio would be larger for pairs of markers that do not contain a QTL. Thus, simulations were repeated where the ratio for marker pairs without the QTL was 10, and for the

marker pair containing the QTL was one. No attempts were made to estimate the ratio in this study.

The number of marker pairs could also have an effect on QTL detection, and so the simulations were repeated using 20 markers (with all variance ratios equal to 1) within a 10 cM chromosome segment. The QTL was placed in the middle of the markers. Thus, there were 76 haplotype effects to be estimated in this scenario.

Comparison Statistics: Let Q be the true position of a QTL and $\hat{Q}$ be the estimated position. The estimated position was the mid point of the marker haplotype interval that gave the largest F-ratio or the largest *H*. Bias of estimation was

$$Bias = \frac{\sum(\hat{Q} - Q)}{n} \quad (7)$$

where *n* is the number of replicates, and the mean squared error was

$$MSE = \frac{\sum(\hat{Q} - Q)^2}{n} \quad (8)$$

Precision, (*P*), was

$$P = \sqrt{MSE - Bias^2}\,. \quad (9)$$

The critical values of the test statistic corresponding to 5% type I error were determined empirically by simulating 1000 replicates based on the null hypothesis of no QTL segregating on the chromosome segment. Another 1000 replicates were simulated under the alternative hypothesis, with the QTL situated in the center of the markers, and the empirical power was computed as the proportion of replicates in which the test statistic exceeded the critical value.

Scenarios: Four scenarios of parameter combinations were studied. Scenario 1 had the ratio of QTL variance to total genetic variance of 0.1, used 100 generations of historic recombination, and started with an allele frequency of 0.5 for all markers and the QTL. Scenario 2 had the ratio of QTL variance to total genetic variance of 0.05, and other parameters were the same as Scenario 1. Scenario 3 was the same as Scenario 1 with the number of generations of historic recombination reduced to 25. Scenario 4 was the same as Scenario 1 with the allele frequency of markers and QTL increased to 0.8.

RESULTS

Linear Regression Two-marker Intervals: Under Scenario 1 (the base combination of parameters), using Set 1 transmission scores, empirical power was 0.77, and using Set 2 the power was 0.82 (Table 2). The only difference was in the scores assigned to a homozygous son from a homozygous sire. Precision was better with Set 2 (1.33 cM versus 1.37 cM for Set 1). However, the probability of a type I error was smaller with Set 1. Results are shown in Fig. 1 for the two sets of indicator variables and for the Knott *et al.* (1996) scores for comparison. With scores of Knott *et al*. (1996), the QTL position could not be detected because the average of 1000 replicates of F-values was below the critical value (1.83).

In Scenario 2, the QTL variance was only half as large as in Scenario 1, and this reduced power, decreased precision, and increased the probability of type I error, but not significantly compared to Scenario 1. Set 2 scores gave better results in this scenario also. Smaller QTL effects gave generally poorer results for both sets of scores.

In Scenario 3, the number of generations of historical recombination was changed from 100 to 25. Empirical power increased from 0.77 to 0.95 with Set 1 indicator variables and from 0.82 to 0.98 with Set 2. Precision was below 1 cM for both sets of scores. Probabilities of type I errors were only 0.04 and 0.10 for Sets 1 and 2, respectively. With 100 generations of random mating, there was more chance for the marker (and QTL) alleles to drift to fixation, than within 25 generations. This might explain the differences observed.

In Scenario 4, allele frequencies of markers and QTLs were changed from 0.5 to 0.8. Empirical power decreased from 0.77 to 0.54 and 0.50 for Sets 1 and 2, respectively, and the QTL location was less precise at over 1.35 cM, but the probability of type I error was very low. Bias and MSE were much greater than the other scenarios. Set 2 scores had slightly smaller Bias and MSE than Set 1, but lower power and poorer precision.

Linear Regression Four-marker Intervals: When four markers were used to establish haplotypes for the linear regression model, results are given in Table 3. Degrees of freedom were expected to be less because of fewer intervals being considered. Empirical power was the same or greater than with the two-marker intervals, but the probabilities of type I errors were slightly greater. Precision was better in all scenarios with four markers rather than two.

Comparison of scenarios followed the same pattern as for the two-marker intervals. However, the precision, bias, and MSE in scenario 4 when the allele frequencies changed from 0.5 to 0.8 were better with the four-marker intervals than with the two-marker intervals. The best scenario was 3 having only 25 generations of historical recombination.

