Journal of Statistical Software

August 2012, Volume 50, Issue 6. *http://www.jstatsoft.org/*

dlmap: An R Package for Mixed Model QTL and Association Analysis

B. Emma Huang
CSIRO

Rohan Shah
CSIRO

Andrew W. George
CSIRO

Abstract

dlmap is a software package capable of mapping quantitative trait loci (QTL) in a variety of genetic studies. Unlike most other QTL mapping packages, **dlmap** is built on a linear mixed model platform, and thus can simultaneously handle multiple sources of genetic and environmental variation. Furthermore, it can accommodate both experimental crosses and association mapping populations within a versatile modeling framework. The software implements a mapping algorithm with separate detection and localization stages in a user-friendly manner. It accepts data in various common formats, has a flexible modeling environment, and summarizes results both graphically and numerically.

Keywords: interval mapping, linear mixed models, quantitative trait loci, R, association mapping, F2.

1. Introduction

Genetic studies to identify the underlying causes of complex traits such as disease resistance and yield have been prevalent for decades. Such studies range from linkage mapping approaches primarily focused on the inheritance of traits in families to association mapping approaches primarily focused on inheritance throughout a broader population. Irrespective of the study design, however, the ultimate goal is to detect quantitative trait loci (QTL), which are regions of the genome associated with the trait under consideration.

Mapping QTL in experimental cross families has been and is still a major focus of genetic research. Initial methods tested for association at individual markers (Soller and Brody 1976), but shifted to forms of interval mapping (Lander and Botstein 1989) to make use of map information and flanking marker genotypes to gain greater power in QTL detection. These research efforts have resulted in a large number of QTL mapping packages, including such commonly used software as **MapQTL** (Van Ooijen and Maliepaard 1996), R/**qtl**

(Broman, Wu, Sen, and Churchill 2003) and **QTL Cartographer** (Wang, Basten, and Zeng 2005). Perhaps the most common form of QTL mapping now is composite interval mapping (Zeng 1993; Jansen 1993) and its many variants (e.g., Kao, Zeng, and Teasdale 1999; Li, Ribaut, Li, and Wang 2008), where additional marker covariates are incorporated into the model to account for background genomic variation. However, a major limitation of many methods and software is their inability to handle random effects.

Random sources of variation often arise in plant and animal studies; failure to account for these correctly can bias linkage findings. Thus researchers have begun to turn to mixed model approaches to account for all sources of variation. This is made evident by the increasing availability of mixed model interval mapping approaches in packages such as **GenStat** (VSN International 2011) and **QTLMapper** (Wang, Zhu, Li, and Paterson 1999). This allows for more flexible modeling of gene-trait association, including modeling QTL as random rather than fixed effects (e.g., Lee and Van der Werf 2006; Verbyla, Cullis, and Thompson 2007).

Recently, detection localization (DL) mapping (Huang and George 2009) was proposed as a new approach for mapping QTL. It takes a whole-genome approach to the detection of QTL which involves modeling the influence of all markers simultaneously, rather than separately as in interval mapping techniques. The simulation results in Huang and George (2009) show that the DL mapping approach is a powerful tool for detecting QTL, and can outperform traditional methods such as composite interval mapping (CIM). The DL mapping method also requires no assumptions about the number of QTL present, while the requirement of specifying the number of cofactors in CIM can affect linkage findings.

DL mapping is built on a linear mixed modeling framework, which allows easy and efficient incorporation of different sources of variation into a unified QTL mapping process. It was initially presented as a QTL mapping method for the analysis of data collected from F2, backcross, recombinant inbred line and doubled haploid experimental crosses. However, it stands out from other mixed model QTL mapping methods (Verbyla *et al.* 2007; Gilmour 2007) in that the mapping algorithm is easily adjusted to also map QTL in association studies. While there are packages available for mixed model association mapping, such as **TASSEL** (Bradbury, Zhang, Kroon, Casstevens, Ramdoss, and Buckler 2007), these in turn are not designed for experimental crosses.

This paper describes the R (R Development Core Team 2012) package **dlmap**, which implements the DL mapping approach. The package can be downloaded from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=dlmap`. The strengths of the **dlmap** package are (1) compatibility with data from other commonly used mapping software and (2) the ability to map QTL in a variety of experimental populations and study designs, both built upon a flexible modeling environment.

