BIOINFORMATICS

Vol. 20 no. 16 2004, pages 2792–2798
doi:10.1093/bioinformatics/bth331

Qxpak: a versatile mixed model application for genetical genomics and QTL analyses

M. Pérez-Enciso[1,2,*] and I. Misztal[3]

[1]Institut Catalá de Reçerca i Estudis Avançats, Pg Lluis Companys 23, 08010 Barcelona, Spain, [2]Departament de Ciència Animal i del Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain and [3]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA

Received on February 13, 2004; revised on April 16, 2004; accepted on May 5, 2004
Advance Access publication May 27, 2004

ABSTRACT

Motivation: Current methodology and software for quantitative trait loci (QTL) analyses do not use all available information and are inadequate to deal with the huge amount of QTL analyses to be needed in forecoming genetical genomics' studies.

Results: We show that a mixed model statistical framework provides a very flexible tool for QTL modeling in a variety of populations, be it a cross between inbred lines, a within population study, or experiments involving a mixture of populations or crosses. The software allows multitrait and multiQTL analyses, inclusion of infinitesimal genetic value and a batch multitrait option suitable for genetical genomics studies. It also allows massive association studies between single nucleotide polymorphisms and the trait(s) of interest.

Availability: A software (Qxpak), together with a manual and example files, is freely available for research purposes. So far, the compiled program is available for linux systems, the windows version will follow soon. See http://www.icrea.es/pag.asp?id=Miguel.Perez

Contact: miguel.perez@uab.es

INTRODUCTION

The mapping of quantitative trait loci (QTL) is now feasible due to the vast amount of DNA polymorphisms that is being uncovered in all species of interest. Traditionally, quantitative trait loci analyses have been carried out in well-designed experiments, like crosses between inbred lines or within family designs (Liu, 1998). Specific software program is used in each design. For instance, QTL cartographer by Z.B. Zeng and coworkers (http://statgen.ncsu.edu/qtlcart/index.php) allows only data from crosses between inbred lines. Another popular software, QTL express (Seaton *et al*., 2002), has different modules, each appropriate for specific designs, e.g. within family analysis or crosses between inbred lines, but not both. Generally, the software available is limited in modeling flexibility, e.g. multitrait models are not usually implemented or it is not possible to include an infinitesimal genetic effect. Furthermore, there is currently no public software that allows analysis of crosses between outbred lines, i.e. when there is genetic variation between as well within the line. Often, the number of QTL fitted or the number of chromosomes analyzed is limited in available programs. Similarly, the specificities of sex chromosomes are not dealt with. For a recent review on QTL analysis challenges and weblinks containing software available, see Abiola *et al*. (2003).

In addition, the recent advent of microarray technology has spurred the massive search of polymorphisms affecting the amount of mRNA level in the cell (Brem *et al*., 2002; Schadt *et al*., 2003) in what has been called 'genetical genomics' (Jansen and Nap, 2001). This poses new challenges both in terms of computing requirements and in modeling strategies.

Using different approaches for different designs is not only cumbersome but also it is not efficient and theoretically unsatisfactory. It may result in less power and less insight into the genetic architecture of the trait. Here, we present a coherent methodology for QTL analyses that is also suitable for genetical genomics studies. The method is based on the mixed model theory, which provides a flexible and elegant modeling tool.

The Qxpak package presented implements multitrait, multiQTL options, and can be applied to populations of any complexity, using all marker and pedigree information jointly. Different models per trait can be fitted and missing data is allowed for automatically. QTL effects can be modeled as fixed, random or mixed. Sex chromosome-linked QTL can also be analyzed, as dosage compensation and different chromosome lengths can be accommodated.

*To whom correspondence should be addressed.

