Stochastically Guaranteed Global Optima in Multi-Dimensional QTL Searches

Carl Nettelblad, Sverker Holmgren
Division of Scientific Computing
Department of Information Technology
Uppsala University
Sweden

February 26, 2010

Abstract

The problem of searching for multiple quantitative trait loci (QTL) in an experimental cross population of considerable size poses a significant challenge, if general interactions are to be considered. Different global optimization approaches have been suggested, but without an analysis of the mathematical properties of the objective function, it is hard to devise reasonable criteria for when the optimum found in a search is truly global.

We reformulate the standard residual sum of squares objective function for QTL analysis by a simple transformation, and show that the transformed function will be Lipschitz continuous in an infinite-size population, with a well-defined Lipschitz constant. We discuss the different deviations possible in an experimental finite-size population, suggesting a simple bound for the minimum value found in the vicinity of any point in the model space.

Using this bound, we modify the DIRECT optimization algorithm to exclude regions where the optimum cannot be found according to the bound. This makes the algorithm more attractive than previously realized, since optimality is now in practice guaranteed. The consequences are realized in permutation testing, used to determine the significance of QTL results. DIRECT previously failed in attaining the correct thresholds. In addition, the knowledge of a candidate QTL for which significance is tested allows spectacular increases in permutation test performance, as most searches can be abandoned at an early stage.

1 Introduction

The rapid development of experimental techniques and off-the-shelf technology for molecular genetics has resulted in that dense genetic marker maps can be provided much easier and cheaper than before. In addition, most markers within such maps can be genotyped on large sets of individuals in short time with a limited use of resources. These developments open new possibilities for detailed genetic analysis of large populations, where one highly interesting and challenging field is the genetic mapping of *quantitative traits*. Such traits exhibit a continuous distribution and are affected by the environment as well as the genetic composition. The underlying genetic architecture of a quantitative trait can be described by identifying *quantitative trait loci* (QTL) for a population and attributing effect values to these loci in a suitable model framework.

The standard approach for locating a single QTL is based on *interval mapping* (IM) [13]. Evaluating the standard IM model at a given position in the genome involves solving a maximum-likelihood problem based on genotype and phenotype frequencies for the population studied. In a *QTL search*, the evaluation of the model is repeated for a set of candidate positions in the genome to determine the QTL location that results in the best model fit. Some important lines of development of the initial IM approach include more computationally efficient linear approximations of the IM model [10] and windowing mechanisms attempting to take the genetic background into account to increase the search power [11].

The result of a QTL analysis is only useful if a proper significance threshold can be derived. Already when searching for a single QTL, traditional χ^2 approximations have been shown to have a significant bias [3]. Therefore, randomization testing is frequently used [5], where permuted datasets (genotypes vs. phenotypes) are employed to empirically derive the distribution of the optimal model fit under the null hypothesis of no QTL being present.

In general, it can be assumed that multiple QTL should be included in a model to fully describe the genetic effect on a trait [7]. Even under the assumption that these QTL do not interact (i.e. the true population effects are perfectly additive), the estimated QTL effects will be more accurate if all putative loci are included in a single, multi-loci model. However, using a d-QTL model results in a d-dimensional global optimization problem that has to be solved to locate the set of QTL resulting in the best model fit. The standard approach for solving QTL search problems is to use an exhaustive search over a dense lattice covering the search space. This approach rapidly becomes

computationally intractable for multidimensional searches, and QTL mapping procedures involving true d-dimensional optimization have so far not been widely used for $d > 2$. One popular approach which avoids the multidimensional searches is the so called forward-selection procedure [2]. This scheme attempts to find the optimum location in the d-dimensional search space by recursively performing one-dimensional searches and selecting the best single-locus models. The initial single-locus model is iteratively extended by adding additional loci in a greedy manner, where the model of optimal explanation power so far is the only one to be extended in the following iteration. A main problem with this approach is that the set of QTL might present an interaction pattern that is not properly captured when only considering the loci one by one, leading to a non-optimal subset being selected. The elementary unit in the extension step of the forward selection process can be moved from single loci to pairs of loci to better account for interactions, although this comes at a significant computational cost and still will not uncover the true optimum in more complex gene networks.

A more general and potentially more accurate approach for models involving d QTL is to consider the full, d-dimensional search problem and adapt known algorithms for such problems to the QTL analysis setting. Here, both stochastic methods such as genetic algorithms [4] and deterministic optimization algorithms such as e.g. DIRECT [12] have been suggested. When using a general global optimization approach, the problem of determining the set of QTL resulting in the optimal model fit is separated from the evaluation of the model of the QTL effects. This implies that models based on e.g. both the linear regression approximation and the standard interval mapping maximum-likelihood model can be easily included in an optimization framework. In this paper, we focus on linear regression models since these are much less computationally demanding and lend themselves to the type of transformations that are exploited to derive the efficient and accurate search methods presented.

For native multidimensional search methods, it is possible to directly transfer the methodology for single-QTL model permutation tests to higher dimensionality. However, performing tens of thousands of permutation tests to get a significance threshold, where each iteration itself contains a multidimensional QTL search, is of course a very computationally demanding task, even when an efficient global optimization scheme such as DIRECT is employed. Also, earlier results [14] show that using DIRECT for the permutation tests can result in a bias in the significance thresholds compared to those from an equivalent (but much more computationally demanding) exhaus-

tive search. The reason for this seems to be that the termination condition used in [14] was too restrictive in the very flat optimization landscape present in most permuted cases. The end result is that a putative 95% threshold instead would give 94.8% significance, as the exhaustive search results in a few additional cases exceeding the optimum found by DIRECT. Naturally, DIRECT can show no bias in the other direction, any point found with this algorithm should also have been evaluated and detected in an exhaustive search. It should be noted that no bias was detected in the DIRECT searches for non-permuted phenotype data, probably due to the more structured nature of these optimization landscapes.

