Computational Biology and Chemistry 34 (2010) 34–41

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

Assessing a multiple QTL search using the variance component model

Kateryna Mishchenko[a], Lars Rönnegård[b,*], Sverker Holmgren[c], Volodymyr Mishchenko[c]

[a] School of Education, Culture and Communication, Mälardalen University, Box 883, SE-721 23 Västerås, Sweden
[b] Department of Economics and Social Sciences, Statistic Unit, Dalarna University, Rodavagen 3, 78188 Borlange, Sweden
[c] Division of Scientific Computing, Department of Information Technology, Uppsala University, Sweden

ARTICLE INFO

Article history:
Received 31 March 2009
Received in revised form 7 December 2009
Accepted 8 December 2009

Keywords:
QTL mapping
REML
Variance component estimation
Average information matrix
Forward selection
Hessian approximation
Active set
Primal-dual method

ABSTRACT

Development of variance component algorithms in genetics has previously mainly focused on animal breeding models or problems in human genetics with a simple data structure. We study alternative methods for constrained likelihood maximization in quantitative trait loci (QTL) analysis for large complex pedigrees. We apply a forward selection scheme to include several QTL and interaction effects, as well as polygenic effects, with up to five variance components in the model. We show that the implemented active set and primal-dual schemes result in accurate solutions and that they are robust. In terms of computational speed, a comparison of two approaches for approximating the Hessian of the log-likelihood shows that the method using an average information matrix is the method of choice for the five-dimensional problem. The active set method, with the average information method for Hessian computation, exhibits the fastest convergence with an average of 20 iterations per tested position, where the change in variance components <0.0001 was used as convergence criterion.

1. Introduction

Quantitative trait loci (QTL) are regions on the genome that affect traits measured on a continuous scale. These traits are affected both by several genetic regions and by environmental factors. QTL detection has been a major field of research for several decades (Lynch and Walsh, 1998), where experimental data has shown to be of great importance and has given unique insights to the genetic architecture of quantitative traits (Carlborg and Haley, 2004).

Experimental data, resulting in high power for QTL detection, may be derived by crossing two breeds that are expected to differ genetically. The relationship between trait values and genotypes can be analyzed after two generations of controlled breeding. These experiments are referred to as F_2 intercrosses. A standard statistical tool for analyzing F_2 intercrosses is the simple regression model, which assumes no genetic variation between individuals of the same breed (Haley and Knott, 1992; Broman, 1997; Ljungberg et al., 2002). However, there is often some genetic variation within the two breeds, and this variation may be modeled as a random effect in a more advanced variance component model (Rönnegård and Carlborg, 2007; Perez-Enciso and Varona, 2000).

In a variance component QTL analysis, all the founders of the F_2 intercross are assumed to be unrelated with genes randomly sampled from an outbred population. QTL mapping based on a variance component model is computationally demanding. The computational procedure consists of an *inner problem* and an *outer problem*. In the *inner problem* a variance component model is fitted at a given position in the genome. The value of the likelihood ratio statistic is calculated for this model and is subsequently used in the outer problem. The *outer problem* consists of finding the position, among all tested positions, with highest likelihood ratio value. Hence, the dimensionality of the inner problem is equal to the number of variance components to be estimated, whereas the number of dimensions in the outer problem is given by the number of QTL that we wish to fit simultaneously.

Calculation of the likelihood ratio statistic requires variance component estimation, where restricted maximum likelihood (REML) estimation is used to ensure unbiased estimates of variance components. Variance component estimation consists of a non-linear optimization problem where the computation of the objective function and its derivative is rather costly. Fast variance component estimation programs developed for animal breeding problems (e.g. ASReml (Gilmour et al., 2002) and DMU (Madsen and Jensen., 2008)) are often used in QTL analysis (e.g. Rowe et al., 2009). These variance component estimation programs have been developed to analyze large data sets ($\approx 10^6$ observations) and to compare a moderate number of models (usually < 10). In QTL analysis, however, the size of the data sets are moderate ($\approx 10^3$ observations), whereas the number of models compared are large

* Corresponding author. Tel.: +46 241 30034.
E-mail addresses: kateryna.mishchenko@mdh.se (K. Mishchenko), lrn@du.se (L. Rönnegård), sverker.holmgren@it.uu.se (S. Holmgren).

1476-9271/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2009.12.001

(usually > 1000). Consequently, the variance component estimation program developed for QTL analysis needs to be robust so that the algorithm converges for all fitted models. Once the robustness has been verified, further efforts can be made to reduce the computational cost of the calculations.

Variance component estimation algorithms have also been developed for QTL analysis in human pedigrees consisting of independent families (for instance in the SOLAR software (Almasy and Blangero, 1998)), where the size of each family is small. This gives a block-diagonal structure in the variance component model which results in significant simplifications in the computational algorithms and convergence problems does not seem to be an issue. In the current paper, we focus on large complex pedigrees that do not have this simple structure.

A major problem in variance component estimation is that the parameter space is constrained (since variances are > 0). This fact needs to be accounted for by employing established techniques for constrained optimization (e.g. primal-dual and active set methods (Forsgren and Gill, 1998)). Convergence for variance components on, or close to, the parameter boundary may otherwise not be guaranteed. A commonly used algorithm for variance component estimation in animal breeding is the *average information REML* (Johnson and Thompson, 1995), which has been implemented in the ASReml and DMU software (Gilmour et al., 2002; Madsen and Jensen., 2008). The focus has been on speed rather than estimating parameters close to, or on, the boundary in this algorithm, since it is primarily developed for animal breeding applications. For parameter estimates on, or outside, the parameter boundary DMU combines average information REML with an expectation-maximization (EM) algorithm to enable convergence within the parameter space. ASReml does not allow zero variances and sets a lower limit to the variance components equal to a small positive value. To our knowledge, these methods do not guarantee convergence within the parameter space.

Previously we have investigated the possibilities of using active set and primal-dual methods for the simplest possible model with two variance components (Mishchenko et al., 2008), a QTL variance and a residual variance, where the given correlation structure for the QTL variance is low rank or can be approximated by a low rank correlation structure (Rönnegård et al., 2007). Fast computation of projection matrices and matrix inversions has also been derived for the two variance component problem (Mishchenko and Neytcheva., 2009).

