

Parallel Computing in Interval Mapping of Quantitative Trait Loci

Ö. Carlborg, L. Andersson-Eklund, and L. Andersson

Linear regression analysis is considered the least computationally demanding method for mapping quantitative trait loci (QTL). However, simultaneous search for multiple QTL, the use of permutations to obtain empirical significance thresholds, and larger experimental studies significantly increase the computational demand. This report describes an easily implemented parallel algorithm, which significantly reduces the computing time in both QTL mapping and permutation testing. In the example provided, the analysis time was decreased to less than 15% of a single processor system by the use of 18 processors. We indicate how the efficiency of the analysis could be improved by distributing the computations more evenly to the processors and how other ways of distributing the data facilitate the use of more processors. The use of parallel computing in QTL mapping makes it possible to routinely use permutations to obtain empirical significance thresholds for multiple traits and multiple QTL models. It could also be of use to improve the computational efficiency of the more computationally demanding QTL analysis methods.

Molecular dissection of multifactorial traits is currently a major issue in genome research. Numerous quantitative trait loci (QTL) mapping studies are being performed and, in livestock, this has allowed the detection of several QTL with large and moderate effects (reviewed by Andersson 2001). Statistical QTL analysis can be accomplished with linear regression (e.g., Haley et al. 1994), maximum likelihood (e.g., Zeng 1994), and Bayesian methods (e.g., Sillanpää and Arjas 1999). The computational demand is relatively low for linear regression, but significantly higher for the other two (Hoeschele et al. 1997). The advantages of methods with

low computational demand are that they allow the search for multiple QTL and data permutations to obtain empirical genome-wide significance thresholds for test statistics (Churchill and Doerge 1994).

Haley et al. (1994) developed a method for regression interval mapping using multiple markers in crosses between outbred populations. We have successfully used this method for QTL detection using a wild boar \times large white intercross (Andersson et al. 1994; Andersson-Eklund et al. 1998; Knott et al. 1998), and the method described in this article has been used to obtain empirical genome-wide significance thresholds (Jeon et al. 1999). The material consists of 191 F_2 animals and the linkage map consists of 18 autosomal linkage groups ranging from 46 to 171 cM, spanning 2259 cM (for further details see Andersson et al. 1994; Andersson-Eklund et al. 1998). In the QTL mapping analysis, phenotypic values of the F_2 offspring are regressed onto indicator regressor variables for the additive and dominance effects of a putative QTL at fixed 1 cM intervals across the genome. An F -ratio test is calculated to compare the model with a QTL at this location with a reduced model without a QTL. The most likely position of a QTL is taken to be the location giving the highest F -ratio.

The use of permutations to obtain empirical significance thresholds for detection of QTL increases the number of computations by a factor of the chosen number of permutations. In most cases, the number of permutations has been set to 1000. Today it is hardly possible to make permutations for multiple QTL models using ordinary workstations. This article shows that this problem can be solved by parallel computing in a way that is simple to implement in existing programs for QTL analysis.

Methods and Algorithms

For QTL mapping, we have used the software developed for least-squares interval mapping in outbred line crosses by Haley et al. (1994). To take account of the variation caused by QTL on other chromosomes than the one currently being analyzed, additive and dominance QTL coefficients were fitted for marker locations selected, as described by Knott et al. (1998). The inclusion of these cofactors increases the computational demand, since additional columns are added to the matrices in the least-squares problem. Parallel processing was implemented in the

available program by assigning one or several linkage groups to each processor. When the number of processors used was smaller than the number of chromosomes, an equal number of chromosomes was assigned to each processor. No attempt was made to equalize the number of least-squares analyses per processor. Figure 1 presents the parallel algorithm for mapping and permutation testing for one QTL in each linkage group. When assigning whole linkage groups to processors, the maximum number of processors to be used equals the number of chromosomes in the species. We used a maximum of 18 processors (which equals the number of autosomal linkage groups in our genetic map). The analysis was subdivided by message passing using the message passing interface (MPI), requiring a minimal addition of new program code. To minimize the amount of message passing, the program was changed to write the output from each processor to a separate file, which were merged upon completion of the analysis. The parallel version of the program is available from the corresponding author.