Table1: Two sets of indicators values for the transmission first haplotype of sires to progeny

Sire Haplotype		Son Haplotype		
		$H_i H_i$	$H_i H_x$	$H_x H_x$
Knott *et al.*	$H_i H_i$	0	0	0
	$H_i H_x$	1	0.5	0
Set1	$H_i H_i$	1	1	0
	$H_i H_x$	1	0.5	0
Set2	$H_i H_i$	2	1	0
	$H_i H_x$	1	0.5	0

Table 2: Comparison of four scenarios using linear haplotype regression with two sets of indicator variables using two-marker intervals

	Scenarios			
	1	2	3	4
QTL Ratio	0.1	0.05	0.1	0.1
Generations	100	100	25	100
Allele Freq	0.5	0.5	0.5	0.8
		Set 1 Scores		
Power	0.77	0.76	0.95	0.54
Pr (Type 1)	0.02	0.18	0.04	0.001
Bias (cM)	1.48	1.69	0.91	2.57
MSE (cM^2)	4.06	4.86	1.93	8.49
Precision (cM)	1.37	1.41	0.98	1.38
		Set 2 Scores		
Power	0.82	0.81	0.98	0.50
Pr (Type 1)	0.07	0.24	0.10	0.001
Bias (cM)	1.33	1.41	0.72	2.37
MSE (cM^2)	3.55	3.81	1.27	7.62
Precision (cM)	1.33	1.35	0.88	1.41

Table 3: Comparison of four scenarios using linear haplotype regression with two sets of indicator variables using four-marker intervals

	Scenarios			
	1	2	3	4
QTL Ratio	0.1	0.05	0.1	0.1
Generations	100	100	25	100
Allele Freq	0.5	0.5	0.5	0.8
		Set 1 Scores		
Power	0.82	0.78	0.95	0.54
Pr (Type 1)	0.10	0.24	0.07	0.001
Bias (cM)	1.54	1.59	1.35	1.91
MSE (cM^2)	3.31	3.63	2.61	4.75
Precision (cM)	0.98	1.03	0.90	1.06
		Set 2 Scores		
Power	0.85	0.82	0.98	0.64
Pr (Type 1)	0.15	0.30	0.14	0.001
Bias (cM)	1.42	1.55	1.23	1.80
MSE (cM^2)	2.90	3.37	2.23	4.44
Precision (cM)	0.95	0.99	0.85	1.10

Table 4: Comparison of four scenarios using simultaneous model with two-marker intervals

	Scenarios			
	1	2	3	4
QTL Ratio	0.1	0.05	0.1	0.1
Generations	100	100	25	100
Allele Freq	0.5	0.5	0.5	0.8
Power	0.84	0.79	0.92	0.63
Pr (Type 1)	0.02	0.11	0.03	0.01
Bias (cM)	1.68	1.84	1.89	2.16
MSE (cM^2)	4.58	5.21	5.32	6.51
Precision (cM)	1.33	1.35	1.32	1.36

Table 5: Comparison of four scenarios using simultaneous model with two-marker intervals and the variance ratio was set to 1 for interval with highest H_i, and set to 10 for all other intervals

	Scenarios			
	1	2	3	4
QTL Ratio	0.1	0.05	0.1	0.1
Generations	100	100	25	100
Allele Freq	0.5	0.5	0.5	0.8
Power	0.79	0.65	0.80	0.60
Pr (Type 1)	0.06	0.18	0.04	0.02
Bias (cM)	1.71	1.91	1.93	2.18
MSE (cM^2)	4.52	5.33	5.34	6.47
Precision (cM)	1.26	1.30	1.28	1.36

Table 6: Comparison of four scenarios using simultaneous model with two-marker intervals with 20 markers along a 10cM

	Scenarios			
	1	2	3	4
QTL Ratio	0.1	0.05	0.1	0.1
Generations	100	100	25	100
Allele Freq	0.5	0.5	0.5	0.8
Power	0.92	0.84	0.96	0.99
Pr (Type 1)	0.01	0.08	0.01	0.005
Bias (cM)	1.54	1.84	1.86	2.23
MSE (cM^2)	4.06	5.32	5.21	7.01
Precision (cM)	1.29	1.39	1.33	1.42

