BIOINFORMATICS

Vol. 20 no. 12 2004, pages 1887–1895
doi:10.1093/bioinformatics/bth175

# Simultaneous search for multiple QTL using the global optimization algorithm DIRECT

K. Ljungberg[1],*, S. Holmgren[1] and Ö. Carlborg[2]

[1]Information Technology, Division of Scientific Computing, Uppsala University,
P.O. Box 337,751 05 Uppsala, Sweden and [2]Genetics and Biometry, Roslin Institute,
Roslin, Midlothian EH25 9PS, UK

Received on July 15, 2003; revised on December 18, 2003; accepted on February 26, 2004
Advance Access publication March 25, 2004

## ABSTRACT

**Motivation:** A simultaneous search is necessary for maximizing the power to detect epistatic quantitative trait loci (QTL). The computational complexity demands that the traditional exhaustive search be replaced by a more efficient global optimization algorithm.

**Results:** We have the previously known algorithm adapted DIRECT, to the problem of simultaneous mapping of multiple QTL. We have compared DIRECT with standard exhaustive search and a genetic algorithm previously used for QTL mapping in two dimensions. In all two- and three-QTL test cases, DIRECT accurately finds the global optimum two to four orders of magnitude faster than when using an exhaustive search, and one order of magnitude faster than when using the genetic algorithm. Thus, randomization testing for determining empirical significance thresholds for at least three QTL is made feasible by the use of DIRECT.

**Availability:** The code of the prototype implementation is available at http://user.it.uu.se/~kl/qtl_software.html

**Contact:** Kajsa.Ljungberg@it.uu.se

## INTRODUCTION

Rapid progress in molecular genetics has led to the development of dense genetic maps, which are powerful tools for studying the molecular basis for quantitative genetic variation. One way to dissect the genetic architecture behind quantitative traits, i.e. traits showing a continuous phenotypic distribution and which are often affected by the joint effect of multiple genes and the environment, is to identify quantitative trait loci (QTL), in the genome. A QTL is a chromosomal region, locus, harboring one or several genes that affect the trait under study. The first methods used to locate, or map, QTL focused on detection of QTL by their marginal, i.e. additive and dominance, effects. These methods are presented in Lander and Botstein (1989) and Haley and Knott (1992). They are based on the concept of interval mapping, where the analyzed trait is modeled to depend on the genetic effects of a single QTL in the genome. A one-dimensional (1D) scan is performed using a dense grid covering the genome, and the single QTL model is fitted at each grid-point. The most likely position of the QTL is taken to be the grid-point with the best model fit. In composite interval mapping (Zeng, 1993) and multiple QTL mapping (Jansen, 1992), a window of analysis is introduced in the 1D scan. These schemes still search for the position of a single QTL, but markers outside the window of analysis are included as cofactors in the model. In this way, the problem with variation caused by other QTL is reduced. A randomization test (Churchill and Doerge, 1994) is normally used to derive an empirical significance threshold for a statistical test of the putative QTL. During randomization testing, normally 1000–10 000 genome scans are performed on permuted datasets to obtain a stable distribution of the model fit under the null hypothesis of no QTL. A recent overview of current QTL mapping techniques is given in Doerge (2002). Bayesian QTL mapping, described in e.g. Satagopan *et al*. (1996) and reviewed in Sillanpää and Corander (2002), is conceptually different from the parametric methods used in this paper, and will not be considered further here.

Since most quantitative traits are believed to be affected by multiple genes, it is desirable to simultaneously model the effects of these genes. Furthermore, simultaneous mapping is necessary for finding groups of interacting QTL where all loci involved lack significant marginal effects. Several methods have recently been proposed to simultaneously model the effects of multiple QTL and their interactions (e.g. Kao *et al*., 1999; Wang *et al*., 1999; Jannink and Jansen, 2001; Sen and Churchill, 2001; Carlborg and Andersson, 2002; Boer *et al*., 2002; Kao and Zeng, 2002; Yi and Xu, 2002). A fundamental problem when using a multiple QTL model is that of computational complexity. For a model including $n$ interacting QTL, the 1D scan in a single QTL model is replaced by an $n$-dimensional search for the most likely positions of the interacting loci. When using randomization testing to derive significance thresholds for multiple QTL, the computations become very demanding even for models involving only two QTL.

*To whom correspondence should be addressed.

To reduce the number of combinations of locations to evaluate, several approaches have been suggested. One suggestion (Kao and Zeng, 1997; Kao *et al.*, 1999; Zeng *et al.*, 1999, 2000) is that the computational complexity of the search is decreased by pre-selection of genomic regions with marginal effects. This potentially leads to a reduction in power since regions with primarily epistatic effects are disregarded. Sen and Churchill (2001) propose that a 2D exhaustive search is performed on a sparse grid. This procedure reduces the resolution and would still be computationally burdensome in higher dimensions. To retain the true global search without introducing a prohibitive computational demand, the exhaustive search technique must be replaced by a more sophisticated algorithm for multi-dimensional global optimization. Carlborg *et al.* (2000) suggest that a genetic optimization algorithm is used, and this type of algorithm was shown to be an efficient tool for mapping interacting QTL pairs in simulated data. Subsequently, a procedure for mapping and significance testing for epistatic QTL pairs was derived (Carlborg and Andersson, 2002). This method has recently been used to map QTL in experimental data, where multiple QTL pairs were detected in which neither of the QTL had significant marginal effects (Carlborg *et al.*, 2003). Similar results have been obtained using the method of Sen and Churchill (2001), e.g. in Sugiyama *et al.* (2001) and Shimomura *et al.* (2001).