METHODS

Suppose two breeds, A and B, with genetic effects (g) normally distributed as $g_A \sim N(\mu_A, \sigma_A^2)$ and $g_B \sim N(\mu_B, \sigma_B^2)$, respectively. Now, assume that a quantitative trait has been recorded in a population with an arbitrary pedigree complexity, where individuals can be 'purebred' from either A or B populations, F_1, F_2 or any other combination (e.g. recombinant inbred lines, backcross, advanced intercross and so on). A general explicative model is

$$\mathbf{y} = \mathbf{Xb} + \sum_{k=0}^{N_q} \mathbf{Zg}_k + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is a vector containing the recorded performances, $\mathbf{b}$ contains the fixed effects to be estimated, $\mathbf{g}_k$ contains the genetic (QTL) effects for any of the N_q QTL affecting the trait. By convention, we take $\mathbf{g}_0$ to stand for the infinitesimal genetic effects, i.e. the genetic effects not accounted for by individual QTL. Finally, $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices that relate observations to the parameters in the $\mathbf{b}$ and $\mathbf{g}$ vectors, and $\mathbf{e}$ is the residuals' vector. Typically, $\mathbf{Z}$ is a diagonal matrix with elements 1 at position (i, i) if i-th individual has a record, 0 otherwise. If there are several traits or repeated measures for the same individual and trait, $\mathbf{Z}$ is block diagonal.

The model in Equation (1) is termed 'mixed' because it contains fixed effects, such as sex or age, and random effects, such as the genetic effects, $\mathbf{g}$. Statistical theory for mixed models is well developed (McCulloch and Searle, 2000) and theory dictates that we also have to specify the distribution of the random variables, i.e. their means and variances (see Appendix). In the case of the QTL effects, $\mathbf{g}_k$, the expected value of the i-th indivual at k-th locus is

$$\begin{aligned} E(g_{ik}) = {} & P(g_{ik}^1 \in A, g_{ik}^2 \in A)\mu_{AAk} \\ & + P(g_{ik}^1 \in B, g_{ik}^2 \in B)\mu_{BBk} \\ & + [P(g_{ik}^1 \in A, g_{ik}^2 \in B) \\ & + P(g_{ik}^1 \in B, g_{ik}^2 \in A)]\mu_{ABk}. \end{aligned} \tag{2}$$

Here $P(g_{ik}^1 \in U, g_{ik}^2 \in W)$ is the probability that alleles from k-th QTL at paternal and maternal haplotypes are of breed U and W origins, μ_{WZk} is the mean genetic effect of individuals having received a U and W origin alleles at locus k. The variance of $\mathbf{g}_k$ is a matrix, $\mathbf{G}_k$, that contains the covariance between the i-th and j-th genetic values at k-th locus. The covariance between i-th and j-th genetic values is

$$\begin{aligned} \mathrm{Cov}(g_{ik}, g_{jk}) = {} & \frac{1}{2}\sum_{h=1}^{2}\sum_{h'=1}^{2} P(g_{ik}^h \equiv g_{ik}^{h'} \mid g_{ik}^h \in A)\sigma_{Ak}^2 \\ & + \frac{1}{2}\sum_{h=1}^{2}\sum_{h'=1}^{2} P(g_{ik}^h \equiv g_{ki}^{h'} \mid g_{ik}^h \in B)\sigma_{Bk}^2, \end{aligned} \tag{3}$$

where $P(g_{ik}^h \equiv g_{ik}^{h'} \mid g_{ik}^h \in U)$ is the probability of alleles g_{ik}^h and $g_{jk}^{h'}$ being identical by descent (IBD) and from origin U, superscript h stands for the paternal or maternal phases, numbered 1 or 2, respectively, and σ_{Uk}^2 is the variance of genetic effects of U origin at locus k.

In order to compute the likelihood and carry out standard statistical tests, it suffices to compute quantities (2) and (3) at any desired genome positions for all individuals and plug them into the likelihood function. It is important to notice that exactly the same computing strategy is followed irrespective of the pedigree complexity, number of QTL or traits. For instance, a cross between inbred lines can be modeled setting all elements in (3) to zero. Alternatively, we set elements in (2) to zero in a within population analysis because all genetic values have the same mean, logically, we need to estimate a single genetic variance $\sigma^2 = \sigma_A^2 = \sigma_B^2$ in (3). If we are interested in testing imprinting, the same formulas hold but the maternal ($h = 2$) or paternal ($h = 1$) coefficients $P(\cdot)$ are set to zero. Qxpak currently allows 22 modeling options for each QTL.