Performances for the single and multiple processor variants of the program were measured for analyses of real data on a Cray T3E computer at the National Supercomputer Center at Linköping University, Linköping, Sweden. The Cray T3E is a distributed memory system with 232 DEC Alpha EV5 processors with a clock frequency of 300 MHz (www.cray.com/products/systems/crayt3e/, visited February 8, 2001). Two single-QTL models, with and without cofactors, were compared on 1, 3, 6, 9, and 18 processors. Each analysis consisted of five permutations and the results were extrapolated to 1000 permutations to indicate the computational time required for permutations for each trait prior to publication.

Results and Discussion

Table 1 shows the computer time needed and the relative increase in performance by adding more processors to permutation testing using the two alternate genetic models. The relative increase in performance (measured as the time used for analysis on one processor divided by the time used for multiple processors) continued without reaching a plateau as the number of processors was increased from 1 to 18. The largest increase in performance was seven times the performance of the single processor system, which is in

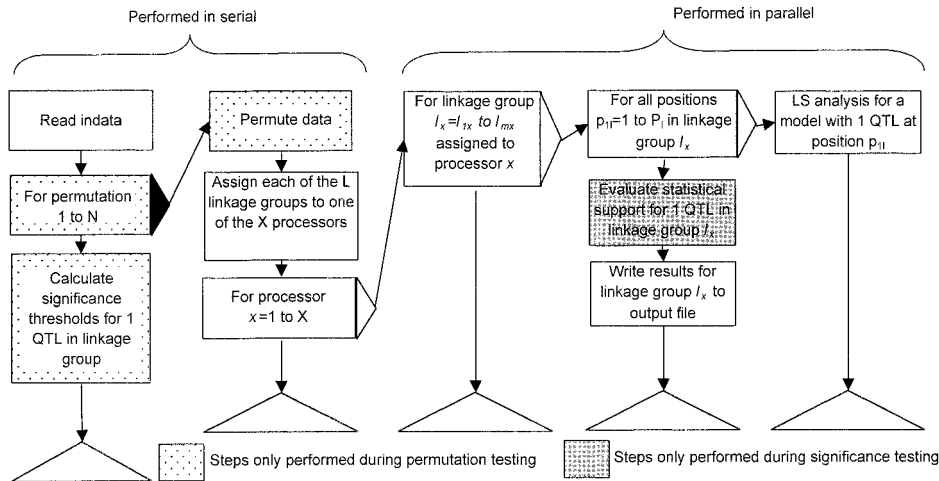


Figure 1. A parallel algorithm with alternate routes for interval mapping and permutation testing for QTL. N = number of permutations to be performed; L = number of linkage groups in the species to be analyzed; X = number of processors used for the analysis; l_{mx} = the m th linkage group assigned to processor x ; P_l = number of positions to evaluate in linkage group l ; p_{11} = position in linkage group l for the QTL, respectively; LS = least squares.

the same range as that reported for general linkage analysis problems by Dwarakadas et al. (1994). The decrease in analysis time for a permutation test is about 10 h. Thus parallel computing makes permutation tests more feasible to obtain genome-wide empirical significance thresholds for all traits in a study. The addition of four cofactors to the analysis increased analysis time by approximately 20%. The relative speed improvement by parallelization is the same for these more complex models, but the actual time gained by parallelization is greater.

The major reason for using chromosome-dedicated processors is the ease of implementation in available analysis programs. Only a few MPI statements need to be added to the existing analysis program: (1) statements to split the chromosomes to different processors before the loop over the number of chromosomes to be analyzed and (2) a chromosome-specific print statement inside the loop. This makes this algorithm a natural starting point for implementing parallel computing for QTL mapping.