To further investigate the evidence for higher order epistasis in experimental crosses, efficient numerical methods are needed for simultaneous mapping of QTL in two- and higher dimensions. In this study, we will explore the properties of a global optimization algorithm named DIRECT to perform QTL searches in two and three dimensions faster and more reliably than when using the genetic algorithm proposed in Carlborg *et al.* (2000). We will show that it is possible to perform simultaneous mapping, including randomization testing, of three fully interacting QTL, using a standard single-processor computer.

## SYSTEMS AND METHODS

### Computations in QTL mapping

There are two main elements in the computations when searching for QTL: the kernel problem and the global optimization problem. In general, any algorithm for the global optimization problem can be used together with any type of kernel algorithm.

The kernel problem consists of evaluating the objective function, i.e. calculating the model fit for a specific combination of putative QTL. Many different genetic models with or without interaction parameters can be used. The model parameters can be determined using e.g. ordinary linear regression (Haley and Knott, 1992; Haley *et al.*, 1994) or maximum likelihood estimation (Zeng, 1994). Both linear regression and maximum likelihood estimation, via the ECM algorithm (Meng and Rubin, 1993), involve solving a least squares problem, which is normally done using standard software library routines. The kernel problem was investigated in Ljungberg *et al.* (2002), where we presented efficient objective function evaluation algorithms based on updated QR factorizations for both linear regression and maximum likelihood kernels.

The global problem consists of optimizing the objective function, i.e. out of all possible QTL combinations finding the one giving the best model fit. It appears in two flavors. When searching the original data, the goal is to find both the most likely positions of the QTL in the set *and* the corresponding value of the parameters and model fit. However, during randomization testing, only the optimal value of the model fit is needed. As long as the value found by the algorithm is sufficiently accurate, the significance thresholds will also be accurate. This is an important observation, since the problem of determining the position of the true global optimum is more difficult for the permuted data where the connection between genotype and phenotype is broken. In this case, the optimization landscape will often have many smaller local optima, scattered over the search space, with almost the same value of the objective function.

### The global optimization problem

When performing simultaneous mapping of a set of $n$ QTL, we search a point $\bar{x}^{\mathrm{opt}} = (x_1^{\mathrm{opt}}, x_2^{\mathrm{opt}}, \ldots, x_n^{\mathrm{opt}})$ in the $n$-dimensional hypercube defined by $0 \leq x_i \leq L$. Here, $L$ is the size of the genome in $cM$ and $x_i$ is the position of the $i$th QTL in the set. The optimal value of the test statistics is independent of the ordering of the QTL in the set. Therefore, the optimization problem exhibits an $n!$-fold symmetry, equivalent to the $n!$ possible orderings of the QTL. This represents a significant reduction of the search space. In QTL mapping, the search space can be divided into boxes where each edge corresponds to one chromosome. Such a chromosome combination box $(c_1, c_2, \ldots, c_n)$ encloses all points where QTL 1 is assumed to be on chromosome $c_1$, QTL 2 is assumed to be on chromosome $c_2$ and so on. The search space symmetry is employed by restricting the search to chromosome combination boxes where $c_1 \leq c_2 \leq \cdots \leq c_n$. Boxes where two or more QTL are located on the same chromosome are also affected by the symmetry, and only part of them need to be considered.

The most likely QTL position combination $\bar{x}^{\mathrm{opt}}$ minimizes an objective function which may be written as (Ljungberg *et al.*, 2002)

$$f(\bar{x}) = \min_b (y - Ab)^{\mathrm{T}} G (y - Ab), \qquad (1)$$

where $y$ is the vector of trait values, $b$ is a vector of regression parameters and $A$ is the matrix of regression indicator variables. The matrices $G$ and $A$ depend on the QTL mapping method being used. When using the linear regression method, $G = I$, and the entries of $A$ are either constants or

**Table 1.** Names and descriptions of parametric models

| Model | Description |
|---|---|
| 2:m | A two-QTL model including fixed effects and additive and dominance *m*arginal effects. |
| 3:m | The three-QTL version of 2:m. |
| 2:m+p | The 2:m model with *p*airwise interaction effects added. |
| 3:m+p | The three-QTL version of 2:m+p. |
| 3:m+p+t | 3:m+p adding the full *t*hree-way interaction. |

continuous functions of $\bar{x}$ within chromosomes. Hence, the objective function $f(\bar{x})$ depends continuously on $\bar{x}$ within every chromosome combination box. However, at the boundaries between chromosomes, $f(\bar{x})$ is normally not continuous. We claim that the same continuity properties also hold for the maximum likelihood objective functions, but present no formal proof. Continuity in this case has been observed experimentally and is supported by a heuristic argument based on the nature of continuous functions.