A particular case occurs when the QTL lies in the differential part of the sex chromosome (X in mammals, Z in birds). At least in mammals, probably also in birds (Ellegren, 2002), dosage compensation exists, which means that genetic effects are different according to whether the QTL allele lies in a female or in a male. We have proposed to include a dosage compensation parameter, ψ, to account for this differential effect (Pérez-Enciso *et al.*, 2002). As a consequence, we need to define different genetic covariances between males, between females, or between male and female. The package accommodates the theory, fully presented elsewhere (Pérez-Enciso *et al.*, 2002). A different issue is sex × QTL interaction, which has been shown to be a common phenomenon, specially for fitness related traits (Leips and Mackay, 2002). Our software also allows testing this interaction or, in general, any interaction between a covariate and a class effect. More complex interactions (epistasis) can equally be modeled but are not implemented in the software yet, see discussion below.

An important novelty of our software is that multitrait options are fully implemented. Multivariate mixed model techniques are well known and developed but have not been applied yet to QTL analyses, although least squares approaches have been published (Knott and Haley, 2000). The advantage of mixed model methods lies in its flexibility compared to least squares. For instance, mixed model theory can accommodate easily missing data, as well as different models per trait (Henderson, 1984), which is not possible in a least squares' approach. In a multivariate setting, we need to define the QTL variances for each trait plus their covariances, i.e.

$$\mathbf{G}_{\text{trait1, trait2}} = \mathbf{G}_0 \otimes \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix},$$

where $\mathbf{G}_0$ contains the $P(\cdot)$ coefficients in equation (3), $\otimes$ stands for the Kronecker product (Searle, 1982), the σ_{ij}^2 term

is the genetic covariance between traits i and j. We have assumed for the sake of simplicity a within breed analysis ($\sigma^2_{ij} = \sigma^2_{Aij} = \sigma^2_{Bij}$) but the same principle applies for separate breeds.

One of the interests of fitting multitrait models lies in disentangling whether a QTL that affects two traits simultaneously is due to a single QTL (pleiotropy hypothesis) or to two distinct QTL (linkage hypothesis) that lies in the same genome region. When the QTL effects are modeled as fixed, ($\sigma^2_k = 0$), these two alternatives can be tested by fitting the same QTL for both traits or fitting two distinct QTL, one for each trait. The difference in likelihoods between the two competing models can be tested using a likelihood ratio test with one degree of freedom if only the additive effect is fitted or two if also the dominant effect is included (Knott and Haley, 2000). This can be done easily with Qxpak (software usage section below). The issue of the significance level is discussed later.

It should be noted, however, that the situation is more complex when the QTL effects are random: the linkage and pleiotropy models are not hierarchical in this case and thus the likelihood ratio test cannot be applied. This occurs because the pleiotropy model contains the locus position, the genetic variances plus the covariance between traits, whereas the linkage model contains two loci positions but does not include the genetic covariance. Thus, other criteria must be used like the Bayesian information criterion or Akaike's information criterion.

We foresee that the unprecedented amount of data provided by microarray technology will be accompanied by numerous studies to dissect the genetic basis of each of the mRNA levels measured (Jansen, 2003). One of the requirements, thus, will be to automate numerous QTL studies, and our package allows this. Although scripts can be written to automate multiple analyses with other QTL packages, Qxpak can fit several of these analyses into one keeping a large modeling flexibility, e.g. an infinitesimal genetic effect or several QTL can be fitted if desired. The only limitation is that the same model should be used for all mRNA levels. This is due to input simplicity rather than computing constraints because it is difficult for the user to specify thousands of potentially different models.

Finally, one of the most active issues of research now in QTL studies is the discovery of causal mutations through the scan of multiple single nucleotide polymorphisms (SNPs). Our software allows the automation of successive SNP tests. Each SNP is tested in turn and whose inclusion in the model gives the highest likelihood is selected as the most plausible causal mutation. There also exists the possibility of testing the additive and/or dominant effects.