In QTL analysis, it is also common to include random polygenic effects as well as QTL effects (Lynch and Walsh, 1998). The correlation structure for polygenic effects (i.e. the additive relationship matrix) is full rank and adds an additional complexity to the variance component estimation problem. Furthermore, possible interaction effects between QTL (i.e. *epistasis*) at several positions on the genome is important to include in the analysis (Carlborg and Haley, 2004). Hence, problems with more than two dimensions for the *inner problem* needs to be studied and will put higher requirements on the computational robustness for the variance component estimation algorithm.

The aim of the current paper is to investigate optimization techniques for the inner problem based on the active set and primal-dual algorithms for constraint optimization, and we apply these schemes for QTL mapping models with 3–5 variance component problems. We wish to find a scheme which is numerically robust and efficient. Moreover, the performance of the schemes using different methods for approximating the Hessian of the log-likelihood are compared. The methods are tested on published data (Carlborg et al., 2006), where the previous analysis was based on a regression model (Haley and Knott, 1992) assuming no within-breed variation. We briefly discuss differences and similarities between our results and these earlier analyses.

2. The restricted maximum likelihood approach

In this section, we consider models where a one-dimensional genome scan is performed for estimating 3–5 variance components. We start by considering a model of a single QTL and additional polygenic effects. Polygenic effects are the combined effects of many genes at different loci each having a small effect (Lynch and Walsh, 1998), whereas a QTL effect is the effect of a restricted part of the genome. The correlation structure for polygenic effects is given by the *additive relationship matrix* and is calculated from pedigree information, whereas the correlation structure for the QTL effect is given by the *identity-by-descent* (IBD) matrix. Elements of the IBD matrix are estimated from pedigree and marker information (Lynch and Walsh, 1998).

2.1. A single QTL and polygenic effects (3D-SCAN)

Variance component analysis for single QTL and polygenic effects is based on a general linear mixed model,

$$y = Xb + Z_1 u_1 + Z_a a + e, \tag{1}$$

where y is a vector of n individual phenotypes of a normally distributed trait, X is an $n \times n_f$ design matrix for fixed effects, Z_1 is an $n \times n_r$ design matrix for random effects, b is a vector of n_f unknown fixed effects, u_1 is a vector of n_r unknown random effects for an individual QTL, Z_a is a $n \times n_a$ design matrix for additional polygenic effects, a is a vector of n_a random polygenic effects, and e is a vector of n residuals of random effects. All random effects are assumed to be normally distributed.

For the QTL analysis setting we also assume that the entries in e are identically and independently distributed and that there is a single observation for each individual. Let Π_1 be the IBD matrix and A the additive relationship matrix, then the variance–covariance matrix for (1) is

$$V = \Pi_1 \sigma_1^2 + A\sigma_a^2 + I\sigma_e^2, \tag{2}$$

where σ_1^2 is the variance of the random QTL effect, σ_a^2 is the variance of polygenic effects and σ_e^2 is the residual variance.

In REML estimation, the parameters σ_1^2, σ_a^2, σ_e^2 are obtained as maximizers of the restricted likelihood function l of the observed data y. This is done by minimizing the restricted log-likelihood function $L(\Sigma)$ associated with (1),

$$L = -2\ln(l) = C + \ln(\det(V)) + \ln(\det(X^T V^{-1} X)) + y^T P y. \tag{3}$$

Here, C is normalizing constant, Σ is the vector of variance components and the projection matrix P is defined by

$$P = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}. \tag{4}$$

In summary, we solve the inner problem, i.e. determine the estimates of $\sigma_1^2, \sigma_a^2, \sigma_e^2$, by solving the optimization problem:

$$\min L(\Sigma) \tag{5}$$

s.t.

$$\sigma_1 \geq 0, \quad \sigma_2 \geq 0, \quad \sigma_3 > 0. \tag{6}$$

Below, we use the notation $\Sigma = (\sigma_1^2, \sigma_a^2, \sigma_e^2) = (\sigma_1, \sigma_2, \sigma_3)$. To determine the main QTL and its effect we need to solve the outer problem and search for the best model fit over the genome. The position τ_0 with the best likelihood value is the most likely position of the main QTL.

2.2. Forward selection for an additional QTL (4D-SCAN)

To solve the problem of finding several QTL, a simultaneous search for them should in principal be performed. For

example, when searching of two QTLs, the outer problem is a two-dimensional global optimization problem. This drastically increases the computational complexity, since the outer problem must be solved in a hypercube with dimensionality defined by the number of QTL under consideration. An efficient approach for detecting multiple QTLs in a single, multi-dimensional search has been developed for the standard least-squares model (Ljungberg et al., 2004, 2005). There, a modified version of the global optimization scheme DIRECT and a hybrid global–local approach were used to solve the outer problem.

In practice, a technique based on forward selection has traditionally been widely used for searching for several QTLs. Here, a one-dimensional search for the position of the main QTL is first performed, and the computed effect of the QTL is subsequently included in the model. Using the extended model, a new one-dimensional search for another QTL is made. An attractive feature of this technique is that only one-dimensional outer problems need to be solved, and the computational complexity is normally much smaller than if a multi-dimensional simultaneous search is performed. However, it is not clear how accurate this technique is for general models. For example, it has been shown, in Carlborg and Haley (2004) that a forward selection scheme can be ineffective in detecting interacting QTL. Hence, the results from forward selection analysis should be used with some care.

In the current paper, we postpone the introduction of simultaneous search for several QTL to future research, and we focus on using the forward selection procedure which gives a one-dimensional outer problem. In this case, an additional random QTL effect u_2 is added to the model (1):

$$y = Xb + Z_1u_1 + Z_aa + Z_2u_2 + e, \tag{7}$$

where the corresponding variance–covariance matrix of y now is

$$V = \Pi_1\sigma_1^2 + A\sigma_a^2 + \Pi_2\sigma_2^2 + I\sigma_e^2. \tag{8}$$

Here, Π_2 is the IBD-matrix for the putative QTL at position τ_0 and σ_2^2 is the variance of u_2.