In this paper, we revisit the use of the DIRECT algorithm for QTL analysis. We show that by transforming the objective function corresponding to the linear regression QTL model we can derive a significantly improved scheme for multidimensional QTL searches. The basis for this new algorithm is that DIRECT relies on an implicit assumption of Lipschitz continuity of the objective function, i.e. a bounded variation within a bounded distance from any combination of loci where the model is evaluated. For QTL mapping problems we show that, under some reasonable conditions, it is possible to derive a bound on the Lipschitz constant for the transformed objective function corresponding to analysis of an ideal infinite-size population. This bound is then used to prune the search tree in DIRECT, resulting in that finite termination of the search at a global optimum can be achieved without examining all possible combinations of loci like in an exhaustive search scheme. These asymptotic results are then shown to correspond to a Lipschitz-like behavior also for finite-size data sets, and allow relevant empirical bounds to be constructed by adding a statistically motivated "safety distance" to the results for infinite-size populations.

These insights in how DIRECT can be improved for QTL search problems are finally developed further to present a modified permutation test where the presence of a set of QTL candidate locations is assumed. By using the Lipschitz bound and only performing the search in each permutation test iteration deep enough to determine which of the pre-existing candidates the current permutation would tend to prove/disprove, extensive performance gains are achieved compared to using the corresponding exhaustive search algorithm. This also allows enough iterations to be performed in sensitive cases to remove the bias due to shallow searches found in the previous application of DIRECT for QTL permutation tests in [14]. The large performance gains for the permutation tests can be considered to be the most important result in this paper, since it enables high-

confidence mapping of multiple QTL on a regular basis without the need of massive computational resources.

2 The DIRECT Algorithm for QTL Searches

The global optimization algorithm DIRECT [12] is based on a divide-and-conquer approach where the search space is successively divided into smaller and smaller boxes and the search effort is focused in the most promising regions. When applied to a QTL search, a set of QTL is defined as a point in a d-dimensional hypercube. In a traditional exhaustive search, the residual sum of squares (RSS) of the linear regression model is evaluated at every point of a fine lattice in this hypercube (possibly excluding some points by exploiting symmetry, corresponding to the ordering of the QTL positions being irrelevant). If DIRECT is run to completion using a minimum resolution criterion matching the step length in the exhaustive search lattice, it will have explored exactly the same points as the corresponding exhaustive search. However, as we will show in this paper, it is normally possible to terminate the DIRECT process at a much earlier point, while still guaranteeing that the same global optimum is found as for the exhaustive search.

The original DIRECT algorithm initially creates a single search box covering the full search space, and the objective function, i.e. in our case the RSS, is evaluated in the center of this box. This box is then split into three equally sized boxes along the majoring dimension. This trinary split results in the centroid of the original box coinciding with the centroid of one of the new boxes. Therefore, only two additional function evaluations are required for the three resulting boxes. DIRECT then continues by iteratively splitting the boxes. At the end of each DIRECT iteration, the convex hull is determined among the remaining boxes, in a space of box radii vs. objective function values. This hull determines which boxes to split in the following iteration. The selection is done based on the principle indicated above; if tracing along the box radii, the RSS value for the sequence of boxes is monotonously decreasing. Figure 1 illustrates a few iterations of DIRECT in a simple one-dimensional space. The hull is "peeled off" when those boxes are split, making new boxes available in the next iteration. This process is repeated until a suitable termination condition has been reached.

As has been remarked above, DIRECT implicitly assumes that the objective function is Lipschitz continuous. This means that a

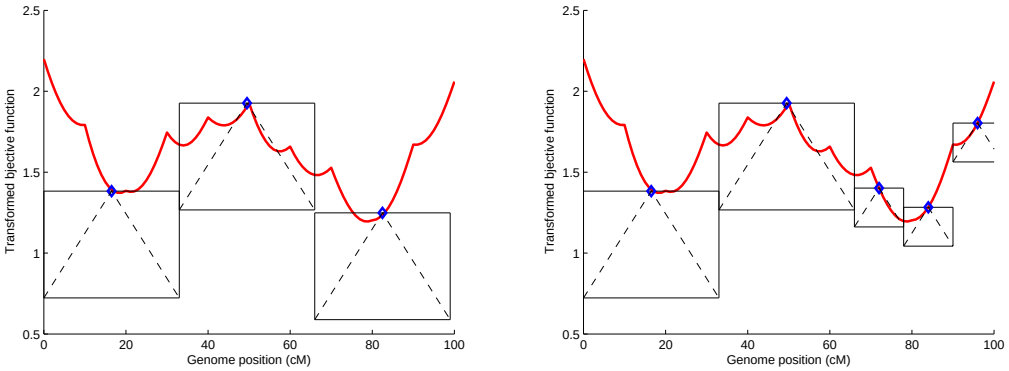

Figure 1: To the left, three boxes, in the optimization space. As all boxes are of equal size, only one will be selected for splitting in the convex hull, resulting in the boxes to the right. If the splitting would continue, two boxes would be split, one from each size, as the smallest function value in the smaller box size is slightly lower than the smallest value in the larger one. Dashed lines indicate possible minimum function values found at each distance from the box centroid, assuming a strict Lipschitz bound of $K = 0.04$.

constant K can be found such that no partial first-order derivative of the RSS will exceed K in any position. If we accept an approximation where the RSS is represented by a trapezoid between the lattice points used in the exhaustive search that we are setting out to improve, this property will be true for any QTL model. However, some care should be taken when using this argument. The Lipschitz condition for the RSS has its origin in the fact that the probability of co-inheritance of two genomic loci is related to their physical distance from each other on the chromosome. The standard unit for genetic distances, mapping distance, is even directly defined from such probabilities. It is also clear that separate chromosomes segregate independently and can be considered to have an infinite mapping distance. Even if it is possible to line up all genome locations as a single line, one chromosome after another, there is a clear disjunction at the chromosome borders. This corresponds to a discontinuity of the RSS at these boundaries. In [14], this situation is handled by introducing what is essentially several DIRECT searches governed by a common priority queue. In this version of the algorithm, each chromosome combination is considered to correspond to a separate search space, called a chromosome combination box (cc-box). In the first step of the DIRECT scheme, the RSS is evaluated at the centroids of all these cc-boxes.

If a bound on the Lipschitz constant K is known it is possible to compute upper and lower bounds for the RSS within any box in the search space. However, when using the original DIRECT algorithm, K is unknown. Note that it is still possible to impose a partial ordering of boxes; if a box A is both smaller and has a larger value of RSS than another box B, then no value of K can result in a smaller minimum bound within A than within B.