ALGORITHM AND IMPLEMENTATION

The algorithm consists of two main steps: first, the IBD probabilities, i.e. the terms $P(\cdot)$, in Equations (2) and (3), are computed and second, the likelihood is maximized. The algorithm used to compute IBD probabilities is a Monte Carlo Markov chain (MCMC) algorithm largely based on the ideas presented (Pérez-Enciso *et al.*, 2000) but with significant improvements: the algorithm updates several marker phases simultaneously thus making convergence faster, especially when markers are tightly linked; and missing marker information is allowed by using the closest informative marker available. The IBD files are saved in disk so they can be reused in later analyses, saving computing time.

IBD coefficients depend on the genome positions at which they are fitted. The algorithm first identifies all possible positions needed and compute a likelihood for every combination, e.g. if there are three QTL fitted, each in one chromosome of lengths 40, 80 and 60 cM and the positions are scanned every cM, there will be $40 \times 80 \times 100 = 240\,000$ different likelihoods to be maximized. Thus, it is imperative to have an efficient algorithm to maximize the likelihood.

In the second step, the mixed model equations (MME, see Appendix) are built according to the desired model and maximum-likelihood estimates are obtained via the EM algorithm. The EM formulas for single traits are:

$$\hat{\sigma}^2_k = \frac{\hat{\mathbf{g}}'_k \mathbf{G}_k^{-1} \hat{\mathbf{g}}_k + \hat{\sigma}^2_e \,\mathrm{tr}[\mathbf{G}_k^{-1}(\mathbf{Z}'\mathbf{Z} + \mathbf{G}_k^{-1}/\hat{\sigma}^2_k)^{-1}]}{n_k}, \quad (4)$$

where $\hat{\sigma}^2_k$ is the estimate of genetic variance for the k-th random effect, $\hat{\mathbf{g}}_k$ contains the prediction of the k-th QTL genetic values (see Appendix), $\hat{\sigma}^2_e$ is the residual variance estimate, 'tr' stands for trace (Searle, 1982) and n_k is the rank of $\mathbf{G}_k$ matrix (i.e. the number of individuals in the pedigree). The residual variance is estimated via

$$\hat{\sigma}^2_e = \frac{\mathbf{y}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \sum_k \mathbf{Z}\hat{\mathbf{g}}_k)}{n_r}, \quad (5)$$

where $\hat{\mathbf{b}}$ contains the estimates of the fixed effects (see Appendix) and n_r is the number of records. These two formulas need to be applied iteratively until convergence. These formulas are easily extended to multivariate models when the individual has measures for all traits. When there are missing traits, the formulas are more complex, see documentation of BLUPF90 package described below. Our program fully supports multivariate models with any pattern of missing data.

The program also provides the likelihood, which can be used to test the desired effects (e.g. 2 QTL versus 1 or 1 QTL versus none) via a likelihood ratio test. Minus twice the log-likelihood is given by

$$-2\ln(L) = \text{constant} + \ln|\mathbf{R}| + \ln|\mathbf{G}| + \ln|\mathbf{C}^\mathrm{R}| + \mathbf{y}'\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \sum_k \mathbf{Z}\hat{\mathbf{g}}_k\right),$$

where $\mathbf{R}$ and $\mathbf{G}$ are the variance of the residuals and of the genetic effects, respectively, $\mathbf{C}^\mathrm{R}$ is the submatrix of the MME