The remainder of this paper is organized into the following parts. Section 2 briefly outlines the theory behind the DL mapping method. Section 3 gives the typical workflow involved in using **dlmap**, and lists some of the useful functions included in the package. Section 4 describes the computational structure of the package and implementation issues. Section 5 gives three worked examples: the first is an introductory simulated example to illustrate the basic use of the package; the second is a more complex example exploring some of the issues underlying analysis of a real experimental cross; and the third considers real genetic data from an association mapping population. Analysis and visualization of results are demonstrated in each case.

2. Theory

The theory behind DL mapping has been published previously (Huang and George 2009); hence we will give an overview of the process rather than full detail of the statistical theory here.

The DL mapping method involves two stages. The first stage is detection, where the number of QTL and their approximate positions are determined. This is done first by selecting whole chromosomes that appear to contain QTL, and then by selecting a single marker on each of these chromosomes which appears to be linked to a QTL. The process then repeats, to allow for multiple QTL on a single chromosome.

The second stage is localization, which estimates the positions of the QTL according to the number of QTL per chromosome found in the detection stage. After the locations have all been determined the effects are determined jointly.

In both stages, we assume a population consisting of n individuals genotyped on a total of M genetic markers spanning C chromosomes. We are attempting to find genomic regions which impact a vector $\mathbf{y}$ of phenotypic trait observations. Note that the length of $\mathbf{y}$ may be larger than n if repeated measurements are taken on individuals. We assume the genetic markers $\mathbf{m}_1$, $\mathbf{m}_2$, …, $\mathbf{m}_j$, …, $\mathbf{m}_C$, are the sets of M_j markers known to reside on each of the C chromosomes. The marker data $\mathbf{Z} = [\mathbf{Z}_1 \ldots \mathbf{Z}_C]$ are partitioned as data collected on C chromosomes where $\mathbf{Z}_j$ is a $(n \times M_j)$ matrix of observed genotypes on $\mathbf{m}_j$. Furthermore, let $\mathbf{Z}_{-j} = [\mathbf{Z}_1 \ldots \mathbf{Z}_{j-1}\mathbf{Z}_{j+1} \ldots \mathbf{Z}_C]$ denote marker data collected on all C chromosomes except the j-th chromosome.

Throughout the models described below, we add terms $\mathbf{X}_e\beta_e$ and $\mathbf{Z}_e\mathbf{u}_e$ to the notation previously defined in Huang and George (2009) to represent the non-genetic fixed and random effects. The non-genetic random effects $\mathbf{u}_e$ are assumed to be uncorrelated with other random effects and distributed as a multivariate normal density with $\mathbf{u}_e \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{G})$.

2.1. Detection

In each iteration i of the detection stage, we start with a collection $\mathcal{R}^{(i-1)}$ of markers that have previously been identified as being linked to QTL on a subset $\mathcal{G}^{(i-1)}$ of chromosomes. Let r indicate the size of the collection $\mathcal{R}^i$. At the start of the detection stage $\mathcal{R}^0 = \{\}$ and $\mathcal{G}^0 = \{1, 2 \ldots C\}$ so that initially all chromosomes are considered to potentially contain QTL, and no QTL have yet been identified.

First, for each chromosome in $\mathcal{G}^{(i-1)}$ we compare a reduced mixed model to a full mixed model. The full mixed model is given by:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}^{i-1}\mathbf{u}^{i-1} + \mathbf{X}_e\beta_e + \mathbf{Z}_e\mathbf{u}_e + \mathbf{e} \tag{1}$$

while the reduced model for the j-th chromosome is given by:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_{-j}^{i-1}\mathbf{u}_{-j}^{i-1} + \mathbf{X}_e\beta_e + \mathbf{Z}_e\mathbf{u}_e + \mathbf{e} \tag{2}$$