DIRECT is part of the general family of Lipschitz optimization schemes. Here, the more traditional schemes [19, 18, 8, 15] assume that a bound of the Lipschitz constant is known. DIRECT deals with the constant being unknown by always splitting the largest remaining box, as a large enough value for K would allow the minimum point to be located within the volume of that box. If a bound on K is known when using DIRECT, known values of RSS in small boxes will also enable exclusion of larger boxes, hereby introducing a pruning of the search tree. In contrast to the peeling effect of successive convex hulls in DIRECT, volumes excluded due to a Lipschitz criterion are permanently irrelevant for further evaluation of the RSS.

The natural termination condition for such a modified DIRECT algorithm is given by the finite resolution criterion corresponding to the lattice in the underlying exhaustive search. This results in an error bound of Kh on the value of the RSS, where h is the step length in the lattice. Thus, a bound on the Lipschitz constant K leads to improved performance (by more efficient exclusion), a well-defined error bound, and to a guarantee that the result is equivalent to the corresponding exhaustive search.

3 Linear Regression QTL Models

In this paper we consider QTL analysis for experimental populations with known relations between individuals, and where the founder individuals (the first generation in the pedigrees) have some origin-defining feature. This can be a matter of a known genetic relation (a set of common inbred or outbred lines) between the founders, or some founders expressing a specific phenotype.

For ease of presentation, let us consider the typical case where we have a F_0 generation of individuals that can be considered to belong to either out of two lines, 0 and 1, so we consider each individual to carry the genotype 00 or 11. If we make a cross of such individuals between the lines, we get a F_1 generation. All those individuals will be of 01 genotype (ordering aside), and so all variation between those individuals would be environmental in nature.

From the F_1 population, two other structures are commonly

considered. One is the *backcross*, where F_1 individuals are crossed to F_0 individuals from either line (say the 11s). This results in that all individuals from that cross will be of 01/11 genotype, so there are only two classes of individuals, allowing for considerable model simplicity. The other common structure results from crossing F_1 individuals with each other, an *intercross*, yielding an F_2 generation with all genotypes 00/01/10/11 represented, in equal proportions (so half of all individuals will be of either heterozygote genotype).

If full genotype information is available the genotype information at a specific locus can be expressed as an $n \times m$ indicator matrix $\mathbf{Z}$, for n individuals and m possible genotypes. This mathematical formalism is consistent with the description in [1, 20]. The genetic effects are then modeled as

$$\mathbf{G}^* = \mathbf{ZSE} + \varepsilon. \tag{1}$$

Here, the design matrix is the product of $\mathbf{Z}$ and $\mathbf{S}$. The separation into two matrices allows us to describe $\mathbf{S}$ independently of the specific number of individuals in the population and their respective genotypes. The matrix $\mathbf{S}$ is ideally chosen to result in an orthogonal model, i.e. a model where independent estimates of variance can be assigned to the included effects (parameters), and where effects can be removed from the model with no change to the estimated values and variances for the remaining effects. To achieve this, the design matrix will differ slightly between different loci, as the random sampling of alleles in the analyzed population will not be perfectly uniform [16, 1].

The quality of a model can be determined by the portion of the total variance in the phenotype values that is explained by the model. A larger explained variance is equivalent to a smaller RSS for the regression, which is given by $\sum \varepsilon^2$ in (1) above.

3.1 Explainable Variance as a Function of Genetic Distance

We now consider several QTL search objective functions and examine their behavior at, and in the vicinity of, a QTL. The explainable genetic variance is as such a natural objective function since the position with minimum residual variance can be defined as the location of a putative QTL. The total phenotypic variance can thus be decomposed as

$$V_{tot} = V_g + V_\epsilon. \tag{2}$$

This definition assumes that all genetic variance is attributable

to the QTL. The residual variance, V_r, at the QTL position is then simply $V_r = V_\epsilon$, or $V_r = V - V_g$. It is trivial to compute V, for any data set, using only the phenotype observations. Hence, we can use the function $f(x) = V - V_r(x)$ as the objective function, just as well as $V_r(x)$ (which is proportional to the RSS) can be used directly. The expected value of $V_r(x)$ can also be computed for any position x if we know the recombination frequency p between the position denoted by x and the actual QTL, as well as the full variance V_g attributable to the QTL.

Loci far apart on the same chromosome, or loci on different chromosomes, will exhibit no linkage at all and any match will be random. If, for simplicity, we assume a diallelic QTL in a backcross population with perfectly ideal allele frequencies, the probability p of a second indicator matching an ideal indicator at the QTL will be 0.5. If the two indicators are identical (infinitely close), then p will be 1.0. All other situations are somewhere in between[1].

If we now assume that the phenotype values for the two QTL genotypes are 1 and 0, the variance with a null model (average only) will be $0.5(1-0.5)^2+0.5(0-0.5)^2 = 0.25$. If we use an indicator of the real genotype with accuracy p, we get two symmetrical classes, each being a mixture of both actual genotypes. Below follows a derivation for the residual variance of the "high" class (dominated by individuals with a 1 phenotype). Due to the symmetric structure, this is equal to the total variance

$$V_r = p(1-\mu)^2 + (1-p)\mu^2 = p(1-2\mu+\mu^2) + (1-p)\mu^2, \quad (3)$$

$$\mu = 1.0(p) + 0.0(1-p) = p,$$

$$V_r = p - p^2.$$

Here, μ is the average value within the class as defined by the indicator. This is the "target" for the linear regression. As the variance is 0.25 under the null hypothesis, the relative reduction in variance possible through the indicator is given by

$$\frac{V - V_r}{V} = \frac{0.25 - V_r}{0.25}. \quad (4)$$

In an actual genome with a known mapping distance x between the loci, p is equivalent to the complement to the recombination fraction.

[1]This assumes ignoring the possibility that some locus due to selection pressure actually tends to show an inverted preferred heritage structure relative to the QTL. Furthermore, such a locus pair would not have a well-defined mapping distance, so this possibility is frequently ignored at an earlier step in the experiment setup, i.e. the construction of the marker map.