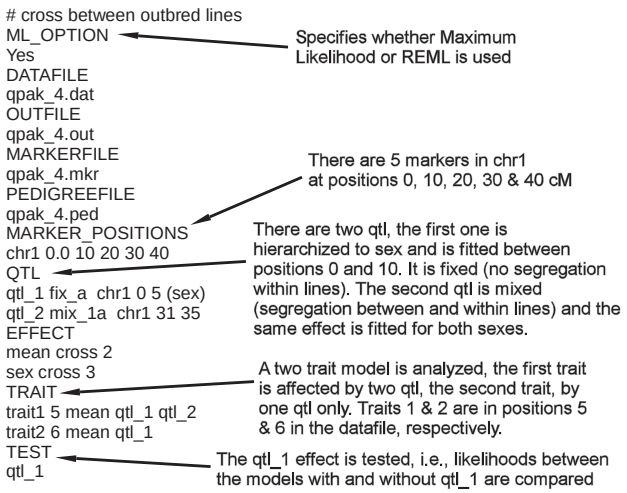

Fig. 1. Example of a parameter file. Section QTL specifies type of QTL (up to 22 modeling options are currently available, check the manual for details). Here it is assumed that alleles are fixed within each parental line (i.e. $\sigma_A^2 = \sigma_B^2 = 0$) for the first QTL, whereas in the second QTL, the dominance effect is set to 0 but σ_A^2 and σ_B^2 are estimated. The effects 'mean' and 'sex' are in columns 2 and 3 of datafile qpak_4.dat, respectively.

coefficient matrix that corresponds to the random effects, inverted (see Appendix).

Programming was in Fortran 95 and we used modules from package BLUPF90 (Misztal *et al.*, 2002), available at http://nce.ads.uga.edu/∼ignacy/newprograms.html. These modules were designed to simplify operations with sparse and dense matrices. Module SPARSEM supported selected operations on sparse matrices, creation of a matrix, extracting blocks, solving, and computing traces and quadratic forms. Module FSPAK90, which is an interface to FSPAK written in Fortran 77 (Pérez-Enciso *et al.*, 1994), supported sparse Cholesky decomposition, sparse finite solving and sparse inversion. In the implementation, the left-hand side of the mixed model equations and each $\mathbf{G}_i^{-1}$ were stored as sparse matrices. Solutions to the mixed model equations were obtained via sparse Cholesky factorization. Calculation of the trace involved the use of a sparse inverse as detailed (Misztal and Perez-Enciso, 1993).

SOFTWARE USAGE

All these theoretical developments are implemented in the Qxpak software, which is freely available to the scientific community at www.icrea.es. A detailed manual, together with a series of examples can be downloaded. Here, we simply outline some of the main usage features. The main options are entered via an ascii parameter file that can contain a series of sections (e.g. Fig. 1). The most relevant sections are the `QTL`, `EFFECT` and `TRAIT` sections. In the `QTL` section, the user must specify the QTL name, the QTL type and the scanning positions on which the QTL should be fitted. The default scanning step is 1 cM, but it can be changed by the user to any desired step width. Each QTL must be specified in a separate line. The `EFFECT` section contains the effect name, its type (covariate or class effect) and the position in the datafile. Additionally, a class effect can be defined as random with a diagonal covariance matrix, an infinitesimal genetic effect, or a user-defined covariance matrix. The `TRAIT` section allows us to model each trait as desired. The user must specify the trait name, the position of the effect values in the data file and the effects (including the QTL) affecting each of the trait. Not all effects or QTL defined need to be actually used in modeling the traits. Each trait is detailed in different lines, and all traits listed are analyzed jointly (a multitrait analysis). Different QTL can be specified for different traits. For instance, these lines in the parameter file:

```
TRAIT
trait_1 10 mean qtl_1
trait_2 11 mean qtl_1
```

result in a model with a single QTL affecting both traits (located in columns 10 and 11 in the datafile). The following lines:

```
TRAIT
trait_1 10 mean qtl_1
trait_2 11 mean qtl_2
```

result in a model with two QTL, each affecting a different trait. Logically, `mean`, `qtl_1` and `qtl_2` must have been defined previously. Effects, QTL or traits can be commented out with the ! or # signs, e.g.

```
TRAIT
trait_1 10 mean qtl_1
!trait_2 11 mean qtl_1
```

is equivalent to

```
TRAIT
trait_1 10 mean qtl_1
```

This makes it easy to run different models, with only minor changes in the parameter file. The program also requires a datafile containing the traits' performances and effects, a pedigree file with the father and mother of each individual, and a marker file where the marker alleles are detailed. Missing genotypes and trait records are coded as 0 (zero). The output file specifies the likelihood at each QTL position and the estimation of effects and their errors of estimation. Furthermore the program computes, for every individual, the probability of a marker interval being of breed origin A or B. This is relevant for fine mapping, as the individuals to be typed are those recombinant with phases known. Thus, if the QTL lies in marker interval 3, say, one should type individuals with

high probability of origin A in interval 1 and high probability of origin B in interval 3, or viceversa. An option allows also to write out the phases sampled in each MCMC iteration, this can be used to determine which is the most likely phase.