The estimates for the forward selection model (7) are obtained as minimizers of the problem:

$$\min L(\sigma_1, \sigma_2, \sigma_3, \sigma_4) \tag{9}$$

s.t.

$$\sigma_1 \geq 0, \quad \sigma_2 \geq 0, \quad \sigma_3 > 0, \quad \sigma_4 \geq 0. \tag{10}$$

This is a four-dimensional optimization problem where $\Sigma = (\sigma_1^2, \sigma_a^2, \sigma_e^2, \sigma_2^2) = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$.

2.3. Forward selection for an additional QTL and interaction effects (5D-SCAN)

An interaction effect between the two QTL in (7) can be added to model pair-wise epistatic interaction between them. The corresponding models are

$$y = Xb + Z_{12}u_{12} + Z_1u_1 + Z_2u_2 + Z_aa + e, \tag{11}$$

where the variance–covariance matrix of y is given by

$$V = \Pi_1\sigma_1^2 + A\sigma_a^2 + \Pi_2\sigma_2^2 + \Pi_{12}\sigma_{12}^2 + I\sigma_e^2, \tag{12}$$

and Π_{12} is calculated as the Hadamard product between Π_2 and Π_1 (Stern et al., 1996; Rönnegård et al., 2008). A large difference in the likelihood ratio of models (7) and (11) suggests that the QTL at positions τ_0 and τ do not act additively and that the interaction effect between the QTL is significant.

The minimization problem for finding the estimates $(\sigma_1^2, \sigma_a^2, \sigma_e^2, \sigma_3^2, \sigma_{12}^2) = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5)$ is now five-dimensional:

$$\min L(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) \tag{13}$$

s.t.

$$\sigma_1 \geq 0, \quad \sigma_2 \geq 0, \quad \sigma_3 > 0, \quad \sigma_4 \geq 0, \quad \sigma_5 \geq 0. \tag{14}$$

3. Optimization methods for the inner problem

In this section, we review two local optimization schemes for constrained problems; the active set and the primal-dual methods. We also introduce three approaches for approximating the Hessian of the log-likelihood.

Both the active set and primal-dual schemes are of Newton type, and hence based on the ability to compute exact or approximative derivatives of the objective function. The gradient of the log-likelihood can be computed analytically as a function of the matrices V and P, see e.g. Lynch and Walsh (1998):

$$\frac{\partial L}{\partial \sigma_i} = \mathrm{tr}\left(\frac{\partial V}{\partial \sigma_i}P\right) - y^TP\frac{\partial V}{\partial \sigma_i}Py, \quad i = 1, \ldots, m. \tag{15}$$

Here, m is the number of variance components in the model.

The Hessian can also be expressed using an analytical formula:

$$H(\sigma_i, \sigma_j) = -\mathrm{tr}\left(\frac{\partial V}{\partial \sigma_i}P\frac{\partial V}{\partial \sigma_j}P\right) + 2y^TP\frac{\partial V}{\partial \sigma_i}P\frac{\partial V}{\partial \sigma_j}Py, \quad i, j = 1, \ldots, m. \tag{16}$$

3.1. The active set method

We use the active set method described in Mishchenko et al. (2008), given by

$$p^{k+1} = -N[N^T \cdot \nabla^2 L(\Sigma^k) \cdot N]^{-1} \cdot N^T \nabla L(\Sigma^k),$$
$$\Sigma^{k+1} = \Sigma^k + \alpha^k \cdot p^{k+1}. \tag{17}$$

The optimality condition is checked by computing the Lagrangian multipliers λ at the potential optimum Σ^*:

$$\lambda^k = A_r^T \nabla L(\Sigma^*), \tag{18}$$

where A_r is right inverse of the matrix A.

We use a line search procedure where the current step α^k is, if needed, reduced so that the next point lies exactly on the relevant constraint.

3.2. The primal-dual interior point method

The primal-dual method is also based on Newton's method, using both primal (computation of Σ) and dual (computation of the Lagrangian multipliers λ) steps. We employ the nonlinear primal-dual method where the iterative scheme is given by

$$\Sigma^{k+1} = \Sigma^k + \alpha^k \Delta\Sigma^k, \qquad \lambda^{k+1} = \lambda^k + \alpha^k \Delta\lambda^k \tag{19}$$

and current step $(\Delta\Sigma_k, \Delta\lambda_k)$ is determined as a solution of the following system of equations:

$$\begin{pmatrix} \nabla^2 L(\Sigma^k) & -A(\Sigma^k)^T \\ \Lambda^k \cdot A(\Sigma^k) & M(\Sigma^k) \end{pmatrix} \cdot \begin{pmatrix} \Delta\Sigma^k \\ \Delta\lambda^k \end{pmatrix} = -\begin{pmatrix} \nabla L(\Sigma^k) - A(\Sigma^k)^T \cdot \lambda^k \\ M(\Sigma^k) \cdot \lambda^k - \mu \cdot \mathbf{1} \end{pmatrix}, \tag{20}$$

where $M(\Sigma^k)$ is a diagonal matrix with the constraints $m(\Sigma^k)$ on the diagonal, Λ^k is a diagonal matrix with the Lagrangian multipliers on the diagonal, and $A(\Sigma^k)$ is a matrix of the gradients of constraints. Since we have linear bound constraints, $A(\Sigma)$ is constant, containing only the values 1, 0 and -1.

In the algorithm, λ_0 is chosen as $\lambda_0 = \mu_0/m(\Sigma_0)$, where μ_0 and $m(\Sigma_0)$ are given at the initial step and μ_k forms a decreasing sequence ($\mu_k \geq 0$). The line search procedure is the same as for

the active set method. Here, the condition $\lambda_k \geq 0$ is checked and if it does not hold we choose α^k to fulfil it.

3.3. Computation of the Hessian of the log-likelihood

The direct usage of (16) is costly, and it may be beneficial to approximate the true Hessian matrix by some entity which is cheaper to compute. In the average information REML method, the Hessian is substituted by the average information matrix, which is the average of the Hessian (16) and its expected value. The average information matrix is given by

$$AI(\sigma_i, \sigma_j) = y^T P \frac{\partial V}{\partial \sigma_i} P \frac{\partial V}{\partial \sigma_j} P y \quad i, j = 1, \ldots, m. \tag{21}$$

For two-dimensional optimization problems, corresponding to a single QTL without any additional effects, the average information REML method (based on Newton iteration) has been shown to be efficient for those cases when minimum was found inside the feasible region (Mishchenko et al., 2008).