In empirical data, linkage groups are of unequal size. This leads to unequal workloads of processors when assigning whole

linkage groups to the processors. In our example the processors assigned to the largest and smallest linkage groups perform 171 and 46 least squares analyses, respectively. Several alternative ways exist to obtain more equal workloads, and they require various additional changes to the analysis program. Several smaller linkage groups could be assigned to the same processor as long as their total size is smaller than that of the largest linkage group. Thus a smaller number of processors is needed to obtain the same decrease in analysis time. In our example the decrease in analysis time in Table 1 can be obtained on 5 instead of 6, 7 instead of 9, and 12 instead of 18 processors by doing this. Division of the computations into equally sized map segments, each involving 125 analyses, leads to a 50% reduction in analysis time using the same number of processors. It is also possible to increase the number of processors in the analysis and thereby further decrease the analysis time. This method requires more programming, and mapping of more than one QTL complicates the programming. The major advantage with the three methods above is that the reduction in analysis time is obtained for both QTL mapping and permu-

Table 1. The expected computer time, in hours, needed for 1000 permutations for two different genetic models, and the relative increase in performance from using up to 18 processors

QTL analysis model	Number of processors					Relative increase in performance using 18 processors
	1	3	6	9	18	
1 QTL, no cofactors	11.6	4.6	3.0	2.4	1.7	6.46
1 QTL, 4 cofactors	13.7	5.8	3.6	2.8	2.1	7.04

tation analyses. To decrease the analysis time for permutation analysis, parallelization could be implemented by separating the permutations on alternate processors. We have achieved a 23 times increased performance using 25 processors on the Cray T3E by doing this (Carlborg Ö, unpublished data).

Parallel computing can dramatically increase the efficiency in least-squares QTL mapping. The analysis time using a multiprocessor system can be decreased to less than 15% of that of a single-processor system by minor changes to existing analysis code. The parallel algorithm presented here should also be applicable to the more computationally demanding methods based on maximum likelihood and Bayesian theory. The use of parallel processing should be of even greater importance for these approaches, as it could make it computationally possible to use these methods for the exploration of more complex genetic models.

From the Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala Biomedical Center, Box 597, S-751 24 Uppsala, Sweden. Sincere thanks are due to Drs. Chris Haley and Sarah Knott for providing the source code for their analysis program. The work was supported by the Food 21 project funded by the Foundation for Strategic Environmental Research (MISTRA) and the National Graduate School in Scientific Computing (NGSSC). The analysis was carried out using computer resources at the National Supercomputer Center (NSC), Linköping University, Linköping, Sweden. Address correspondence to Örjan Carlborg at the address above or e-mail: orjan.carlborg@hgen.slu.se.

© 2001 The American Genetic Association

References

- Andersson L, 2001. Genetic dissection of phenotypic diversity in farm animals. *Nat Rev Genet* 2:130-138.
- Andersson L, Haley CS, Ellegren H, Knott SA, Johansson M, Andersson K, Andersson-Eklund L, Edfors-Lilja I, Fredholm M, Hansson I, Håkansson J, and Lundström K, 1994. Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* 263:1771-1774.
- Andersson-Eklund L, Marklund L, Lundström K, Haley CS, Andersson K, Hansson I, Moller M, and Andersson L, 1998. Mapping quantitative trait loci for carcass and meat quality traits in a wild boar \times large white intercross. *J Anim Sci* 76:694-700.
- Churchill GA and Doerge RW, 1994. Empirical threshold values for quantitative trait mapping. *Genetics* 138:963-971.
- Dwarakadas S, Schaffer AA, Cottingham RW Jr, Cox AL, Keleher P and Zwaenepoel W, 1994. Parallelization of general-linkage analysis problems. *Hum Hered* 44:127-141.
- Haley CS, Knott SA, and Elsen JM, 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* 136:1195-1207.
- Hoeschele I, Uimari P, Grignola FE, Zhang Q, and Gage KM, 1997. Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics* 147:1445-1457.
- Jeon J-T, Carlborg Ö, Törnsten A, Giuffra E, Amarger V,

- Chardon P, Andersson-Eklund L, Andersson K, Hansson I, Lundström K, and Andersson L, 1999. A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nat Genet* 21:157–158.
- Knott SA, Marklund L, Haley CS, Andersson K, Davies W, Ellegren H, Fredholm M, Hansson I, Hoyheim B, Lundström K, Moller M, and Andersson L, 1998. Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* 149:1069–1080.
- Sillanpaa MJ and Arjas E, 1999. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* 151:1605–1619.
- Zeng ZB, 1994. Precision mapping of quantitative trait loci. *Genetics* 136:1457–1468.

Corresponding Editor: Bruce S. Weir