## Models

Throughout this work, we have used the Haley–Knott regression method for experimental crosses between outbred lines (Haley *et al*., 1994). Table 1 describes the five genetic models that are used. The indicator variables for the marginal additive and dominance effects are determined as described in Haley *et al*. (1994) for an outbred line cross. The pairwise interaction variables are obtained by multiplying the respective additive/dominance indicator variables for the QTL in the pair as in Haley and Knott (1992), and the three-way interaction indicator variables are obtained analogously.

In this paper, we have not evaluated the power to detect epistatic QTL using the different models, nor have we investigated the best choice of model for the datasets used. This question will be addressed separately. The purpose of the current study is to compare the computational methods in terms of speed and their ability to find the global optimum of the objective function using real data, and the models were chosen with the intention to give a varied set of optimization landscapes.

## Data

We have tested the computational methods on data from two mapping populations. The first population consists of 191 pigs from an $F_2$ intercross between European Wild Boar and Large White domestic pigs (Andersson *et al*., 1994). The genome size is ∼2300 cM and we used phenotypic data for six growth-related traits. The second population consists of 852 chickens from an $F_2$ intercross between red jungle-fowl and White Leghorn chickens described in Schütz *et al*. (2002). The genome size is ∼2500 cM, and phenotypic data for nine different growth traits were used. We leave out further details about the phenotypes since we are not currently looking for new QTL.

In addition to optimizing the objective function for various models in the original datasets, it is relevant to compare empirical significance thresholds derived when using the three methods. For this purpose four sets of randomized data were generated, 1000 randomizations each of two chicken and pig traits.

# ALGORITHMS

## Exhaustive grid search

The standard method for solving the global optimization problem is to use an exhaustive grid search, evaluating the objective function for every possible QTL combination using steps of e.g. 1 cM. We have performed exhaustive 2D and 3D searches for all test cases. The symmetry of the search space was easily exploited. To make the computations feasible, the exhaustive searches were performed on a parallel computer. We measure the accuracy of DIRECT and GA as their ability to find the same optimum as the one found by exhaustive search, which is the global optimum.

## The DIRECT algorithm

The original DIRECT algorithm was presented in Jones *et al*. (1993). It searches for the global minimum $\bar{x}^{\mathrm{opt}}$ of multidimensional Lipschitz continuous functions $f(\bar{x})$ with the same type of constant constraints as the QTL mapping problem described above. The practical interpretation of a function $f(\bar{x})$ being Lipschitz continuous is that the slope of $f(\bar{x})$ is limited by some constant $K$ everywhere.

DIRECT systematically divides the search space into smaller and smaller boxes, see Figure 1. The Lipschitz continuity condition is used for deterministically determining which boxes to select for further division in each iteration. Suppose that the search space at iteration $i$ has been divided into $L$ boxes, and that $f(\bar{x})$ has been computed at the center of each box. Given $K$, a lower bound on $f(\bar{x})$ in each box could be computed, and the box with the lowest bound would be selected for further division. In practice, $K$ is unknown, so DIRECT divides all boxes where $f(\bar{x})$ has the lowest bound for *any* value of $K$ from zero to infinity. The center point of each new box is sampled, and the selection procedure is repeated. The box selection step is very fast. It should be noted that the Lipschitz continuity condition is only used for bounding $f(\bar{x})$ within each box, which is important for the application of the algorithm to QTL mapping problems.

In the original formulation of the algorithm, no box is ever discarded from the search. A box not considered potentially optimal in one iteration can be chosen for division in a later iteration. If the algorithm is run for a sufficiently long time, it is possible to prove that the global optimum will always be found (Jones *et al*., 1993). In practice, the global optimum is normally found after a rather small number of iterations. However, a general problem for global optimization

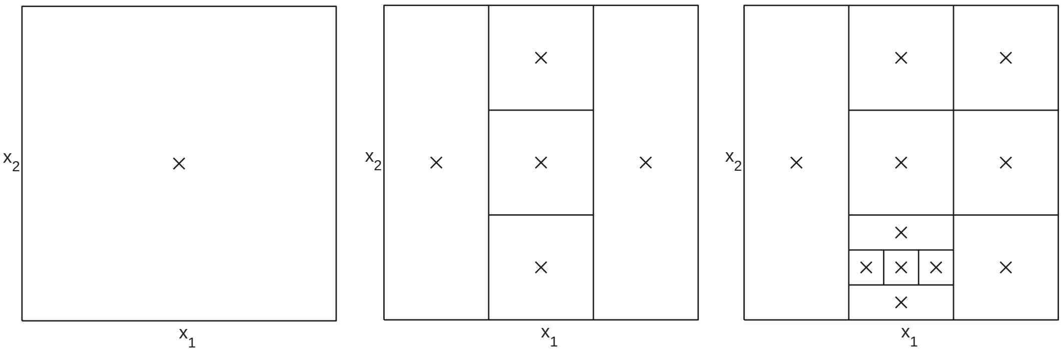

**Fig. 1.** Illustration of DIRECT search space division.

algorithms is how to determine when to stop the iterations. In the original paper (Jones *et al.*, 1993), it was suggested that a fixed number of function evaluations be used.

The original algorithm has been modified to fit the QTL search problem. As observed above, $f(\bar{x})$ is a continuous function of $\bar{x}$ within every chromosome combination box. However, at the boundaries between chromosomes, $f(\bar{x})$ is normally not continuous. To guarantee that the continuity condition of the algorithm is fulfilled, the search is initiated by sampling the center point of all chromosome combination boxes in the search space. In the original algorithm, only the center point of the complete search space is to be sampled at initiation. Also, we do not normalize the $x_i$ coordinates as in the original algorithm, and do not divide boxes with edges smaller than 1 cM.