We use three approaches for approximating the Hessian using: average information (Johnson and Thompson, 1995), BFGS, and average information–BFGS combined. BFGS is a numerical approximation of the Hessian matrix (Nocedal and Wright, 1999) as described below.

(1) *Average information*: Here, we approximate the Hessian by the average information matrix (21). For our problems, we have $m = 3$, 4 and 5. Thus, the average information matrix is 3×3, 4×4 or 5×5.
(2) *BFGS with damping*: Here, we use the BFGS updating technique (Nocedal and Wright, 1999), including a damping parameter θ which is used to ensure that the Hessian approximation is always positive definite, see Algorithm 1. We use two versions of this approach: in the first implementation (BFGS), the initial value of the approximated Hessian is set to the identity matrix ($H^0 = I$).
(3) *Average information–BFGS*: Here, we use the average information formula for computing the initial value ($H^0 = AI$) and otherwise use the BFGS algorithm as described in Algorithm 1.

Algorithm 1. Hessian updating by the damped BFGS formula.

$$s^k = \Sigma^k - \Sigma^{k-1}$$
$$y^k = \nabla L(\Sigma^k) - \nabla L(\Sigma^{k-1})$$
if
$$s^{k^T} \cdot y^k \geq 0.2 \cdot s^{k^T} \cdot H^{k-1} \cdot s^k \tag{22}$$
then
$$\theta^k = 1$$
else
$$\theta^k = (0.8 \cdot s^{k^T} \cdot H^{k-1} \cdot s^k)/(s^{k^T} \cdot H^{k-1} \cdot s^k - s^{k^T} \cdot y^k)$$
end if
$$r^k = \theta^k \cdot y^k + (1 - \theta^k) \cdot H^{k-1} \cdot s^k \tag{23}$$
$$H^k = H^{k-1} - \frac{H^{k-1} \cdot s^k \cdot s^{k^T} \cdot H^{k-1}}{s^{k^T} \cdot H^{k-1} \cdot s^k} + \frac{r^k \cdot r^{k^T}}{s^{k^T} \cdot r^k} \tag{24}$$

Note that the computational complexity for the BFGS update is much smaller than the complexity for computing the average information matrix. The convergence criterion used in all algorithms was $\left|\Sigma^k - \Sigma^{k-1}\right| < 10^{-4}$.

4. Numerical results

We use data from a selection experiment in chicken (Carlborg et al., 2006). The number of observations is $n = 767$, and the fixed effects are population mean and sex effect. The genome consists of 20 chromosomes of lengths 50–410 cM. To perform the genome scan, the IBD matrices are computed on a mesh with 5 cM step length and the best model fit is found by an exhaustive search. The 5% genome-wide significance thresholds is 3.84 for model (1) and 5.99 for model (11), which are approximations given by George et al. (2000). In Carlborg et al. (2006), a QTL with main effect was found near the 85 cM position on chromosome 7 using a simple regression model.

We perform the genome scan three times: the first search (3D-SCAN) is applied using model (1) with polygenic effects, the second (4D-SCAN) is applied to model (7), and in the final scan (5D-SCAN) we evaluate model (11). When performing 3D-SCAN, we solve the three-dimensional local optimization problems and detect the position τ_0 for the main QTL. This result is used as input for 4D-SCAN, where four-dimensional local optimization problems are solved. Finally, in 5D-SCAN, five-dimensional optimization problems are solved.

We have verified that for each scan, all methods used produce very similar log-likelihood values and variance components. Hence, all tested schemes are reasonably robust and produce accurate solutions. Below, we compare the performance of the methods and also present a more detailed study of robustness (i.e. sensitivity to

Table 1
Statistics based on the number of iterations performed by the different optimization schemes for 3D-SCAN, 4D-SCAN and 5D-SCAN.

	Method					
	Active set			Primal-dual		
	AI [a]	AI–BFGS [a]	BFGS [a]	AI [a]	AI–BFGS [a]	BFGS [a]
3D-SCAN						
Max ♯iter	32	60	23	55	68	32
Min ♯iter	13	30	8	21	42	15
Mean ♯iter	16.03	47.33	13.55	26.58	52.81	22.27
Std ♯iter	2.07	5.47	4.22	3.78	5.01	2.63
4D-SCAN						
Max ♯iter	29	78	42	49	85	32
Min ♯iter	13	49	9	23	50	18
Mean ♯iter	17.04	64.97	16.41	27.04	69.34	24.07
Std ♯iter	2.14	6.06	5.48	3.24	6.16	2.34
5D-SCAN						
Max ♯iter	51	98	60	93	100	53
Min ♯iter	11	63	9	23	51	20
Mean ♯iter	18.18	81.33	19.11	32.50	85.81	27.28
Std ♯iter	4	6.65	6.27	6.11	6.59	3.39

[a] Hessian.

Table 2
Cpu timing for 5D-SCAN (s).

	Method					
	Active set			Primal-dual		
	AI [a]	AI–BFGS [a]	BFGS [a]	AI [a]	AI–BFGS [a]	BFGS [a]
Aver time per iter	1.057	1.037	1.034	1.056	1.036	1.034
Aver time	19.2	19.7	84.1	34.3	28.2	88.7

[a] Hessian.