If we assume no recombination interference, we can relate this to mapping distances through the Haldane mapping function[9], with x in cM (centimorgan)

$$p = 1 - 0.5(1 - e^{-\frac{2x}{100}}) = 0.5 + 0.5e^{-\frac{2x}{100}}. \quad (5)$$

Inserting (5) for p in (3) and then inserting the result in (4), we arrive to a relative reduction in variance of

$$\frac{V - V_r(x)}{V} = e^{-\frac{4x}{100}}. \quad (6)$$

The result will be similar if a different mapping function is used [13]. Thus, the variance at any point x, measured as the distance from the QTL defined in (2), is

$$V_r(x) = V - V_g e^{-\frac{4x}{100}}. \quad (7)$$

3.2 A Bound on the Lipschitz Constant

The RSS, which is simply a scaling of the unexplained variance, has earlier been used as the objective function when DIRECT is applied to QTL analysis. The Lipschitz assumption is then translated into an assumption that the RSS is significantly correlated to the RSS in the (macroscopic) vicinity of that point, within the same chromosome. We now suggest an alternative objective function where we can derive a bound on the Lipschitz constant for an ideal infinite-size population. Later we will consider the possible deviations in actual finite-size populations.

Ignoring a scaling factor, the RSS is equivalent to the residual variance V_r. $V_r(x)$ is in this idealized case of a single QTL equivalent to (7).

Adding a constant will not affect the location of minima, so we can instead consider $g(x) = V_r(x) - V = -V_g e^{-\frac{|4x|}{100}}$. Furthermore, the function $g(x)$ is always negative, so $f(x) = -\ln(-g(x))$ will always be defined and share the locations of minima with $g(x)$.

$$\frac{df(x)}{dx} = \frac{d}{dx} - \ln(V_g e^{-\frac{|4x|}{100}}) = \pm 0.04 \quad (8)$$

Equation (8) directly shows that the derivative is bounded. At $x = 0$, the limits from both directions will also be within the bound.

We have now presented an objective function $f(x)$ which has minima at the correct locations and which also has a computable, bounded derivative. The use of $|x|$ in the definition of $g(x)$ is related to the fact that x is defined as the *distance* from the QTL, while

a position in the chromosome can naturally be both upstream and downstream from this position. It is possible to shift the function by introducing the true QTL position y, resulting in $f(x) = \ln V_g - 0.04|x - y|$.

We will now generalize the result above and demonstrate that the bound on K is maintained. First, we have assumed that all genetic variance was attributable to a single locus (at y). We can now assume a single-locus model for analysis, but that the true QTL, with respective components of V_g are represented as a vector $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, resulting in the following expression for the residual variance (assuming all y_i to be unlinked):

$$V_r(x) = V - \sum_{i=1}^{n} V_{g_i} e^{-\frac{4|x-y_i|}{100}} \tag{9}$$

If all QTL are indeed unlinked, the positions y_i relative to any reference will be $\pm\infty$, except for at most one y_j. Since linkage is transitive, the observation position x can at most be linked to a single QTL. Thus, (9) reduces to

$$V_r(x) = V - V_{g_j} e^{\frac{4|x-y_j|}{100}}, \tag{10}$$

and the derivative bound on (10) follows from the result in (8).

The next extension is to make the search landscape itself multi-dimensional, replacing the scalar x with a vector $\mathbf{x} = \{x_1, x_2..., x_m\}$. Assume that each x_i is defined with a point of reference in linkage with the corresponding QTL y_i. Modeled locations not in linkage with any true QTL will result in no explainable variance, and therefore do not need to be considered. As the numbering of both vectors is essentially arbitrary, all other cases are also symmetrical to this one. Using the earlier result, each x_i will only have a term for the corresponding y_i, as all other mapping distances $|x_i - y_j|$ would be $+\infty$. This results in

$$V_r(\mathbf{x}) = V - \sum_{i=1}^{n} V_{g_i} e^{-\frac{4|x_i-y_i|}{100}}. \tag{11}$$

We then have that

$$f(\mathbf{x}) = -\ln(-V_r(\mathbf{x}) + V), \tag{12}$$

where the logarithm does not affect the location of minima.

Equation (11) can be rewritten as

$$V_r(\mathbf{x}) = V - V_{g_j} e^{-\frac{4|x_j-y_j|}{100}} - \underbrace{\sum_{i=1, i\neq j}^{n} V_{g_i} e^{-\frac{4|x_i-y_i|}{100}}}_{C}, \tag{13}$$

for any arbitrary $j \in N, j \leq n$. Based on (13), (12) becomes

$$f(\mathbf{x}) = -\ln(V_{g_j} e^{-\frac{4|x_j - y_j|}{100}} + C). \tag{14}$$

From basic calculus, we know that $|\frac{d}{dx}\ln(e^{kx} + C)| = |\frac{ke^{kx}}{e^{kx}+C}| \leq \frac{d}{dx}\ln(e^{kx}) = \frac{ke^k x}{e^k x} = k$ for any $C \in R, C \geq 0$. Hence, all partial derivatives $\frac{\partial f}{\partial x_j}$ are confined by the bound given for the single-dimensional first derivative presented earlier. Thus, $f(\mathbf{x})$ as defined above will have a well-defined Lipschitz bound.

For a multi-locus case including interacting QTL (*epistasis*), the important observation to make is that the probability of a fully correct indication of the QTL genotype is p^n (if assuming all x_i approximately equal, or at least of the same order of magnitude), where p is again the probability of a correct indication at a specific locus at a specific one-dimensional distance x. The logarithm is then approximately given by

$$\ln V_G - \frac{4x_1 n}{100}. \tag{15}$$

This results in a constant derivative of 0.04 for a single locus, 0.08 for two interacting loci, 0.12 for three interacting loci, and so on. This is a very pessimistic approximation, basically assuming there are no detectable main effects at all, all genetic variance being attributable to the epistatic variance components when doing an orthogonal decomposition of variance [6].

For linked QTL, the picture is more complex. The effects from different QTL are, at least partially, confounded, as there is only a single variable (the indicator position within the linkage group), relating back to both components. Among other things, this means that the total explainable variance can be 0 at some point in the region between two linked QTL if the effects at the QTL have opposite signs. However, outside of the interval between the linked QTL, the behavior is completely identical to the presence of a single QTL at the position of the closest QTL in the set, with an effect equivalent to the combined average effect of all the linked QTL observed from that position. This can intuitively be understood from the memory-less nature of the exponential function.