We present no proof that $f(\bar{x})$ is Lipschitz continuous as well as continuous within the chromosome combination boxes. However, a simple argument along this line can be appled to the computations. By construction, $G$ is positive semidefinite (Ljungberg *et al.*, 2002). Thus, $f(\bar{x})$ can neither be smaller than 0 nor exceed $y^{\mathrm{T}}Gy$, which is finite. When performing the search for the set of QTL, a resolution limit of typically 1 cM is used, and thus there exists a practical Lipschitz constant which is bounded by $y^{\mathrm{T}}Gy$.

The only parameter in DIRECT with a significant influence on performance is the number of function evaluations allowed. In 2D searches, we performed 6000 evaluations, and in 3D searches, 46 000 function evaluations plus 10 000 in the intermediate refinement step. Using these settings we found the global optimum in all test cases using non-randomized data.

We have observed, in accordance with other authors (e.g. Cox *et al.*, 2001; Bartholomew-Biggs *et al.*, 2002), that DIRECT quickly locates the region of the global optimum but that local convergence is rather slow. We therefore finish the search by performing a local exhaustive search, ±5 cM in each dimension, around the best point. This is similar to the procedure suggested in Cox *et al.* (2001). In the 3D searches, we also use an intermediate refinement step. After a set number of iterations, the chromosome combination box containing the best point is located, and a number of additional iterations are performed in this box only, before the final local exhaustive search.

## The genetic algorithm

We have compared DIRECT with a genetic algorithm, GA, from a library named PGAPack (Levine, 1996). The same GA was used in Carlborg *et al.* (2000), where a position in the search space is encoded as a string of $2n$ real numbers representing the chromosomes and the chromosome positions of the $n$ QTL. One QTL position string is called a GA-chromosome, and the fit of a chromosome is given by the objective function value at the corresponding position in the search space. A GA-population is a set of GA-chromosomes, and in each iteration new GA-chromosomes are created by mutation and crossover among the existing ones, selecting for best fit. The GA is, thus, partly related to forward selection in the sense that mutation and/or crossover on a good candidate GA-chromosome often results in keeping one QTL position fixed and changing the other. The symmetry of the search space is exploited by not allowing the algorithm to evaluate the reflection of a position already visited. After the GA is finished, a local exhaustive search ±5 cM is performed around the found optimum in the same way as for DIRECT.

A significant effort has been made in tuning the parameters to obtain the best possible accuracy for all test cases. Table 2 shows the different parameter settings chosen for this study. We refer to the settings chosen in Carlborg *et al.* (2000) as GA(20k), the name reflecting the approximate number of function evaluations performed. The best parameter choice found was a modified version of GA(20k) which we call GA(75k). GA(6k) is the settings giving only the same number of function evaluations as DIRECT in 2D. The parameterization used in the 3D searches is called GA(1M).

# IMPLEMENTATION

All objective function evaluations were done using the efficient kernel algorithm presented in Ljungberg *et al.* (2002). The experiments showed that, in practice, the only factors

**Table 2.** Parameter settings for the GA

| Name | Number of populations | Iterations/ population | Population size |
|---|---|---|---|
| GA(6k) | 3 | 980 | 20 |
| GA(20k) | 10 | 1000 | 20 |
| GA(75k) | 25 | 1500 | 20 |
| GA(1M) | 25 | 2000 | 200 |

determining the CPU time for the three methods are the number of function evaluations performed and the time required for a single evaluation. The CPU time for one evaluation depends on the model and dataset, but not on the optimization method since they use the same kernel algorithm.

All code was written in Fortran90, and the computations were done on SPARC UIII, 900 MHz processors. The exhaustive searches were performed on a parallel computer using message passing interface (MPI) (http://www.mpi-forum.org) library routines, and the CPU times reported are the sums of the CPU times for each processor, not including overhead time for the parallelization.

## RESULTS

### Original, non-randomized, data

The accuracy is reported as the percentage of successful localizations of the exact global optimum out of the total number of searches. Since the GA has a random element, the result will depend on the random seed. Therefore, using this method, each search was repeated 15 times to give a reasonable statistic. DIRECT is deterministic and gives the same result every time.

We again stress that we do not present any new QTL mapping method in this paper. The parameter estimates at a given position depend only on the choice of objective function, e.g. linear regression or maximum likelihood, and are completely independent of the optimization algorithm. In the context of this study, returning the correct result means returning the set of QTL positions that gives the best value of the test statistic for the chosen model and mapping method.

First we report the results for searches in two dimensions. We have tested the methods for the 2:m and 2:m+p models in Table 1 on all datasets described in the Data subsection, which gives a total of 30 tests.