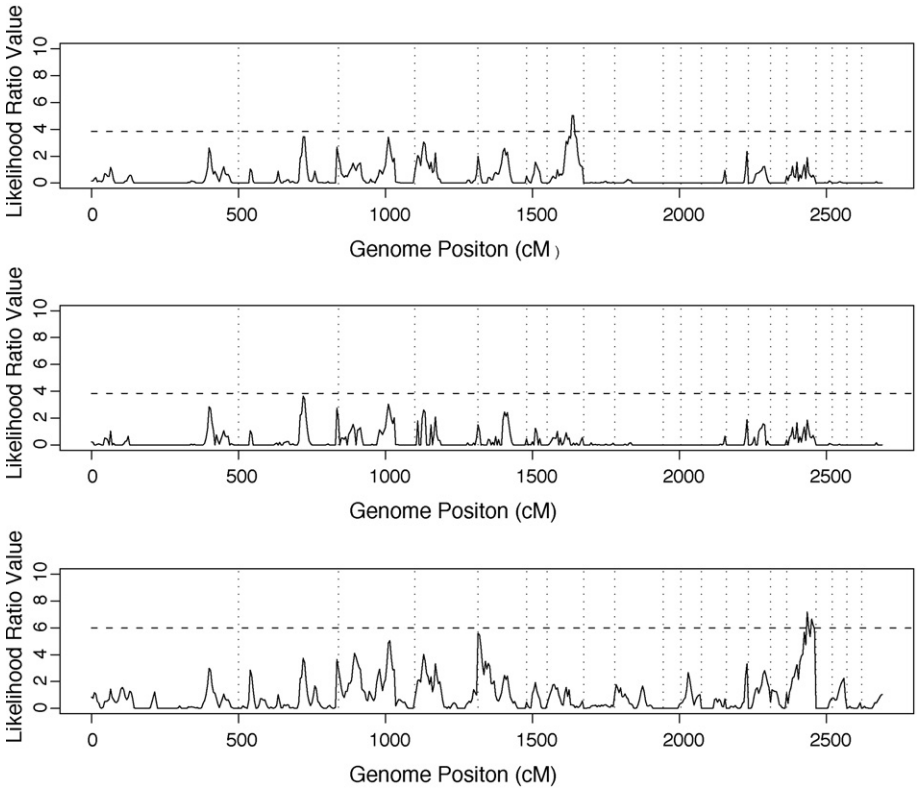

Fig. 1. Likelihood ratio values for the three scans. Borders between chromosomes are shown as dotted vertical lines. The approximate 5% genome-wide significance threshold is given by the dashed horizontal line.

problem parameters such as the position in the genome and the dimensionality of the optimization problem).

The active set scheme, together with the average information Hessian approximation results in few iterations and a small variation among positions (Table 1). For all schemes, the dependence on the dimensionality of the optimization problem is small.

In Table 2, CPU timings for 5D-SCAN are presented as average over all positions in the genome. We conclude that the active set method for constraining the parameter space together with an approximation using the average information matrix should be used for variance component QTL models with up to five variance components.

The highest peak of the likelihood ratio curve for 3D-SCAN was found to be located at position 85 cM on chromosome 7 (Fig. 1), which is consistent with the earlier results using the least-squares model (Carlborg et al., 2006). The curve for 4D-SCAN, derived using the forward selection model (7), is similar to the one for 3D-SCAN with the exception that the peak corresponding to the main QTL is no longer present. The highest peak in the curve for 4D-SCAN is located at 220 cM on chromosome 4. Hence, this is the location of the potential second QTL when interaction effects are not included in the model. Also note that, for these data, the highest peak using (7) is the second highest peak from the genome scan using model

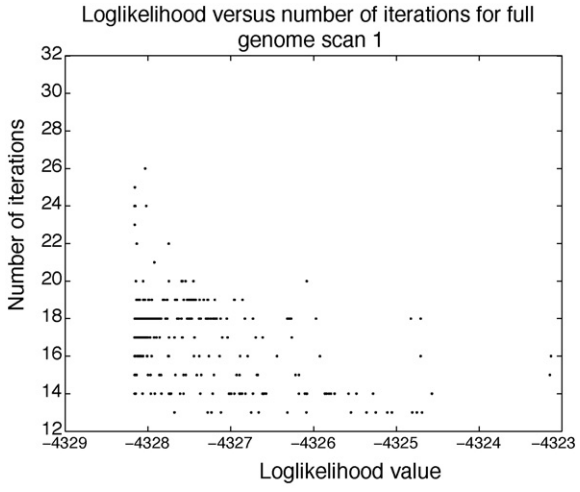

Fig. 2. The number of iterations for 3D-SCAN as a function of the maximal value of the log-likelihood.

(1) and that the differences in the top two graphs of Fig. 1 are small apart from the fact that the peak on chromosome 7 (which was included in the model of (7)) is not found in the middle graph of Fig. 1.

The highest peak in the likelihood ratio curve for 5D-SCAN is located at 65 cM on chromosome 16. Hence, this is the location of the potential second QTL when interaction effects are included in the model. The large differences between 4D-SCAN and 5D-SCAN are expected since large epistatic effects were found for these data in Carlborg et al. (2006).

The number of iterations used increases for positions on the genome with small likelihood ratios. Positions with small likelihood ratios have either flat log-likelihood surfaces or QTL variance components close to zero. This was also observed in our results where the number of iterations increases as the maximal log-likelihood value decreases (Fig. 2 gives results for 3D-SCAN with the active set + average information method).

5. Summary

We present efficient and robust optimization schemes for solution of the REML problem for variance component estimation in the setting of QTL analysis. We consider three important model settings, resulting in REML optimization problems of dimension 3–5. The most robust and efficient scheme is the active set method combined with approximating the Hessian with the average information matrix, and possibly updating this approximation during the iterations using the BFGS procedure with damping. We also consider using the primal-dual optimization method, but this scheme is slightly less efficient, especially when it is taken into account that a linear system of equations must be solved using QR factorization at each iteration.

Although the speed of the calculations has been of secondary importance in our study, it is essential the methods are fast enough to perform full genome scans. In the Appendix we derive formulas for increasing the computational speed and show how to compute the inverse to the variance-covariance matrix using an algorithm which requires $C_1 kn^2$ arithmetic operations, compared to $C_2 n^3$ operations if a standard factorization is used.

The aim of our paper was to derive efficient and robust algorithms for variance component based QTL analysis, but there are also some interesting similarities and differences between our results and those found by Carlborg et al. (2006). In the study of Carlborg et al. (2006) a regression model was used assuming no within-breed variation. Using the same data they found a QTL network between chromosomes 1–4, 7 and 20. In our study, the main peaks of model (1) are found on chromosomes 1–4, 7 and 16. Furthermore, the QTL showing clearest interaction effects with the QTL on chromosome 7 in our study is located on chromosome 16. These differences may be explained by the differences in the assumptions of the two QTL models.