Software limitations

The software presented is general but has some limitations. The most important one is related to missing marker information. If a marker is not known, that marker is disregarded for that particular individual and no attempt is made to infer it from typed ancestors or descendants. This can cause biased IBD probabilities if the pedigree contains many untyped individuals. Similarly, if a marker treated as a SNP is missing, that individual is removed from the analysis for that particular SNP (but not for those where the SNP is available). The user is also cautioned against heavy departures from normality. Maximum-likelihood techniques are quite robust but still the validity of P-values depends on normality and are exact only asymptotically.

DISCUSSION

We have presented a versatile tool for the genetic dissection of complex traits that is implemented in the freely available software Qxpak. The package provides ample modeling QTL flexibility without special restrictions on the kind of pedigrees that can be handled. Our approach is able to cope with three broad situations:

(1) *QTL analysis:* Here we include all classical QTL experiments (crosses between inbred lines, within family analyses) as well as complex populations without any predetermined experiment design. The method proposed can deal with any number of traits, or QTL, including sex-linked QTL. It can also accommodate missing markers and missing observations, i.e. for some individuals we may not have recorded all traits. It also allows different models for each trait.

(2) *Association studies:* We are interested in assessing, from among a large panel of SNPs, which have an effect on the trait(s) of interest in order to determine the causal mutation efficiently and with the minimum of manual intervention.

(3) *Genetical genomics:* In this setting, we are doing thousands of QTL analyses, one for each mRNA level, and we want to automate the process as much as possible.

Of course, real situations may involve a mixture of the settings above, e.g. one can do a massive association study in many traits successively (Brem *et al*., 2002). This is also possible with our software.

An important issue is determining the correct significance threshold. This has been, perhaps, one of the most highly debated and controversial issues in QTL studies. Currently, the most popular approach is to use permutation to obtain the genome wide or chromosome wide significance P-values (Churchill and Doerge, 1994). Permutation is a simple and effective strategy, yet it cannot be adapted easily to all situations. The drawback of permutation tests is that the family and data structures are broken. Thus, its implementation in models containing more than one QTL or an infinitesimal genetic effect is not straightforward. Similarly, permutation tests cannot be easily applied in multivariate tests when one is interested in testing the effect of a QTL on a single trait. Other authors (Lander and Kruglyak, 1995) have also proposed rules to compute deterministically the significance thresholds. Again, these rules cannot be applied to all possible situations that can be encountered in QTL analyses. These authors suggested very stringent thresholds, although the research community now advocates for a less conservative approach in order not to miss important sources of genetic variation (Abiola *et al*., 2003). Qxpak provides the likelihood ratio test between two hierarchical models and the nominal P-value using a χ^2 approximation. Exact significance P-values could be obtained with permutation for simple models but this has not been implemented because of computing costs and the limitations of permutation cited above. All in all, and although no method can be claimed to be entirely satisfactory, we recommend the χ^2 approximation for the likelihood ratio test but with a more strict P-value than the nominal significance level along guidelines published (Abiola *et al*., 2003).