Figure 2 shows the average CPU times and accuracy over the nine phenotypes of the chicken dataset using the 2:m and 2:m+p models. The results for pig data and the same models were very similar. It should be noted that the CPU times essentially equal the number of function evaluations performed multiplied with the time for each objective function evaluation. The CPU time for function evaluation is largely determined by the number of individuals and fixed effects, and the important result here is the relative change in computation time, not the absolute values. An exhaustive search with the 2:m model requires about 20 min, and 46 min with the 2:m+p model. DIRECT finds the global optimum in less than 3 and 7 s, respectively, which represents a speed-up of between two and three orders of magnitude. GA(75k) gives the global optimum at close to 100% of the runs, with CPU time 34 and 76 s. Using GA(6k), the genetic algorithm with the same number of function evaluations and thus practically the same CPU time as DIRECT, reduces the accuracy from close to 100% to around 60%. GA(20k), the settings of Carlborg *et al.* (2000), gives intermediate results. The GA has more difficulties finding the global optimum when epistasis is included in the model. It was observed already in Carlborg *et al.* (2000) that the GA sometimes failed when a QTL pair lacked significant marginal effects. This can be explained by the forward selection property of the algorithm.

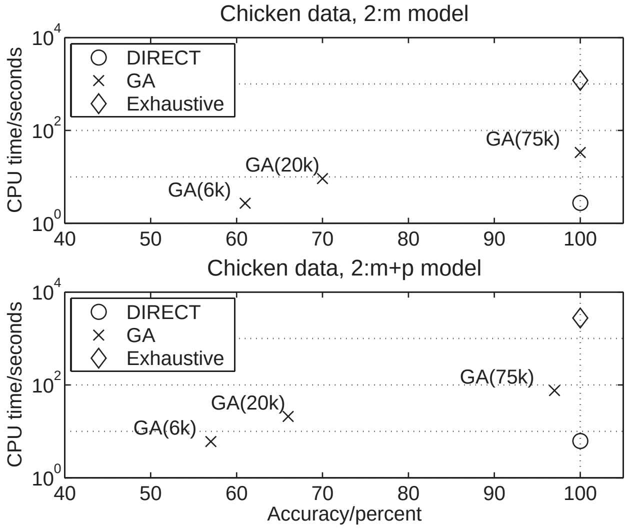

**Fig. 2.** CPU time for 2D searches as a function of the percentage of successful localizations of the global optimum.

Now we turn to three-QTL results. We have used the 3:m and 3:m+p models combined with four chicken traits, one of which was also used with the 3:m+p+t model, giving nine tests in total.

Figure 3 shows the average CPU times and accuracy over four phenotypes of the chicken dataset using the 3:m and 3:m+p models, and one phenotype using the 3:m+p+t model. The chicken 3:m, 3:m+p and 3:m+p+t exhaustive searches would take ∼25, 60 and 142 days, respectively, on a single processor computer. The gain in using DIRECT over exhaustive search is more than four orders of magnitude in speed, the searches taking 0.5, 3 and 6 min, while not losing accuracy. Using GA(1M) gives high accuracy for the 3:m and 3:m+p models but lower accuracy for 3:m+p+t and is over one order

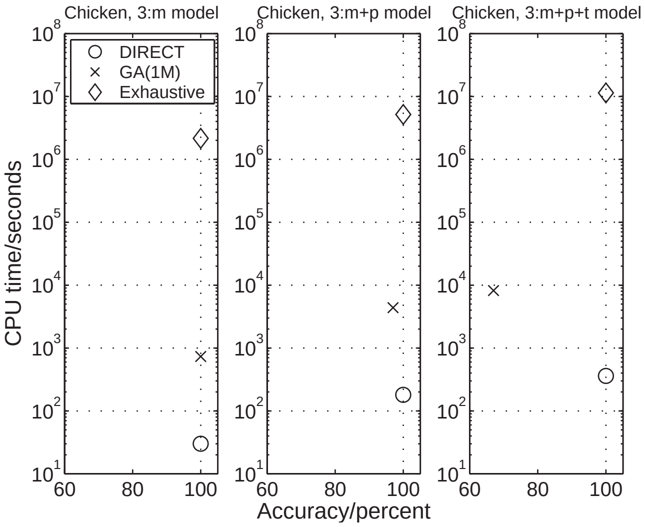

**Fig. 3.** CPU time for 3D searches in chicken data as a function of the percentage of successful localizations of the global optimum.

of magnitude slower than DIRECT, the searches requiring 12, 73 and 136 min, respectively.

Figures 4 and 5 illustrate the difference in search patterns between DIRECT and the GA. Here we show results from model 2:m+p with pig data. The two figures show the sampling pattern after a complete run, i.e. 6000 function evaluations, using DIRECT (Fig. 4) and GA(6k) (Fig. 5). The locations where the objective function has been evaluated are marked with '×' in contour plots of the objective function around the four largest peaks. For clarity, most contours for lower levels are not shown. DIRECT uses the function evaluations very efficiently. It gives even coverage of the search space with dense clusters of function evaluations around the largest peaks. This indicates that the algorithm can locate the global optimum for original data also in difficult cases when there are many local optima of similar magnitude. Using the same number of function evaluations, the GA sometimes does not find the global optimum, even if the regions around all the four largest peaks are sampled. If many peaks are of similar height, the best position found so far when the local search is initiated might be at the wrong peak. Or the right peak might have been found, but local exhaustive search ±5 cM is not a good enough method to localize the very best position on the peak. The GA samples the search space stochastically to a large extent.

## Randomized data

Finding the global optimum can be expected to be more difficult in a randomized dataset, since the optimization landscape will be smoothed out and the peaks smaller for most of the randomizations when the connections between genotype and phenotype is broken.