4 Using a practical bound

The bound derived does not apply directly to the derivative of the actual residual variance in the data, but to the expected value of

the derivative, corresponding to the relation between the mapping distance and the expected number of crossover events. Depending on what recombinations are actually present (i.e. in which individuals, with accompanying phenotype values), the actual residual variance can, and will, be different from the relationship predicted. This is an effect of sampling, which should decrease with an increasing size in population and vanish at a theoretical infinite population size. However, experimental populations tend to be rather small, and thus the infinite-size approximation cannot be used directly.

There is also another class of errors, due to insufficiencies of the model. For example, the fact that the Haldane mapping function is not completely accurate can eventually cause the bound on the derivative to be broken. Even if the marker map and genotype probabilities in non-marker positions are constructed based on the Haldane function, the fact that the function is just an idealization of the actual biological recombination behavior implies that mapping distances are never completely additive, so estimates of recombination rates based on those mapping distances will not even be asymptotically accurate, sampling errors aside. The experimental populations used are also rarely a result of random mating (e.g. the presence of large sibsets from matings of the same parents in the F_2 generation), which means that the sampling errors will be more severe compared to an ideal model-conforming population of equal size. Any approach relying on the properties of the model to improve efficiency, ours included, will also be more sensitive in cases where the model is not fully valid.

The point of deriving a Lipschitz bound for the DIRECT algorithm is to determine a minimum bound on the values that might exist within a fixed-size neighborhood of a pre-defined point $\bar{x}$. The stochastic effects from sampling of phenotypes, and sampling of genotypes (specific recombination events vs. the asymptotically expected frequencies from mapping distances) need to be taken into account. The latter can possibly be handled by introducing a "virtual" distance measure, inflating actual distances to account for the tail end of the probability distribution for the actual number of recombination events. In this paper, we have found that a simpler approach, taking both effects into account by allowing a well-chosen fixed extra threshold, is easier to implement, while still highly efficient. Assume we have computed $q = f(\bar{x})$, where $\bar{x}$ is a genome location in some dimensionality n. We then propose to use a bound of the following form in practice:

$$f(\bar{x} + \bar{\delta}) \geq \min(q, c_0 - c_1 d) - 0.04D - c_2 \tag{16}$$

$$\bar{\delta} = (\delta_1, \ldots \delta_i, \ldots \delta_n) \tag{17}$$

$$D = \sum \delta_i \tag{18}$$

Here, D is the Manhattan distance in the genome, which is a distance measure compatible with eqn. (15). Also, c_0 is a constant describing a level above which sampling noise is so strong that any regular patterns break down. The constant c_2 can be seen as a correction for phenotype sampling, and c_1 as a correction for uncertainty in each genome dimension. In the following, we discuss how to choose c_0, c_1, and c_2, and whether they can be chosen in such a way that they both give a bound that for practical uses can be considered strict, and is still restrictive enough to allow significant acceleration when comparing DIRECT with an exhaustive search approach. First, we will elaborate on the identified sources of deviation from the bound, concluding with how these relate to the terms in our simplified bound for finite-size populations.

4.1 Virtual distances

As has been noted above, we need to get a practically usable bound that still retains the performance increases obtained.

Knowing the expectation value of the Lipschitz constant, we can give a one-sided 50% confidence bound. To achieve significant results, we need to estimate additional properties of the distribution of this constant.

A partial model of the uncertainty of recombination can be represented by a binomial distribution. Again assuming the presence of a single QTL and an indicator at some distance x cM, the explainable variance is described by (3), where the central component is p, which is the probability of the indicator allele matching the QTL allele. In the derivation in section 2, this was assumed to correspond to the mapping distance. If we view the mapping distance as a correct estimate of the recombination probability, the total count of correctly indicated individuals in a specific population is described by the distribution $Bin(n, p)$. From this, we can define the random variable for the actual indicator match frequency, $p' = \frac{Bin(n,p)}{n}$. As the relation between recombination frequency and mapping distance by Haldane's function is bijective, a value for p' can be transformed back into a "safe mapping distance" with a chosen threshold.

This virtual distance provides a guideline for reasonable bound estimates for the objective function. As an example, in a population

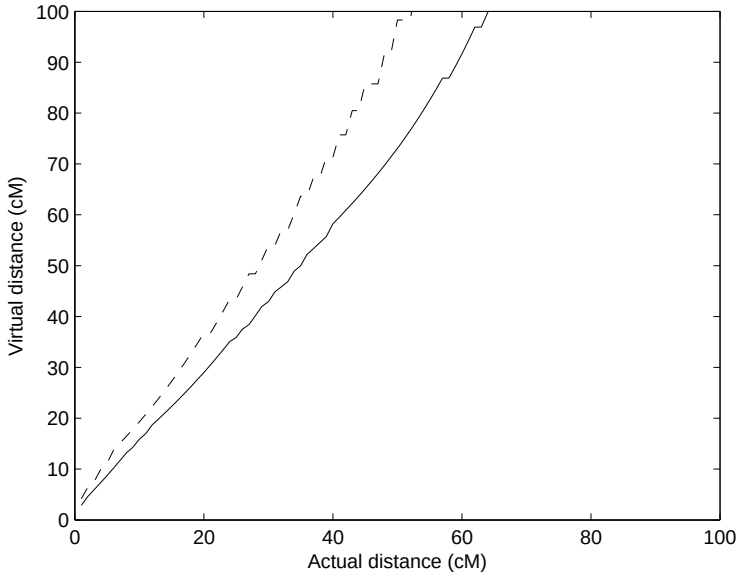

Figure 2: Virtual distances for 99% confidence in recombination counts versus actual distances, for a backcross with 500 (solid) and 200 (dashed) individuals, respectively.

with 500 individuals and a box evaluated at 25 cM from the actual QTL, this box will have an explainable variance that is no worse than that of an ideal 36 cM distance box, with a probability of 99%. Hence, the virtual distance is 36 cM, and this distance could be used in the Lipschitz bound calculations. Figure 2 shows how the virtual distance varies depending on population size and threshold value. Figure 3 shows the same data, but normalized to illustrate the ratio between real and virtual distances. This model is based on the assumption that all individuals of each genotype have the exact same phenotype, or at least that there are no sampling errors within the sets. These effects will introduce another noise component that is not compensated by (this realization of) virtual distances.