With notable exceptions, gene interaction remains largely unexplored despite its interest and relevance (Fijneman *et al*., 1996). Our approach is easily generalized to deal with epistasis. Suppose that interaction between loci k and k' are to be investigated, the terms needed to compute the likelihood can be modeled as

$$E(g_{ik} \times g_{ik'}) = \sum_{W=A,B} \sum_{Z=A,B} \sum_{W'=A,B} \sum_{Z'=A,B} P(g^1_{ik} \in W, g^2_{ik} \in Z, g^1_{ik'} \in W', g^2_{ik'} \in Z')\mu_{WZk,W'Z'k'}$$

that is, we need to consider the alleles received at the two loci, k and k', simultaneously. Similarly,

$$\mathrm{Cov}(g_{ik} \times g_{ik'}, g_{jk} \times g_{jk'}) = \sum_{W=A,B} \sum_{Z=A,B} \left[\frac{1}{2} \sum_{h=1}^{2} \sum_{h'=1}^{2} P(g^h_{ik} \equiv g^{h'}_{ik}, g^h_{ik'} \equiv g^{h'}_{k'} | g^h_{ik} \in W, g^h_{ik'} \in Z)\sigma^2_{WZkk'} \right]$$

The only changes required relate to computing the IBD probabilities and can be easily accommodated with our MCMC strategy, the remaining of the computing strategy remains unchanged. This option is not implemented in the software yet, although it is planned for future releases.

Our method is based on mixed models and maximum-likelihood estimation. Certainly, maximum-likelihood methods are more expensive computationally than least squares (Seaton *et al*., 2002) but our implementation, based on state of the art algorithms, makes it this approach computationally feasible for QTL experiments with up to a few thousands when the QTL is modeled as random, depending on how sparse the relationship matrix is. There is basically no limit if the QTL fitted are fixed effects (i.e. crosses between inbred lines). This should cover most of the current day experiments. As an example, a scan of 100 positions in a 700 individual pedigree takes <10 min on a Pentium IV 2.66 GHz PC running Linux. It should be mentioned that the most expensive part (computing the IBD probabilities and the matrix factorization and ordering) is done only once, so running additional traits increases only marginally the CPU time required.

ACKNOWLEDGEMENTS

We thank Miguel Toro, Hee-Bok Park and Jordi Estellé for testing the programa, and Jesús Piedrafita for comments on the manuscript. M.P.-E. thanks a startup grant from the Catalan Institute for Research and Advanced Studies (ICREA, www.icrea.es) and from Universitat Autònoma de Barcelona and Armand Sánchez. Part of this work was developed using the Catalan Supercomputing Center (CESCA, www.cesca.es) facilities.

REFERENCES

Abiola,O., Angel,J.M., Avner,P., Bachmanov,A.A., Belknap,J.K. Bennett,B., Blankenhorn,E.P., Blizard,D.A., Boliver,V., Brockmann,G.A. *et al*. (2003) The nature and identification of quantitative trait loci: a community's view. *Nat. Rev. Genet.*, **4**, 911–916.

Brem,R.B., Yvert,G., Clinton,R. and Kruglyak,L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

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

Ellegren,H. (2002) Dosage compensation: do birds do it as well? *Trends Genet.*, **18**, 25–28.

Fijneman,R.J., de Vries,S.S., Jansen,R.C. and Demant,P. (1996) Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse. *Nat. Genet.*, **14**, 465–467.

Henderson,C.R. (1984) *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario, CA.

Jansen,R.C. (2003) Studying complex biological systems using multifactorial perturbation. *Nat. Rev. Genet.*, **4**, 145–151.

Jansen,R.C. and Nap,J. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388–391.

Knott,S.A. and Haley,C.S. (2000) Multitrait least squares for quantitative trait loci detection. *Genetics*, **156**, 899–911.

Lander,E. and Kruglyak,L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.*, **11**, 241–247.

Leips,J. and Mackay,T.F. (2002) The complex genetic architecture of *Drosophila* life span. *Exp. Aging Res.*, **28**, 361–390.

Liu,B.H. (1998) *Statistical Genomics*. Boca Raton, Florida.

McCulloch,C.E. and Searle,S.R. (2000) *Generalized, Linear, and Mixed Models*. Wiley, New York.

Misztal,I. and Perez-Enciso,M. (1993) Sparse matrix inversion in restricted maximum likelihood estimation of variance components by expectation–maximization. *J. Dairy Sci.*, **76**, 1479.