We used the 2:m+p model for the four randomized datasets. We determined the 1.0, 5.0, 10 and 20% genome-wide significance thresholds for 0 against 2 QTL using an exhaustive search. The thresholds were also calculated using DIRECT and GA on the same data. In Table 3, we report the true levels (as given by the exhaustive search) of the thresholds derived using DIRECT, GA(20k) and GA(75k) intended to give the 1.0, 5.0, 10 and 20% significance levels. A number 5.6% in the 5.0% column means that in 5.0% of the randomizations, a global optimum better than $x$ ($x$ not reported) was found when using the global optimization algorithm, i.e. the 5.0% significance threshold would be taken to be $x$, while in reality 5.6% of the true global optima, obtained using exhaustive search, were better than the same $x$. A threshold that is too low, i.e. at 5.6% instead of 5.0%, gives a slight increase in the type I error rate. This could in part explain the increased rate of type I errors in Carlborg and Andersson (2002), where the genetic algorithm is used.

Looking at the individual runs, it can be seen that DIRECT finds the wrong position in about 9% of the randomizations. The function values are, however, accurate enough to give nearly the same threshold values as exhaustive search, and they are calculated between two and three orders of magnitude faster. Using GA(20k) the wrong position is found in 23–35% of the cases. This is about the same error rate as was found with non-randomized data. The error rate is about 1–14% when using GA(75k), which gives very accurate threshold values at the cost of increased CPU time.

There is a tendency for the 1 and 5% computed thresholds to be more accurate than the 10 and 20%. This reflects that it is easier for both algorithms to find large peaks, while the randomizations giving a 'smeared' landscape with many smaller peaks are more difficult from an optimization point of view.

## DISCUSSION

This study has shown that DIRECT is a fast and accurate algorithm for global optimization in QTL mapping. The exact optimum is found in real datasets, and searches in randomized data are accurate enough to give almost the same empirical significance thresholds as exhaustive search. 2D searches require a few seconds, and 3D searches are finished in a few minutes. DIRECT makes randomization testing of two QTL models faster, and randomization testing of three QTL models fully feasible. This opens the possibility to thoroughly investigate the power of simultaneous search to detect triplets of interacting QTL, which will be done in future research.

We have implemented DIRECT to simultaneously search for up to 15 interacting QTL. In practice, a search in more than five or six dimensions will necessitate some strategy to reduce the search space, which otherwise will be prohibitively large even if DIRECT is used. Possible strategies include imposing conditions such as having at least one QTL on each chromosome in a set. This is essentially a forward selection procedure,

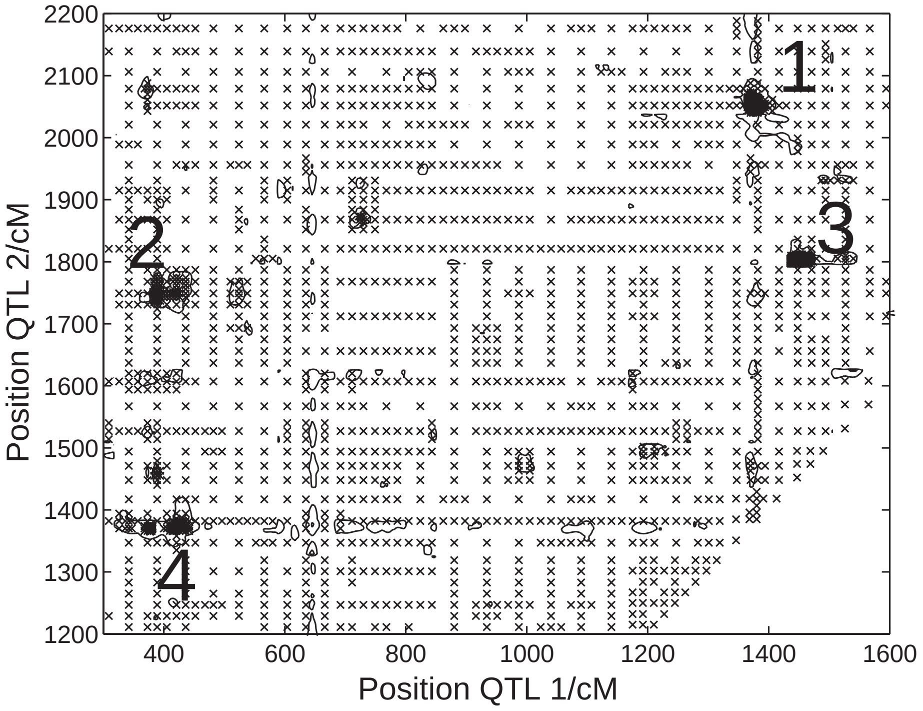

**Fig. 4.** Search pattern after 6000 function evaluations with DIRECT in the region around the four largest peaks, numbered 1–4 according to their relative ranks.