In both models, $\mathbf{X}$ is an $(n \times r)$ design matrix whose columns contain the marker genotypes for the selected markers, β is an $(r \times 1)$ vector of additive QTL effects for QTL linked to the markers, $\mathbf{Z}^{i-1} = [\mathbf{Z}_1^{i-1}\mathbf{Z}_2^{i-1} \ldots \mathbf{Z}_{\mathcal{G}^{i-1}}^{i-1}]$ denotes the marker data on the chromosomes under investigation excluding data on any of the r selected markers, $\mathbf{u}^{i-1}$ is a vector of random QTL effects associated with the markers on the C^{i-1} chromosomes under investigation excluding any of the r selected markers, and $\mathbf{e}$ is a residual vector. The random effects $\mathbf{u}^{i-1}$ and $\mathbf{e}$ are assumed to be uncorrelated and distributed as multivariate normal densities with $\mathbf{e} \sim \mathbf{N}_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ and $\mathbf{u}^{i-1} \sim \mathbf{N}_{C^{i-1}}(\mathbf{0}, \sigma^2 \mathbf{S})$, where $\mathbf{S}$ is a block diagonal covariance matrix with $\mathbf{S}_{jj} = \gamma_j \mathbf{I}_{M_j^{i-1}}$ $(j = 1, \ldots, C^{i-1})$.

In other words, we are comparing the fit of a model containing a separate variance component for each chromosome under investigation to one which leaves out the j-th chromosome. The random effects for markers on each chromosome model the chromosome's genetic contribution, so by comparing the full and reduced models we can determine whether the j-th chromosome has a significant genetic contribution. The r markers are included in the models as fixed effects in order to reduce the potential for the detected QTL to obfuscate the discovery of unknown QTL.

We test for the presence of QTL on the j-th chromosome by calculating the residual log likelihood ratio (lr_j). That is, for the j-th indexed chromosome, we calculate $lr_j = -2(\log\mathrm{L}_{full} - \log\mathrm{L}_j)$ where $\log\mathrm{L}_{full}$ and $\log\mathrm{L}_j$ are the residual log likelihoods under the full and j-th reduced models, respectively. We can calculate an appropriate significance threshold for the statistic by considering its distribution under the null hypothesis that the two models are equivalent. For a single chromosome, the distribution is known to follow a mixture of chisquared distributions (Self and Liang 1987). However, for multiple chromosomes, the distribution is unknown and we must either use a correction for multiple testing such as Bonferroni or Holm, or assess the significance empirically with permutation (Churchill and Doerge 1994). Note that as the individual tests are for independent chromosomes, the Bonferroni correction is not overly conservative. Chromosomes for which the likelihood ratio test is not significant are dropped from $\mathcal{G}^i$, and will not be considered again in subsequent iterations.

For each chromosome found in the i-th iteration to contain a QTL, we identify the marker on this chromosome that explains the most variation. Suppose the j-th indexed chromosome has been found to contain undetected QTL. For each marker on this chromosome we construct a linear mixed model where the marker is treated as a fixed effect, together with the r markers already identified as being associated with QTL. The remaining markers on the other $C^{i-1} - 1$ chromosomes are treated as random effects where as before, markers on the same chromosome have common variance. Markers on the j-th indexed chromosomes are excluded from the random effects in the model.

The linear mixed model for the x-th marker on the j-th indexed chromosome is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{z}_{j;x}\beta_x + \mathbf{Z}_{-j}\mathbf{u}_{-j} + \mathbf{X}_e\beta_e + \mathbf{Z}_e\mathbf{u}_e + \mathbf{e} \qquad (3)$$

where $\mathbf{z}_{j;x}$ is a $(n \times 1)$ vector of marker data collected on the x-th marker on the j-th indexed chromosome, and β_x represents an additive fixed effect for a QTL at the marker. All other terms are as described previously. We calculate the Wald statistic of β_x and select the marker associated with the model giving the largest Wald statistic. This fixed effect is then incorporated into the full model in subsequent iterations. Note that we are not using the Wald statistic to assess the significance of the fixed effect; it is only used to identify the "best" marker.

As the detection procedure is iterated, chromosomes with non-significant likelihood ratio tests are removed from $\mathcal{G}^i$ and the selected markers are added to $\mathcal{R}^i$ and therefore incorporated as fixed effects to the full and reduced models. The detection stage terminates when no chromosomes have significant likelihood ratio test statistics, thus indicating that all genetic variation has been explained by the fixed effects already in the model.