The optimization methods presented produce accurate solutions in a robust way, and using the schemes presented in our paper enables the usage of more complex variance component models for genetic analysis in a production setting. Using the proposed active set scheme, it is possible to regularly solve REML optimization problems with 2, 3, 4, or 5 variance components in about 20 iterations. This was not possible earlier using the standard average information-REML scheme, where convergence problems may occur (MacGregor, 2003). It could be argued, however, that convergence problems for positions on the genome with estimates on the boundary of the parameter space are not of interest, because such positions do not generally contain a QTL. This argument is not satisfactory, since the reason for divergence can not be uniquely identified and the result can not be interpreted. General variance component software should not produce such non-conclusive results.

Appendix A. Numerical algorithms for evaluating derivatives of the log-likelihood

The efficiency of the solution of the inner problem depends both on the efficiency of the optimization scheme as such, i.e. the number of iterations required, and the efficiency of the algorithms used for computing the derivatives used inside the optimization loop. To optimize the overall procedure, we consider efficient algorithms for the derivative computations.

The algorithms used for the numerical linear algebra operations needed for computing the gradient and Hessian approximation should be adapted to the formulation of the problem, the properties of the covariance matrices, and the number of variance components included in the model. In Johnson and Thompson (1995); Callanan and Harville (1991), efficient approaches have been developed for the case when the problem is formulated and solved in terms of mixed model equations. These methods cannot be utilized for QTL analysis problems, since the IBD matrices are normally singular. Efficient algorithms for computing the log-likelihood derivatives for simple QTL models with two variance components, including a singular IBD matrix, were developed in Mishchenko et al. (2007); Mishchenko and Neytcheva. (2009). In this case, the variance–covariance matrix V is given by

$$V = \Pi_1 \sigma_1 + I\sigma_2, \tag{25}$$

where Π_1 is a positive semi-definite IBD matrix of size $n \times n$ and σ_1, σ_2 are the variances of the random QTL effects and the residual.

The main observation leading to the methods developed in Mishchenko et al. (2007); Mishchenko and Neytcheva. (2009) is that the derivatives of the log-likelihood are given by expressions that include V^{-1}. Thus, a factorization of V as a product of constant matrices, or matrices that can be easily updated, can potentially be used to reduce the cost of computing the log-likelihood derivatives.

In Mishchenko et al. (2007), the specific structure of the IBD matrix was used in two ways. First, the case when the IBD matrix is given explicitly as a product of low rank matrices,

$$\Pi_1 = 1/2\tilde{\Pi}\tilde{\Pi}^T, \tag{26}$$

was considered. Here, $\tilde{\Pi}$ is a rectangular matrix of size $n \times m$, $m \gg n$. In Rönnegård and Carlborg (2007), it has been shown that at locations in the genome where complete genetic information is available, a representation like (26) can always be determined. Also, the rank of the IBD matrix as such locations depends only on the size of base population. The case when the genetic information is not fully complete was also considered. For such problems the IBD matrix can normally still be well approximated by a low-rank matrix. To compute such an approximative representation, a truncated spectral decomposition was used, where

$$\Pi_1 \cong W_t \Lambda_t W_t^T. \tag{27}$$

Here, W_t is the truncated matrix of eigenvectors of Π_1 of size $n \times k$, $k < n$, corresponding to k largest eigenvalues forming the diagonal of the truncated matrix of eigenvalues Λ_t.

The factorizations (26) and (27) were subsequently used to compute the inverse of the variance-covariance matrix (25) by employing the Woodbury formula. The projection matrix (4) was computed and used for the computations of the factors involved in the formulas for the gradient and the average information matrix.

In Mishchenko and Neytcheva. (2009), an alternate algorithm for the case when the IBD matrices are given in general form was presented. There, the spectral decomposition of the IBD matrices was also used, but neither V^{-1} nor P were computed explicitly.

Instead, the action of P, i.e. Py, was computed directly by using a factorization of the variance-covariance matrix given by

$$V = W \cdot D \cdot W^T, \tag{28}$$

$$D = \Lambda\sigma_1 + I\sigma_2, \tag{29}$$

where Λ and W are the diagonal matrix of eigenvalues and the orthogonal matrix of eigenvectors of Π_1. In the algorithm, the factorization (28) is subsequently used to compute the action of V and P by solving systems of equations with multiple right-hand sides. As a result of the special structure of the factorization (28) and the presence of constant factors when forming Py, the algorithm developed in Mishchenko and Neytcheva. (2009) in general has a lower computational complexity that the method developed in Mishchenko et al. (2007). Moreover, it was shown that the most efficient method for computing the derivatives of the log-likelihood is to combine the idea of approximation of the IBD matrices using truncated spectral decomposition and direct evaluation of the action of P without forming V^{-1} and P. The method has a smaller computational complexity already if the rank of the approximated IBD matrix in (27) is reduced by 20%. Finally, the trace computations occurring in the formulas for the derivatives were simplified by utilizing the eigenvalues of the inverse of the variance-covariance matrices in (28).

The advantage of the factorization (28) combined with solving a system of equations with multiple right hand sides can be fully utilized only for models where the variance–covariance matrix has the simple form (25). However, the idea of factorizing the IBD matrix into a product of a matrices of low rank can still be utilized for models where the matrix V has a more complicated structure, like the models that have been introduced earlier in the paper. Below, we present some generalizations of the methods mentioned above methods to two cases where the components in the matrix V have specific structure.

A.1. A model with polygenic effects

For problem (5), the variance–covariance matrix V is the sum of three terms, $V = \Pi_1\sigma_1 + A\sigma_2 + I\sigma_3$, where Π_1 is a positive semi-definite IBD matrix and A is a positive definite matrix of polygenic effects. In this case, the matrix A^{-1} is also explicitly known.