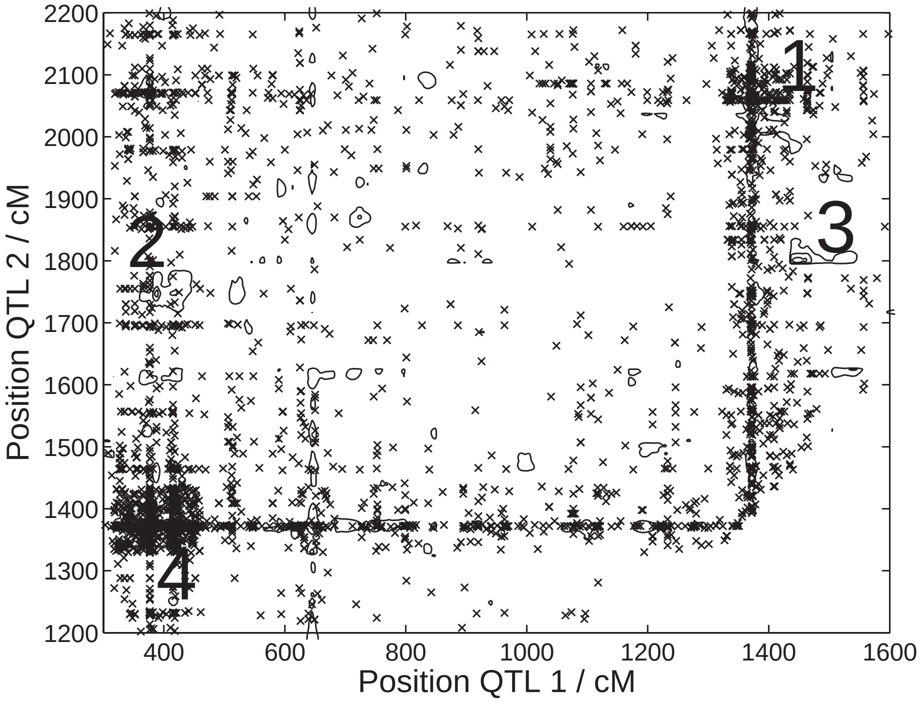

**Fig. 5.** The GA sampling pattern after 6000 function evaluations, GA(6k), in the region around the four largest peaks.

**Table 3.** Empirical thresholds with CPU times

| | 2:m+p model<br>Computed thresholds (%) | | | | Time |
|---|---|---|---|---|---|
| Chicken data algorithm | | | | | |
| Exh. search | 1.0 | 5.0 | 10 | 20 | 32 days |
| DIRECT | 1.0 | 5.2 | 10 | 21 | 1.7 h |
| GA(20k) | 1.0 | 5.6 | 12 | 24 | 5.8 h |
| GA(75k) | 1.0 | 5.3 | 10 | 21 | 21 h |
| Pig data algorithm | | | | | |
| Exh. search | 1.0 | 5.0 | 10 | 20 | 17 days |
| DIRECT | 1.0 | 5.1 | 10 | 21 | 57 min |
| GA(20k) | 1.0 | 5.4 | 11 | 22 | 3.5 h |
| GA(75k) | 1.0 | 5.0 | 10 | 20 | 14 h |

but with the advantage that the positioning of the 'known' QTL is unrestricted within the boundaries of the pre-selected chromosomes. Also, the significant reduction in search space using this approach makes it possible to simultaneously add more than one free QTL in each forward selection step. It is a statistical problem to investigate whether the resulting improvement in objective function value motivates adding QTL to the model. Backward selection can be performed by restricting the number of QTL on each chromosome to at most the same number as with the optimal $n$ QTL model, and then optimizing an $n-1$ QTL model in the resulting search space. Again this has the advantage of free QTL positioning within chromosome boundaries. These options are implemented, but so far not evaluated.

DIRECT is developed to find the optima of Lipschitz continuous functions, i.e. functions where the rate of change of the objective functions is limited everywhere by some constant $K$, where $K$ is normally unknown. We gave a motivation for Lipschitz continuity of the QTL mapping objective function based on that $0 \leq f(\bar{x}) \leq y^{\mathrm{T}}Gy$, and that the resolution is limited. A more interesting observation is that genetic distance is a measure of change, a measure of recombination events. Recombinations are reflected by change in the indicator variable matrix $A$ and consequently in $f(\bar{x})$. The magnitude of the change in $f(\bar{x})$ depends not only on the genetic distance but also on the phenotype values of the individuals switching genotype between the flanking markers, but there still exists a limit on the possible rate of change in $f(\bar{x})$. No such limit is assumed in the calculations, but we believe it is the explanation for the good performance of DIRECT.

The optimization landscape will change, and consequently the convergence properties of DIRECT will change, if a different objective function is chosen. According to Kao (2000), the differences between maximum likelihood mapping and linear regression are minor if the marker map is dense, but become larger, e.g. as the size of marker intervals and the proportion of the variance explained by a QTL increase. DIRECT can be applied in combination with maximum likelihood methods, since the condition of a practical Lipschitz constant is fulfilled, but we have not investigated the exact performance of the algorithm on those types of objective functions. Also, when analyzing data from other types of experimental crosses, the optimization landscape will probably be different than in this study. However, we believe that an $F_2$ cross between outbred lines is one of the most difficult cases, since the objective function will contain more noise and less distinct peaks. In addition, we have used models with many parameters, and that too will make the peaks smaller and more difficult for the optimization algorithms to find. When adapting DIRECT to other experimental designs, an advantage is that the only parameter necessary to adjust is the number of function evaluations allowed, as opposed to GA which is dependent on the settings of a large number of parameters. We have used 6000 function evaluations in the 2D searches and 56 000 evaluations in the 3D search, which corresponds to 0.2% and 0.002% of the total number of positions.