4.2 Sampling of trait values

If the heritability (explainable variance) of a putative QTL would be complete, 100%, the binomial distribution resulting in the model of virtual distances would account for all effects originating in the sampling of a finite-size population. However, many real traits of interest, even in heavily controlled experiments, have a total genetic

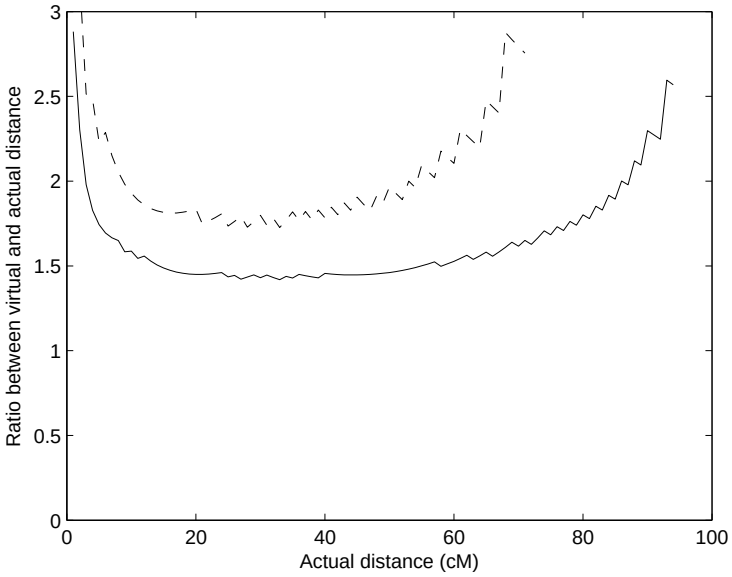

Figure 3: Virtual distances for 99% confidence in recombination divided by actual distances, showing relative increase (1 meaning identical distances). Shown for a backcross with 500 (solid) and 200 (dashed) individuals, respectively.

variance component of far less than 50%. Therefore, it is not enough to consider what proportions of individuals correspond to the true QTL genotype at a specific analyzed position. The recombination process will introduce a shift in genotype in a random sample of the individuals. The size of the sample is determined by the (virtual) recombination distance. However, the actual phenotype values of the individuals shifted is also relevant. If the samples of individuals within the genotype groups (as defined by their true QTL genotype) have means that tend *away* from the other genotype, e.g. if the sampled individuals are overrepresenting the tails of the total phenotype distribution, the observed within-class means will get closer to each other and thus the explainable variance will decrease further.

We can describe this by considering the set of all the population to be $\mathbf{P}$, of cardinality n. We once again consider the backcross, so we have individuals in subsets $\mathbf{P}_0, \mathbf{P}_1$, reflecting their actual genotype at some putative QTL. There is also observed genotype data at some distance away, resulting in a division along another axis, in all resulting in subsets $\mathbf{P}_{00}, \mathbf{P}_{01}, \mathbf{P}_{10}, \mathbf{P}_{11}$. For each set $\mathbf{P}_i$, there is a corresponding mean μ_i, cardinality n_i and variance σ_i^2. The genetic effect is $\mu_1 - \mu_0 = a$, and by translation we can assume $\mu_0 = 0, \mu_1 = a$. By scaling, we can assume the residual variance to be $\sigma_0^2 = \sigma_1^2 = 1$. We can also assume the distribution at the locus itself to be ideal, $n_0 = n_1 = n/2$. The previously introduced recombination frequency p can be used to determine the cardinality of the double-indexed subsets, $\{n_{01}, n_{10}\} = pn/2, \{n_{00}, n_{11}\} = (1-p)n/2$. In practice, the symmetry in size $n_{01} = n_{10}$ does not hold, due to the sampling effects of recombination partially modeled through virtual distances.

The sampling of values will be directly manifested in the values of μ_{01} and μ_{10}. In the expectation/median case, we will have simply $\{\mu_{01}, \mu_{10}\} = \{\mu_0, \mu_1\}$. According to the assumption that the residual variance is normal, we want to know the distribution for the mean of a random sample of specific size n_{01} from a finite size normal population of size n_{00}. Thus, the normal distribution for the sample mean is $\mu_{01} = N\left(\mu_0, \frac{1}{n_{01}} \frac{n_{00}}{n_0 - 1}\right)$, the second factor being the so-called Finite Population Correction Factor. We can continue the assumption $n_{01} = n_{10}$ by assuming $\Delta\mu = \mu_{01} - \mu_0 = \mu_1 - \mu_{10}$. The concept of an identical $\Delta\mu$, and the value at all being a single constant, is naturally a simplification, but this theoretical modeling could be extended. The sampling effects for recombination frequency, and the sampling effects due to varying phenotype means in the subsets of individuals actually showing recombinations over the distance considered, are closely related. At this time we do not have a full analytical expression for these interactions.

4.3 Motivating our approximate bound

The critical observation to motivate the approximate bound 16 is that the sampling effects are expected to be strong for small sets, i.e. short distances. When few individuals are sampled, the freedom is much greater. For very short distances, this is true. The question can be whether any recombination is occurring at all, and the change from e.g. 2 to 3 individual recombinations will affect the objective function radically. The Lipschitz bound rather assumes recombinations to be a continuous event, as would be the case in an infinite-size population. By including a constant threshold term in each dimension, we can account for both change in recombination counts, and an RSS-inflating or RSS-deflating random sampling of the identity of the recombining individuals.

Furthermore, sampling effects are generally not expected to all go in the same direction at the same time. The general safety threshold applied at all points should not scale with the total number of dimensions. This makes sense, since the real source of the trait value sampling uncertainty is originating from the environmental variability, which is not specific to our model-dimension. A correction might be needed for small population sizes, though.

Strong sampling effects due to small individual counts are not only present in the case of short distances, but also very long distances. The effects from a single individual recombining in the correct manner versus 50/50 proportions at unlinked positions can be drastic. Therefore, we also include the upper bound q, which should be chosen so that the explained variance found, compared to a model with no explainable variance (i.e. a $+\infty$ objective function value), is exceeding a few individuals. The chance for such a division increases with a higher dimensionality, which is why the upper threshold should decrease with the number of loci.