Straight-forward algorithms for computing the log-likelihood can be based on computing a factorization, e.g. the Cholesky factorization, of the matrix V in each iteration. The factorization can also be used for computing the entries of (15) and (21) explicitly. This corresponds to computing P, Py, $\Pi_1 Py$, APy according to the algorithm developed in Mishchenko and Neytcheva. (2009). The remaining terms, $\mathrm{tr}(P)$,$\mathrm{tr}(\Pi_1 P)$ and $\mathrm{tr}(AP)$ are computed by explicit matrix–matrix multiplications and trace computations. An alternative is to compute the spectral decomposition of V, which will slightly simplify the trace computations since the eigenvalues of V can be used for computing the trace of V^{-1}. However, using this type of approach is computationally demanding. The factorization of V in each iteration has complexity Cn^3, and the computation of P involves matrix–matrix multiplications which are also of the same complexity. Moreover, the advantage of knowing the inverse of A explicitly is not utilized.

We now consider an alternative scheme which reduces the costs of computing (15) and (21). Here, we use the matrix A^{-1} explicitly, and we exploit a factorization of V^{-1} based on the spectral decomposition described in Mishchenko et al. (2007). The algorithm is described in detail in (Mishchenko and Neytcheva., 2009, Section 3).

We start with the spectral decomposition of $\Pi_1\sigma_1 + I\sigma_2$, which gives

$$V = WDW^T + A\sigma_3, \tag{30}$$

where W and D are defined above. Then, we exploit the Woodbury formula for computing the inverse of V,

$$V^{-1} = \sigma_3^{-1}A^{-1} - \sigma_3^{-2}A^{-1}W[D^{-1} + \sigma_3^{-1}W^TA^{-1}W]^{-1}W^TA^{-1}. \tag{31}$$

Note that, despite the fact that A^{-1} is known and the terms $W^TA^{-1}W$ and W^TA^{-1} are constant and should be computed only once in (31), the matrix $[D^{-1} + \sigma_3^{-1}W^TA^{-1}W]$ of size $n \times n$ still needs to be inverted (or the correspondent system of equations should be solved) in each iteration in the optimization loop, due to the presence of the matrix D which depends on σ_1 and σ_2. So, the efficiency of the procedure for computing V^{-1} depends on how we solve the following system of equations with multiple right-hand sides:

$$[D^{-1} + \sigma_3^{-1}W^TA^{-1}W]x = b, \tag{32}$$

where x is an unknown matrix and b is known. Here, several possibilities can be considered. We suggest to employ the truncated spectral decomposition of the matrix Π_1. Here, V and V^{-1} can be written as

$$V \cong W_tD_tW_t^T + A\sigma_3, \tag{33}$$

$$V^{-1} \cong \sigma_3^{-1}A^{-1} - \sigma_3^{-2}A^{-1}W_t[D_t^{-1} + \sigma_3^{-1}W_t^TA^{-1}W_t]^{-1}W_t^TA^{-1}, \tag{34}$$

where W_t is the truncated matrix of eigenvectors of Π_1 and $D_t = \Lambda_t\sigma_1 + I_t\sigma_2$.

By employing (32), we arrive at a smaller system:

$$[D_t^{-1} + \sigma_3^{-1}W_t^TA^{-1}W_t]x = b. \tag{35}$$

If $k \ll n$, solving this system, using e.g. Cholesky factorization, is significantly cheaper than solving (32).

In summary, the procedure for computing the approximate inverse of V^{-1} expressed by formula (35) is:

(1) Compute $\tilde{A} = A^{-1} \cdot W_t$ and $\hat{A} = W_t^T \cdot \tilde{A}$.
(2) Compute $D_t^{-1} + \sigma_3^{-1}\tilde{A}$.
(3) Solve the system of equations with multiple right-hand sides:
$$[D_t^{-1} + \sigma_3^{-1}\hat{A}] \cdot E = \tilde{A}, \tag{36}$$
where E is a matrix of size $n \times k$.
(4) Compute $\tilde{A} \cdot E$.
(5) Compute $\sigma_3^{-1}A^{-1} - \sigma_3^{-2}\tilde{A} \cdot E$.

Using this algorithm, the computational complexity for evaluating V^{-1} in each iteration is $2n^2k + 4nk^2 + 1/3k^3$, plus $n^3 + 2n^2k$ for the spectral decomposition and computing the product $A^{-1}W_t$, which is done once at the start of the whole computational procedure.

The computations of the matrix P is done according to the method described in Section 3.2 of Mishchenko and Neytcheva. (2009). The only modification needed is that the factor $M = V^{-1} \cdot X$ should be computed directly as a matrix–matrix product or as a product of V^{-1} presented in (33), depending on size k of matrix W_t.

Since the matrix P is computed explicitly, $\mathrm{tr}(P)$ is easily obtained. $\mathrm{tr}(PA)$ is computed using the representation (4) and V^{-1}, using formula (34):

$$\mathrm{tr}(AP) = \mathrm{tr}(A \cdot V^{-1}) - \mathrm{tr}(A \cdot V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}). \tag{37}$$

The fist term in (37) is

$$\mathrm{tr}(A \cdot V^{-1}) = \mathrm{tr}(\sigma_3^{-1}I - \sigma_3^{-2}W_t \cdot E), \tag{38}$$

where E is computed as a solution of equations according to (36).

The second term of (37) is computed multiplying A by $V^{-1}X$, and then by multiplying the result by $(X^TV^{-1}X)^{-1}X^TV^{-1}$. Here, $(X^TV^{-1}X)^{-1}X^TV^{-1}$ and $V^{-1}X$ were computed already in the computation of P and can be reused. The total computational complexity

for the trace is of order $4n^2r$, where $r \equiv n_f$ is the number of columns in X.

For the computation of $tr(\Pi_1 \cdot P)$ we again use (4) and (34). We start with computing the matrix-matrix product $\Pi_1 \cdot V^{-1}$ and compute its trace. The second term is computed by performing the matrix-matrix multiplication $\Pi_1 V^{-1} X$ and then multiplying the result by $(X^T V^{-1} X)^{-1} X^T V^{-1}$. The computational complexity is again of order of $4n^2r$.

The complete algorithm for the model with a single QTL and polygenic effects is given below:

Algorithm 2. Computation of gradient DL and average information matrix AI.