## ACKNOWLEDGEMENTS

Ö.C. gratefully acknowledges financial support from The Knut and Alice Wallenberg foundation. We thank Leif Andersson for contributing the datasets. K.L. is affiliated with The Linnaeus Centre for Bioinformatics and The Graduate School in Mathematics and Computing, both at Uppsala University.

## REFERENCES

Andersson,L., Haley,C., Ellegren,H., Knott,S., Johansson,M., Andersson,K., Andersson-Eklund,L., Edfors-Lilja,I., Fredholm,M. and Hansson,I. (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science*, **263**, 1771–1774.

Bartholomew-Biggs,M., Parkhurst,S. and Wilson,S. (2002) Using DIRECT to solve an aircraft routing problem. *Comp. Optim. Appl.*, **21**, 311–323.

Boer,M., ter Braak,C. and Jansen,R. (2002) A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics*, **162**, 951–960.

Carlborg,Ö. and Andersson,L. (2002) The use of randomization testing for detection of multiple epistatic QTL. *Genet. Res.*, **79**, 175–184.

Carlborg,Ö., Andersson,L. and Kinghorn,B. (2000) The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics*, **155**, 2003–2010.

Carlborg,Ö., Kerje,S., Schütz,K., Jacobsson,L., Jensen,P. and Andersson,L. (2003) A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res.*, **13**, 413–421.

Churchill,G. and Doerge,R. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Cox,S., Haftka,R., Baker,C., Grossman,B., Mason,W. and Watson,L. (2001) A comparison of global optimization methods for the design of a high-speed civil transport. *J. Global Optim.*, **21**, 415–433.

Doerge,R. (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.*, **3**, 43–52.

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

Haley,C., Knott,S. and Elsen,J.-M. (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, **136**, 1195–1207.

Jannink,J.-L. and Jansen,R. (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics*, **157**, 445–454.

Jansen,R. (1992) A general mixture model for mapping quantitative trait loci by using molecular markers. *Theor. Appl. Genet.*, **85**, 252–260.

Jones,D., Perttunen,C. and Stuckman,B. (1993) Lipschitzian optimization without the Lipschitz constant. *J. Optim. Theory App.*, **79**, 157–181.

Kao,C.-H. (2000) On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics*, **156**, 855–865.

Kao,C.-H. and Zeng,Z.-B. (1997) General formulae for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics*, **53**, 653–665.

Kao,C.-H. and Zeng,Z.-B. (2002) Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics*, **160**, 1243–1261.

Kao,C.-H., Zeng,Z.-B. and Teasdale,R. (1999) Multiple interval mapping for quantitative trait loci. *Genetics*, **152**, 1203–1216.

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

Levine,D. (1996) *Users Guide to the PGAPack Parallel Genetic Algorithm Library*. Argonne National Laboratory, Argonne, IL.

Ljungberg,K., Holmgren,S. and Carlborg,Ö. (2002) Efficient algorithms for quantitative trait loci mapping problems. *J. Comput. Biol.*, **9**, 793–804.

Meng,X.-L. and Rubin,D. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Satagopan,J., Yandell,B., Newton,M. and Osborn,T. (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, **144**, 805–816.

Schütz,K., Kerje,S., Carlborg,Ö., Jensen,P. and Andersson,L. (2002) Analysis of a red junglefowl × white leghorn intercross reveals trade-off in resource allocation between behavior and production traits. *Behav. Genet.*, **32**, 423–433.

Sen,S. and Churchill,G. (2001) A statistical framework for quantitative trait mapping. *Genetics*, **159**, 371–387.

Shimomura,K., Low-Zeddies,S., King,D., Steeves,T., Whiteley,A., Kushla,J., Zemenides,P., Lin,A., Vitaterna,M., Churchill,G. and Takahashi,J. (2001) Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res.*, **11**, 959–980.

Sillanpää,M. and Corander,J. (2002) Model choice in gene mapping: what and why. *Trends Genet.*, **18**, 301–307.

Sugiyama,F., Churchill,G., Higgins,D., Johns,C., Makaritsis,K., Gavras,H. and Paigen,B. (2001) Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics*, **71**, 70–77.

Wang,D., Zhu,J., Li,Z. and Paterson,A. (1999) Mapping QTLs with epistatic effects and QTL × environment interactions by mixed linear model approaches. *Theor. Appl. Genet.*, **99**, 1255–1264.

Yi,N. and Xu,S. (2002) Mapping quantitative trait loci with epistatic effects. *Genet. Res.*, **79**, 185–198.

Zeng,Z.-B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl Acad. Sci., USA*, **90**, 10972–10976.

Zeng,Z.-B. (1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.

Zeng,Z.-B., Kao,C.-H. and Basten,C. (1999) Estimating the genetic architecture of quantitative traits. *Genet. Res.*, **74**, 279–289.

Zeng,Z.-B., Liu,J., Stam,L., Kao,C.-H., Mercer,J.M. and Laurie,C.C. (2000) Genetic architecture of a morphological difference between two drosophila species. *Genetics*, **154**, 299–310.