2.2. Localization

In the localization stage of our QTL mapping strategy, we use an iterative procedure to position multiple QTL on a chromosome. Each iteration involves a linear mixed model approach analogous to regression-based interval mapping. Once a QTL has been positioned, it is included in subsequent scans of that chromosome as a fixed effect. We also include markers on other chromosomes as random marker effects (as we did in our detection procedure). After all QTL have been positioned along a genome, we construct a final multiple regression model to estimate the sizes of the QTL.

Assume C^* chromosomes have been found to contain QTL in our detection stage where $C^* \leq C$. We arbitrarily index these chromosomes $1, 2, \ldots, C^*$. QTL on each chromosome are positioned independently of other chromosomes. Consider the i-th iteration for the j-th chromosome: $i-1$ interval mapping scans have already been performed, and $i-1$ QTL have been positioned.

Then essentially we fit separate models at every hypothesized QTL position along the chromosome in order to determine the best position for the new QTL. Let θ_Q indicate the hypothetical QTL position. At each point we fit the linear mixed model

$$\mathbf{y} = \mathbf{X}\beta + E(\mathbf{g}|\theta_Q, \mathbf{z}_L, \mathbf{z}_R)\beta_{\theta_Q} + \mathbf{Z}^*_{-j}\mathbf{u}^*_{-j} + \mathbf{X}_e\beta_e + \mathbf{Z}_e\mathbf{u}_e + \mathbf{e} \tag{4}$$

This model is similar in structure to Equation 3 that we used in our detection procedure to select markers. $\mathbf{X}$ is an $(n \times (i-1))$ matrix of expected QTL genotypes for the $(i-1)$ positioned QTL and β_{θ_Q} is the additive effect of a putative QTL at position θ_Q. $\mathbf{u}^*_{-j}$ denotes the vector of random effects associated with markers on chromosomes where QTL have been detected, excluding the j-th of these. Correspondingly, $\mathbf{Z}^*_{-j}$ is the matrix of marker genotypes on all C^* chromosomes except for the j-th indexed chromosome and any of the markers flanking the $i-1$ previously positioned QTL. $E(\mathbf{g}|\theta_Q, \mathbf{z}_L, \mathbf{z}_R)$ is the expected QTL genotype conditional on its position θ_Q and the genotypes of its left $(\mathbf{z}_L)$ and right $(\mathbf{z}_R)$ flanking markers. See Whittaker, Curnow, Haley, and Thompson (1995) for further details. We assume $\mathbf{e}$ and $\mathbf{u}^*_{-j}$ are uncorrelated and distributed as multivariate normal densities as before.

For each such model, we calculate the Wald statistic for β_{θ_Q}. The hypothesized position yielding the mixed model with the highest Wald statistic is the estimated location of the QTL. We repeat this process until all QTL have been positioned along the genome.

Once the C^* chromosomes have been iteratively scanned and the QTL positioned, we construct a final model to estimate the additive effects of the QTL. To estimate the additive effects of the located QTL we fit

$$\mathbf{y} = \mu + \sum_{j=1}^{C^*} \sum_{m=1}^{t_j} E(\mathbf{g}|\theta_{j;m}, \mathbf{z}_{j;m-1}, \mathbf{z}_{j;m+1})\beta_{j;m} + \mathbf{X}_e\beta_e + \mathbf{Z}_e\mathbf{u}_e + \mathbf{e} \tag{5}$$

where μ is the intercept, and for the j-th indexed chromosome, $\theta_{j;m}$ is the position of the m-th QTL, t_j is the number of detected QTL, and $\mathbf{z}_{j;m-1}$ and $\mathbf{z}_{j;m+1}$ are the $(n \times 1)$ vectors of genotypes collected on the markers flanking the m-th QTL. This final model jointly estimates the effects of all QTL on all chromosomes.

3. Procedure

A typical analysis using the **dlmap** package is performed in the following stages.

1. For experimental crosses, read genetic data into R as a `cross` object using `read.cross` from the **qtl** package. This function supports a wide variety of file formats; further information can be found in the documentation for R/**qtl**.

   If the data involves repeated measurements from the same genetic line there will be more phenotypic information than genetic information. In this case a separate matrix of phenotype data will need to be read into R. In order to link the genetic marker data to the phenotypic data, a column of identifiers must be included in both files.