Simultaneous Haplotype Model, Two-marker Intervals: Tables 4 contain the results from the simultaneous haplotype model, applied to the four scenarios of parameter combinations. A plot of the H_i values for scenario 1 for a typical replicate shows the indication of QTL location (Fig. 2). Power for this model was similar to the linear regression model using four marker intervals and Set 1 scores. Probability of type I errors was similar to the linear regression model with two marker intervals and Set 1 scores, except for Scenario 4 where the simultaneous haplotype model was higher. Bias and MSE were greater than either linear regression approach and precision was about the same as the linear regression model with two marker intervals, except for Scenario 3, which gave poorer precision with the simultaneous haplotype model. The simultaneous haplotype model has the advantage that all marker intervals are considered at one time, compared to linear regression where only one marker interval at a time is analyzed. At the same time, this is a

disadvantage in that more effects are being estimated at one time. The net result is that the simultaneous haplotype model does not appear to be better than the linear regression model.

DISCUSSION

The regression model of Knott *et al*. (1996) was not successful in detecting QTL within small regions of 1 to 5 cM. The method works better for large regions of 10 to 20 cM. However, by regression of transmitted haplotypes instead of genotypes, as in this study, and assuming linkage disequilibrium (LD), QTL can be detected within 1 cM regions, if the QTL exists, with considerable power and precision.

In this study the positions and order of the markers need to be known. In practice, the markers will likely not be equally spaced. One possibility is to use SNP genotypes rather than haplotypes assuming linkage disequilibrium exists, and to put 5 or more SNP markers within each cM segment, such that the markers are very near or even within a QTL. The SNP genotypes then approximate the QTL genotypes. The order of the SNPs would not need to be known, nor the distances. The SNP location giving the largest estimated effects would most likely be the location closest to the QTL.

The size of QTL effect is important in accurate mapping as expected, and also the allele frequencies for the QTL. Decreases in empirical power in this study were non significant in going from 10% to 5% of total genetic variance. The decrease in power was much less than in using the simple regression method or variance component method as in [3, 4].

Empirical power of QTL detection increased when the number of generations of historical recombination decreased from 100 to 25. After 100 generations of recombination in a small population, most loci would have drifted to become homozygous and the informativeness of the haplotypes would decrease. Thus, effective population size may have some importance on QTL detection. With 25 generations of recombination, less homozygosity of loci was allowed to occur and the informativeness of haplotypes was consequently greater. There was no allowance for mutations to occur in order to maintain QTL genetic variability.

The most important aspect of the regression methods in this study was the scores used in the regression model. A preliminary study comparing seven different sets of scores was made. The better two from that study were used here. Sets 1 and 2 were not significantly different in performance, but Set 2 gave slightly greater empirical power and better precision. Set 1 was very close to conditional probabilities of inheritance of the sire haplotype, dependent on whether the sire was homozygous or heterozygous for the haplotype. With the Knott *et al*. (1996) method, only one set of scores was given whether the sire was heterozygous or homozygous for the haplotype alleles. There could be an optimum set of scores that could be derived, which might be dependent on the frequencies of different haplotypes.

The simultaneous haplotype model assumes the existence of linkage disequilibrium (LD), and many generations of historic recombination. LD can also be created by selection, but all replacements and matings were random in this simulation. The haplotype interval effects were assumed to be random effects, and in this study a common ratio of residual to QTL variance of 1 was used for all intervals. A comparison was made where the variance ratio was set to 1 for the interval with the highest H_i, and set to 10 for all other intervals (Table 5). Compared to using a ratio of 1 for all intervals, power was lower and bias was slightly greater, while precision was slightly better. The differences were small. Thus, assuming the same ratio for all intervals is easier to apply, and estimation of the appropriate variances by Bayesian or restricted maximum likelihood is not necessary.

A variance ratio of 1 means that the variance for each interval is equal to the residual variance, but in fact, the ratio should probably be very large because the amount of variance accounted for by the QTL is only 0.10 or 0.05 of the additive genetic variance. For a heritability of 0.30 and QTL effects of 10%, the ratio of residual to QTL variance would be 23.33. Thus, using a ratio of 1 is only a small step up from assuming the interval haplotype effects are fixed. The purpose of using a ratio of 1 is to remove dependencies among the possible confounding of interval haplotype effects.