Misztal,I., Tsuruta,S., Strabel,T., Auvray,B., Druet,T. and Lee,D.H. (2002) BLUPF90 and related programs (BGF90). Proceedings of 7th World Congress on Genetics Applied to Livestock Production, Monteplier, France. CD-ROM communication, vol. 28, p. 07.

Pérez-Enciso,M., Misztal,I. and Elzo,M.A. (1994) FSPAK-an interface for public domain sparse matrix subroutines. Proceedings of 5th World Congress on Genetics Applied to Livestock Production, vol. 22, pp. 87–88.

Pérez-Enciso,M., Clop,A., Folch,J.M., Sanchez,A., Oliver,M.A., Ovilo,C., Baragan,C., Varona,L. and Noguera,J.L. (2002) Exploring alternative models for sex-linked quantitative trait loci in outbred populations. Application to an Iberian × Landrace pig intercross. *Genetics*, **161**, 1625–1632.

Pérez-Enciso,M. and Varona,L. (2000) Quantitative trait loci mapping in F2 crosses between outbred lines. *Genetics*, **155**, 391–405.

Pérez-Enciso,M., Varona,L., and Rothschild,M. (2000) Computation of identity by descent probabilities conditional on DNA markers via a Monte Carlo Markov Chain method. *Genet. Sel. Evol.*, **32**, 467–482.

Schadt,E.E., Monks,S.A., Drake,T.A., Lusis,A.J., Che,N., Colinayo,V., Ruff,T.G., Milligan,S.B., Lamb,J.R., Cavet,G. *et al*. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.

Searle,S.R. (1982) *Matrix Algebra Useful for Statistics*. Wiley, New York.

Seaton,G., Haley,C.S., Knott,S.A., Kearsey,M. and Visscher,P.M. (2002) QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics*, **18**, 339–340.

APPENDIX

The distribution of the random variables in model (1) is

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{g}\\ \mathbf{e}\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\mathbf{Xb}+\mathbf{P}\boldsymbol{\mu}\\ \mathbf{P}\boldsymbol{\mu}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{V} & \mathbf{G} & \mathbf{R}\\ \mathbf{G} & \mathbf{G} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{pmatrix}\right],$$

where $\mathbf{P}$ is a matrix containing the $P(g^1_{ik} \in U, g^2_{ik} \in W)$ elements from Equation (2), μ is a vector with the μ_{Wk} elements for every QTL, $\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$. The matrix $\mathbf{R}$ contains the variances and covariances of the residuals. This matrix is diagonal in univariate models and block-diagonal for multitrait analyses. In the case of missing traits, these blocks are different for each individual according to which trait(s) is(are) missing. Matrix $\mathbf{G}$ is equal to $\sum_{k=0}^{N_q} \mathbf{G}_k$, each matrix $\mathbf{G}_k$ is a matrix made up of the terms described in Equation (3). Here, it is assumed that the total genetic variance is the sum of the variances at each loci. This is an approximation, when there is linkage disequilibrium in the outbred populations, but is accurate as long as markers are not very sparsely located

(Pérez-Enciso and Varona, 2000). The corresponding mixed model equations are:

$$\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}}_0 \\ \dots \\ \hat{\mathbf{g}}_{N_q} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \dots & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}_0^{-1} & \dots & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \\ \dots & \dots & \dots & \dots \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \dots & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}_{N_q}^{-1} \end{pmatrix}^{-1} \times \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \dots \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

(Henderson, 1984; McCulloch and Searle, 2000), where $\hat{\mathbf{b}}$ contains the best linear unbiased estimates of the fixed effects, whereas $\hat{\mathbf{g}}_k$ contains the best linear unbiased predictors of the k-th QTL genetic effects. These quantities are required to compute the QTL and residual variances, Equations (4) and (5). The submatrix $\mathbf{C}^{\mathrm{F}}$ required to compute the likelihood is:

$$\mathbf{C}^{\mathrm{R}} = \begin{pmatrix} \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}_{\mathbf{0}}^{-1} & \dots & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \\ \dots & \dots & \dots \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \dots & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}_{N_q}^{-1} \end{pmatrix}^{-1}.$$