   An alternative to using **qtl** to input data is to construct three files (or objects) encoding the genetic marker data, phenotypic trait data, and chromosome assignments for each marker. These are required for an association mapping population, but can be used for experimental crosses as well.

2. Create a `dlcross` object which combines the genetic and phenotypic data into a single object using

   ```
   dlcross(format = c("rqtl","dlmap","other"), genobj, pheobj, mapobj,
     idname, genfile, phefile, mapfile, type, step, fixpos, estmap)
   ```

   The ID name (`idname`) has a default value of `"ID"`; the step size or number of positions within intervals (`step` and `fixpos`) have default values of `0`; and whether to estimate the map or not (`estmap`) has a default value of `TRUE`.

   Input can either be input directly from files (`genfile`, `phefile`, `mapfile`) as mentioned in the previous step, or can be read into R and input as objects (`genobj`, `pheobj`, `mapobj`) to the function. This function creates the input to the main analysis function by merging the genetic and phenotypic values into a single data frame. For an experimental cross with `genobj` in cross format, if there are no repeated measurements the phenotypic values may be included in the `cross` object, and the `pheobj` argument may be omitted.

   In addition, the function computes expected genetic values for a grid of points along the genome. The argument `fixpos = f` specifies that locations for QTL will be considered at markers and `f` positions equally spaced between every adjacent pair of markers. The argument `step = s` specifies that locations for QTL will be considered on an equally spaced grid of positions `s` cM apart. Only one of `step` and `fixpos` should be specified, since they will almost always specify a different set of QTL locations. Neither of these will be used for an association mapping population.

3. Specify a base model accounting for all effects due to sources other than genetic markers (e.g., phenotypic or polygenic). These correspond to the term $\mathbf{u}_e$ in the previous models, and typically relate to environmental variation or the design of the underlying experiment.

4. Use the `dlmap` function to perform the QTL analysis.

   ```
   dlmap(object, phename, baseModel, algorithm = c("asreml", "lme"),
     pedigree, seed, n.perm, multtest = c("holm", "bon"), alpha, filestem,
     ...)
   ```

The random seed (seed) has default value of 1; the number of permutations (n.perm) has default value of 0; the significance level (alpha) has default value of 0.05; the maximum number of iterations in the fitting process (maxit) has default value of 60; and the file stem for output files (filestem) has default value "dl".

The argument object is the output from the dlcross function. The required argument baseModel gives the model for the non-genetic terms and must have class asreml. The argument phename gives the name of the trait of interest for QTL mapping. The argument pedigree allows for the possibility of including a polygenic random effect to account for family structure, either by inputting a pedigree matrix or a kinship matrix; seed and n.perm relate to parameters for permutation testing; multtest performs an adjustment for multiple testing if no permutations are run (with a Bonferroni or Holm correction); alpha sets the significance level for testing; and filestem pertains to the names of output files. The argument maxit controls the maximum number of iterations of the underlying algorithm.

4. Computational details

The **dlmap** package has two primary functions which create novel data structures in order to accommodate genetic and phenotypic data in an integrated manner. As analysis proceeds through the different stages of data input, detection and localization, the structure of the data object alters in order to retain different pieces of information. Table 1 summarizes the analysis and visualization functions for these two types of objects. Below we discuss further some computational details and give recommendations for their usage.

4.1. dlcross

Data is initially converted to dlcross format either from cross format (as described in the **qtl** package), or from external files or objects. This class of object takes the form of a list with several components necessary for later analysis. In particular, the dfMerged and dfMrk components contain merged genetic and phenotypic data; map and mapp contain the original genetic map and a map of all positions to be scanned in localization; genCross contains the original (if input) cross object; nphe is the number of phenotypes; idname is the ID variable common to both genetic and phenotypic data; and loc is a flag indicating whether localization should be performed in ensuing analysis. For the most part these components will only need to be accessed internally during the detection and localization stages.

To explore objects in dlcross format we provide basic summary and plot functions. For experimental crosses, plotting this type of object will display the genetic map and a histogram or barplot of the first few phenotypic variables, depending on whether they are continuous or categorical. For association mapping populations only the phenotypic variables will be plotted. The summary of the object will include information on the number of genotyped samples, the number of phenotypic measurements and variables, and when a genetic map is included, the number of chromosomes and distribution of genetic markers among chromosomes.