These three constants will need to be chosen based on population size, and population structure.

4.4 Application of the modified bound

With a higher genetic variance component, the global minimum will be more extreme. This means that, once a minimum is found, there is a much greater opportunity of aggressive pruning in the search space, based on box sizes and the derivative bound. DIRECT will find the true global minimum, and fewer iterations will be needed for traits with a higher heritability.

While attractive, this property also introduces a real problem, since the most frequent genome scan operation is often a run within a

random permutation test [5]. To get a significance threshold for 99%, it is prudent to do at least $2,000$ runs, compared to a single run for the actual trait data. Since each permutation test run will include a random permutation of the phenotype data, with unmodified genotype data, the expected heritability in those runs is very close to 0. Indeed, the purpose of the test is to determine the expected "false" heritability in random data similar to that actually analyzed.

However, as the purpose of doing the random permutations is to determine a significance threshold empirically, we will not need to locate the exact QTL in the random data. Rather, it is enough to get one single boolean value out of each run: is it possible to find a QTL set in the permuted dataset with a residual variance below that of the actual trait set, or not? The process outlined below can also be adapted to handle the case of several candidate traits with different variance levels. The distribution of outcomes is still completely discrete: the random set may show a higher apparent heritability than all candidates, lie in the region between two of them, or be inferior to all of them (which is generally expected to be the most common result).

This results in one trivial termination condition, even without a value of the Lipschitz constant K: if the residual variance detected in a random set reaches a value below the known residual variance in the trait set, then immediate termination is possible, with a result of true to the boolean test for that specific run. This scenario will be relatively rare, however (5% of the tests if the trait might hit the 95% threshold). Thanks to the value of K, we can also add the same constraint that was used in the search for the actual trait, i.e. boxes where the derivative bound precludes reaching a value below V_r (of the actual trait, not the random set) can be discarded. The only boxes we need to explore are those that still might allow a lower residual variance in the random data.

Thus, with the bound of K and the known V_r for the trait for which we are seeking significance, the same condition to discard boxes that were used in the end of the trait search can be applied from the start. This greatly accelerates the process, while simultaneously allowing a higher number of iterations to be spent in those permuted sets that actually might contain a random QTL above the intended threshold. This second observation is also important, since it might eliminate the problem shown in some earlier studies regarding a bias for DIRECT in random permutation tests [14].

5 Results

The issues of permutation testing found in previous use of DIRECT [14] was only resulting in a very small difference. Doing say $5,000$ permutations with a specific experimental dataset in order to validate our method might render a false positive, as the case where an exhaustive search finds a different optimum is rare.

Meanwhile, using a purely simulated dataset might hide non-ideal properties of experimental datasets, giving a validation of our bound which would not work out in practice. For example, the patterns of missing genotype data can give additional confounding and sampling effects.

For these reasons, we decided to use an unpublished experimental dataset with a combination of microsatellite and SNP markers and varying patterns of missing data. This dataset is described in [17]. Based on inferred haplotypes in the F_0 generation, derived using the tool presented in that paper, new population replicates were constructed and QTL with varying dimensionality simulated accordingly. These replicates could then be analyzed for main effects and permuted runs. By creating hundreds or thousands of replicates, with hundreds of permutations within each, we can expect to discover any deviations between exhaustive search and running DIRECT with a combined termination condition of minimum resolution equivalent to the exhaustive search lattice, and a pruning of impossible split candidates based on our modified Lipschitz-style bound in (16).

5.1 Values for the bound constants

Similar simulations were performed for 2, and 3 dimensions, with the same values for the constants found in (16). First, the "noise ceiling" c_0 needs to be chosen. We propose here a level close to the fraction $\frac{4}{n}$, meaning a portion of variance equivalent to explaining 4 individuals. For our population, including 765 individuals, this translates to 5.25. The choice of c_1 was 0.5, equivalent to choosing $\frac{4^d}{n}$ as the proportion of random explanation. For c_2, 0.50 was also used, but motivated as a safety distance, i.e. equivalent to 12.5 cM.

The criterion has also been verified for 1 dimension, but in one dimension the exhaustive search is so fast that replacing it is not a relevant goal.

5.2 Simulations

The simulations were performed on the isis cluster at the UPPMAX computational resource center, running as single threads on nodes

Table 1: *Size of validation tests for two and three dimensions. Time use for validating exhaustive search runs limits the possible size for 3D. Permutations were done per replicate.*

d	Heritability (h^2)	Number of replicates	Number of permutations
2	0.09	500	1000
3	0.14	500	100

with AMD Optreon 2220 CPUs. The code is parallel, but since a very high number of replicates were used, the serial version, with its higher memory locality and lack of synchronization overhead, was preferred.

For each run, first the non-permuted QTL model of specified size was fit, allowing all levels of interaction (each genotype-phenotype mean was a free parameter). Then, permutations were created. For exhaustive search, all possible candidate loci sets were tried in a 1 cM lattice. For DIRECT, the minimum resolution was that same lattice, but in addition the bound was used to avoid splitting of some boxes, if it was definite that the values within that box could not improve on the minimum found in the original main run. The optimization of terminating the full search when a minimum surpassing the one in the original was found, was not implemented. This occurrence is supposed to be rare, but would reduce the maximum number of function evaluations needed drastically. It would also be possible to stop after only a lower number of permutations, if it is already clear at that point that the original result will not be significant.

Table 1 presents the specific number of replicates, the simulated broad-sense heritability h^2, and the number of permutations done for each replicate. Table 2 presents average, minimum and median number of function evaluations for complete sets of main run plus permutations. Timings show that objective function evaluations exceed 90% of the time used in the DIRECT version. For actual significance tests, a higher number of permutations would be prudent, but in order to properly validate the bound within reasonable time, a higher number of replicates is supposed to contribute to a more varied test set.

Out of our 500 simulations used in 2 dimensions, 34 were less than 99% significant. If these are removed from the computed number of function evaluations, the average number of evaluations decreases to $7.78e5$. Acceleration compared to exhaustive search is only possible when the result from the main run rises above the noise floor established by the constants in our bound.