INITIALIZATION

1. Compute factors W, Λ for spectral decomposition of matrix Π_1: $\Pi_1 = W\Lambda W^T$.
2. Compute factors W_t, D_t for truncated spectral decomposition of matrix $\Pi_1\sigma_1 + I\sigma_2$: $\Pi_1\sigma_1 + I\sigma_2 \cong W_t D_t W_t^T$.
3. Compute factors $A^{-1}W_t$ and $W_t^T \cdot A^{-1} W_t$.

ITERATION LOOP

4. Compute $\tilde{V}^{-1}$ using formula (34). 5. Compute P
6. Compute Py, $\Pi_1 Py$, APy directly as matrix–vector products.
7. Compute $PAPy$ and PPy.
8. Compute $\mathrm{tr}(P)$, $\mathrm{tr}(AP)$.
9. Compute the gradient $\mathrm{DL} = -\begin{pmatrix} \mathrm{tr}(PA) - (Py)^T \cdot APy \\ \mathrm{tr}(P) - (Py)^T Py \\ \mathrm{tr}(PB) - (Py)^T BPy \end{pmatrix}$.
10. Compute the average information matrix

$$AI = \begin{pmatrix} (P\Pi_1 Py)^T \cdot \Pi_1 Py & (P\Pi_1 Py)^T \cdot (Py) & (P\Pi_1 Py)^T \Pi_1 Py \\ (P\Pi_1 Py)^T \cdot (Py) & (PPy)^T \cdot (Py) & (PPy)^T \Pi_1 Py \\ (P\Pi_1 Py)^T Py & (PAPy)^T Py & (PAPy)^T APy \end{pmatrix}.$$

The efficiency of the approach described above depends on the truncation index k. When the IBD matrix Π_1 can be well approximated by a low-rank matrix, the approach based on the truncated spectral decomposition will be efficient.

Additionally, we point out that the cases when σ_3 is equal or close to zero should be considered separately and the whole computational procedure is altered due to more simple structure of the variance–covariance matrix V.

References

Almasy, L., Blangero, J., 1998. Multipoint quantitative trait linkage analysis in general pedigrees. American Journal of Human Genetics 62, 1198–1211.

Broman, K.W., 1997. Identifying quantitative trait loci in experimental crosses. Ph.D. Thesis.

Callanan, T.P., Harville, D.A., 1991. Some new algorithms for computing restricted maximum likelihood estimates of variance components. Journal of Statistical Computation and Simulation 38, 239–259.

Carlborg, Ö., Haley, C.S., 2004. Epistasis: too often neglected in complex trait studies? Nature Reviews Genetics 5, 618–625.

Carlborg, Ö., Jacobsson, L., Åhgren, P., Siegel, P., Andersson, L., 2006. Epistatsis and the release of genetic variation during long-term selection. Nature Genetics 38, 418–420.

Forsgren, Anders., Gill, P.E., 1998. Primal-dual interior methods for nonconvex nonlinear programming. SIAM Journal on Optimization 8, 1132–1152.

George, A.W., Visscher, P.M., Haley, C.S., 2000. Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. Genetics 156, 2081–2092.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J., Thompson, R., 2002. Asreml User Guide. VSN International Ltd., UK.

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

Johnson, D.L., Thompson, R., 1995. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. Journal of Dairy Science 8 (2), 449–456.

Ljungberg, K., Holmgren, S., Carlborg, Ö., 2002. Efficient algorithms for quantitative trait loci mapping problems. Journal of Computational Biology 9 (6), 793–804.

Ljungberg, K., Holmgren, S., Carlborg, Ö., 2004. Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. Bioinformatics 20, 1887–1895.

Ljungberg, K., Mishchenko, K., Holmgren, S., 2005. Using DIRECT for a multidimensional global optimization problem arising during genetic mapping of quantitative traits. Submitted. Alo available as Technical Report 2005-035. IT, Uppsala University.

Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc.

MacGregor, S., 2003. Genetic linkage mapping in complex pedigrees. Ph.D. Thesis.

Madsen, P., Jensen, J., 2008. An User's Guide to DMU: A Package for Analysing Multivariate Mixed Models. University of Aarhus, Aarhus, Denmark.

Mishchenko, K., Holmgren, S., Rönnegård, L., 2008. Newton-type methods for REML estimation in genetic analysis of quantitative traits. Journal of Computational Methods in Sciences and Engineering 8, 53–67.

Mishchenko, K., Neytcheva, M., 2009. New algorithms for evaluating the log-likelihood function derivatives in the AI-REML method. Communications in Statistics 38, 1348–1364.

Mishchenko, K., Holmgren, S., Rönnegård, L., 2007. Efficient implementation of the AI-REML iteration for variance component QTL analysis. Technical Report 2007-4. Research Report Mälardalen University.

Nocedal, J., Wright, S.J., 1999. Numerical optimization. Springer Verlag, New York.

Perez-Enciso, M., Varona, L., 2000. Quantitative trait loci mapping in f2 crosses between outbred lines. Genetics 155, 391–405.

Rönnegård, L., Carlborg, Ö., 2007. Separation of base allele and sampling term effects gives new insights in variance component QTL analysis. BMC Genetics 8 (1).

Rönnegård, L., Mishchenko, K., Holmgren, S., Carlborg, O., 2007. Increasing the efficiency of variance component quantitative trait loci analysis by using reduced-rank identity-by-descent matrices. Genetics 176, 1935–1938.

Rönnegård, L., Pong-Wong, R., Carlborg, Ö., 2008. Defining the assumptions underlying modeling of epistatic QTL using variance component methods. Journal of Heredity 99, 421–425.

Rowe, S.J., Pong-Wong, R., Haley, C.S., Knott, S.A., de Koning, D.J., 2009. Detecting parent of origin and dominant qtl in a two-generation commercial poultry pedigree using variance component methodology. Genetics Selection Evolution 41, 6.

Stern, M.P., Duggirala, R., Mitchell, B.D., Reinhart, L.J., Sivakumar, S., Shipman, P.A., Uresandi, O.C., Benavides, E., Blangero, J., O'Connell, P.P., 1996. Evidence for linkage of regions on chromosome 6 and 11 to plasma glucose concentrations in Mexican Americans. Genome Research 6, 724–734.