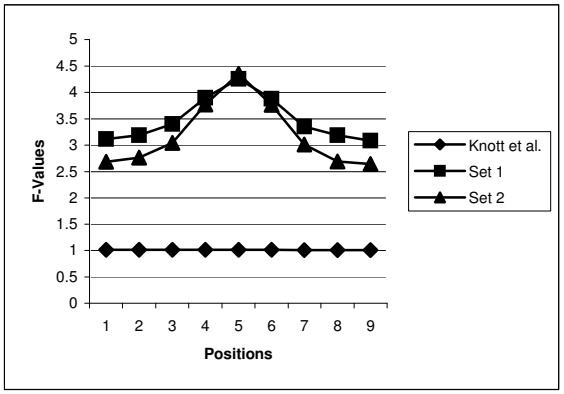

Fig. 1: Comparison of the average of 1000 replicated F-values using the Knott et al. (1996) method with linear haplotype regression model using two sets of transmission scores for two-marker intervals

SNP markers could be located more closely than every 1cM and presumably the precision of estimating QTL location should be better. Another comparison was made using 20 markers along a 10 cM segment of DNA with one QTL (Table 6). Power of detecting QTL was greater than with 10 markers. More SNP markers at closer distances were able to locate the QTL better than having SNPs only every 1 cM apart. Denser sets of markers may be optimum.

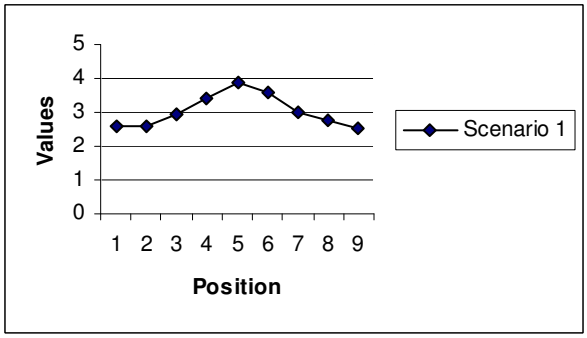

Fig. 2: The sum of absolute values of estimated haplotype effects within each haplotype pair for Scenario 1

Fine mapping of QTL using linear regression on haplotypes or a simultaneous haplotype model is possible using closely linked markers in a small region of a chromosome. Further studies should examine the effects of QTLs having more than two alleles, and having more than one QTLs in a small region at one time with these new methods. The simultaneous haplotype model could be extended to cover the entire length of a chromosome. The model should be able to detect multiple QTLs of varying magnitudes of effects. Xu (2003) describes a similar model for within family analyses where regressions are on marker genotypes (as random variables) for all markers on a chromosome. Using Bayesian methods, Xu (2003) was able to pin-point the locations of QTL without a lot of noise.

REFERENCES

1. Boettcher P.J., Pagnacco G., Stella A. 2004)A Monte Carlo Approach for Estimation of Haplotype Probabilities in Half-Sib **Families.** J. Dairy Sci., 87, 4303-4310.
2. Grapes L., Dekkers J.C.M., Rothschild M.F., Fernando R.L. 2004 Comparing Linkage Disequilibrium-Based Methods for Fine Mapping Quantitative Trait Loci. Genetics, 166, 1561-1570.
3. Kolbehdari D., Jansen G. B., Schaeffer L. R., Allen O. B. 2005 Power of QTL detection by either fixed or random models in half-sib designs. Genet. Sel. Evol., 37, 601-614.
4. Knott S.A., Elsen J.M., Haley C.S. 1996 Methods for multiple-marker mapping of quantitative trait loci in half- sib populations. Theor. Appl. Genet., 93, 71-80.
5. Meuwissen T. H. E., Goddard M.E. 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked markers. Genetics, 155, 421–430.
6. Patil N., Berno A.J., Hinds D.A., Barrett W.A., Doshi J.M., Hacker C.R., Kautzer C.R., Lee D.H., Marjoribanks C., McDonough D.P., et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science, 294, 1719-1723.
7. VanRaden P.M., Wiggans G.R. 1991 Derivation, calculation, and use of national animal model information. J. Dairy Sci., 74, 2737-2746.
8. Windig J.J., Meuwissen T.H.M. 2004 Rapid haplotype reconstruction in pedigrees with dense marker maps. J. Anim. Breed. Genet., 121, 26–39.
9. Xu, S. 2003 Estimating polygenic effects using markers of the entire genome. Genetics, 163, 789-801.
10. Zhang K., Qin Z.S., Liu J.S., Chen T., Waterman M.S. Sun F. 2004 Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Research, 14, 908-916.