Table 2: *Number of function evaluations for two and three dimensions with exhaustive search and DIRECT, respectively. Minimum, mean, median and maximum number of function evaluations per full run of main QTL search followed by the number of permutations given in Table 1 are reported. A very low number of DIRECT runs resulted in full exhaustive searches. Note that the 3-dimensional scan used a lower number of permutations.*

d	Method	Min	Median	Avg	Maximum
2					
	Exhaustive	$7.88e6$	$7.88e6$	$7.88e6$	$7.88e6$
	DIRECT	$4.76e5$	$7.79e5$	$1.84e6$	$7.88e6$
3					
	Exhaustive	$9.99e8$	$9.99e8$	$9.99e8$	$9.99e8$
	DIRECT	$2.81e5$	$1.21e6$	$2.76e6$	$3.98e7$

6 Discussion

A hurdle against wide implementation of efficient global optimization algorithms for QTL searches have been the reluctance against methods which do not guarantee optimal results. The bound presented in this paper, and the underlying theory, is a step towards providing such guarantees.

Previous reported speedups for DIRECT in the QTL space have been several orders of magnitude. The work presented in this paper does not improve on those results. Instead, we provide an option for performing the QTL scan where accuracy and guaranteed results are of higher importance. We do so by letting DIRECT continue executing until a full exhaustive search has in some sense been performed, but with a pruning taking place removing boxes that are not possible optima. This allows DIRECT to be used not only for finding QTL candidates, but also to be reliably used in permutation testing to find the extreme end of the null hypothesis distribution. These computations can be used to compute significance thresholds as well as assessing the extreme value distribution, something which could form a basis for comparisons between models with different dimensionality and parametrization.

We finally propose to use a derivative bound which is of ad-hoc form, but this is based on a thorough analysis and reformulation of the objective function. This reformulation accelerates the performance of DIRECT, even when the pruning step is not added. The performance improvement is easily understood, as DIRECT will perform best if the function is linear (finding the top of a single triangle in only a few

iterations), and the transformation used will in general transform the objective function to be almost piecewise linear. This understanding of the expected local form of the objective function, when full genotype information is available, could also be used to better assess probable QTL locations in cases of partial and incomplete genetic information. We can also note that our empirically determined thresholds for c_0, c_1 correspond quite well to a correction for the number of degrees of freedom in the model. If this can be verified and mathematically proven, the general validity of the bound would be further ensured. The choice of c_2 will still be complicated, as it is used to correct for the non-ideal distribution of recombination events.

It should be noted that the performance for our version of DIRECT is dependent on the heritability. For a trait with no heritability at all, finding the true optimum can in principle only be done by an exhaustive search, as very limited correlations are expected in the objective function between loci. Previous incarnations of DIRECT would have failed in those cases, while our modified implementation is adaptive and will perform more function evaluations. If one knows beforehand that QTL with a heritability below some limit h^2_{min} are not relevant, then such information can be added to our pruning and give better performance even in those cases.

If the goal is to establish the significance of a QTL with very high significance, then our approach will excel. If most null hypothesis permutations for a dataset have a minimum that is inferior to that determined for the main model, the permuted DIRECT runs can exit after only a hundred or so function evaluations, even in multiple dimensions. This allows doing runs equivalent to $100,000$ or more permutations for loci where significance levels above 99.9% would be relevant.

References

[1] J. M. Alvarez-Castro and O. Carlborg. A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. *Genetics*, 176(2):1151–1167, 2007.

[2] M. Bogdan, J. K. Ghosh, and R. W. Doerge. Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. *Genetics*, 167(2):989–999, 2004.

[3] E. A. Carbonell, T. M. Gerig, E. Balansard, and M. J. Asins.

Interval mapping in the analysis of nonadditive quantitative trait loci. *Biometrics*, 48(1):305–315, 1992.

[4] O. Carlborg, L. Andersson, and B. Kinghorn. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics*, 155:2003–2010, 2000.

[5] G. Churchill and R. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–971, 1994.

[6] C. C. Cockerham. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39(6):859–882, 1954.

[7] R. Doerge. Mapping and analysis of quantitative trait loci in experimental populations. *Nature reviews-Genetics*, 3:43–52, 2002.

[8] E. Galerpin. The cubic algorithm. *J of Mathematical Analysis and Applications*, pages 635–640, 1985.

[9] J. B. S. Haldane. The combination of linkage values, and the calculation of distance between the loci of linked factors. *J Genet*, 8:299–309, 1919.

[10] C. S. Haley and S. A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–24, 1992.

[11] R. C. Jansen. Interval Mapping of Multiple Quantitative Trait Loci. *Genetics*, 135(1):205–211, 1993.

[12] D. Jones, C. Perttunen, and B. Stuckman. Lipschitzian optimization without the lipschitz constant. *J. Optimization Theory App*, 79:157–181, 1993.

[13] E. S. Lander and D. Botstein. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*, 121(1):185–199, 1989.

[14] K. Ljungberg, S. Holmgren, and O. Carlborg. Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. *Bioinformatics*, 20(12):1887–1895, 2004.

[15] R. Mladineo. An algorithm for finding the global maximum of a multimodal, multivariate function. *Mathematical Programming*, pages 253–271, 1987.

[16] C. Nettelblad, O. Carlborg, and J. M. Alvarez-Castro. Efficient algorithms for multi-dimensional global optimization in genetic mapping of complex traits. Technical report 2010-005, Division of Scientific Computing,Dept of IT, Uppsala University, 2010.

[17] C. Nettelblad, S. Holmgren, L. Crooks, and O. Carlborg. cnf2freq: Efficient determination of genotype and haplotype probabilities in outbred populations using markov models. In S. Rajasekaran, editor, *BICoB 2008*, volume 5462 of *Lecture Notes in Computer Science*, pages 307–319. Springer, 2009.

[18] J. Pinter. Globally convergent methods for n-dimensional multi-extremal optimization. *Optimization*, 17:187–202, 1986.

[19] B. Shubert. A sequential method seeking the global maximum of a function. *SIAM J. on Numerical Analysis*, pages 379–388, 1972.

[20] Z.-B. Zeng, T. Wang, and W. Zou. Modeling Quantitative Trait Loci and Interpretation of Models. *Genetics*, 169(3):1711–1725, 2005.