4.2. dlmap

After the primary analysis function dlmap has been run, the input dlcross object is stored in

Function/method	Description
`dlcross`	Create input data objects for `dlmap` function. Genetic and phenotypic data are merged and genetic expected values are computed at potential QTL locations.
`dlmap`	Perform QTL detection and localization. Produce an object of class `dlmap` containing estimates of QTL locations and effects and the final fitted model.
`summary`, `print`	Print a summary of the analysis results.
`plot`	Visually display the linkage map and the results of the analysis. QTL locations and flanking markers are highlighted.
`profileplot`	Display the Wald statistic profile for the chromosomes detected to have QTL, from the first iteration of the detection step.

Table 1: A list of functions/methods available in **dlmap** with short descriptions.

the output object, which now additionally inherits class `dlmap`. In addition to the input object, this new list-based structure contains components indicating the number of QTL detected, the final model fit containing all detected QTL, the Wald profile for each chromosome where a QTL is detected, and a summary of the QTL. The latter two components are primarily for use in `summary` and `plot` functions to operate on objects of type `dlmap`. The `plot` function will display the linkage map with QTL positions marked as well as intervals between flanking markers for the QTL. Additionally, the `profileplot` function will display the Wald profile for each chromosome where QTL are detected, which should peak at the position estimate for the QTL. Finally, the `summary` function will print a table containing the chromosome, position in cM, flanking markers, effect and standard deviation estimates, and Z value and p value for all QTL detected. Effects and p values are estimated from a model including all QTL simultaneously.

There are two versions of the `dlmap` function, which use different engines to fit the linear mixed models which form the framework of the algorithm. A specific version is selected based on the value of the argument `algorithm`. For example, `algorithm = "asreml"` uses the R implementation of the software **ASReml** (Gilmour, Gogel, Cullis, and Thompson 2009). This provides a much more general and powerful implementation of the DLMapping algorithm and is the preferred method of analysis. The other option, `algorithm = "lme"`, uses R/**nlme** (Pinheiro, Bates, DebRoy, Sarkar, and R Development Core Team 2011), and is more restricted in its capabilities. It cannot be used to model random effects or covariance structure, cannot handle more than 200 markers, and only allows for a single phenotypic observation per genotype. Also, permutation has not been implemented for this function because it is very slow. However, this version will fit the basic algorithm and is useful should a license for **ASReml** not be available.

In studies where the number of genetic markers is much larger than the number of phenotyped individuals, we transform the genetic data to improve computational efficiency. By doing so we are able to reduce the dimension of the analysis to the number of genetic lines used in the analysis multiplied by the number of chromosomes in the genetic map. This is done in a manner similar to **wgaim**, with thanks to Julian Taylor and Ari Verbyla for the suggestion (Taylor and Verbyla 2011). This transformation is particularly useful for high-dimensional data such as we see in association analysis.

Missing values in `asreml` are replaced with zeros, so it is important to centre any covariates with missing data. This is done internally for all genotypes when `algorithm = "asreml"`. Thus individuals with phenotypic but not genotypic data, which play important roles in field trials, may be included safely. When `algorithm = "lme"` these individuals cannot be included, so the default behavior is to omit observations with missing values.

It is recommended that `n.perm` be set to 0 for initial exploratory analysis, as the permutation analysis may be lengthy. Additionally we recommend that permutation not be used in association analysis, as this is inappropriate when substructure exists (which is often the case). By default, the Holm correction is used to adjust for the number of chromosomes under consideration at each detection stage. While this is a conservative measure it seems to perform well in practice.

Two files are output with names set by the argument `filestem`, which has a default value of `"dl"`. The file "`filestem.trace`" contains **ASReml** licensing and likelihood convergence output which otherwise would be dumped to the screen and possibly obscure other messages. Errors, warnings and other messages will still appear on the screen. Some warnings which appear may be passed through from an ASReml call and output on exit. These may generally be ignored. This file is not created if `algorithm = "lme"` is used.

The file "`filestem.det.log`" is a record of iterations in the detection stage. For each iteration the restricted maximum likelihood ratio test testing for genetic variation on each chromosome is output, along with adjusted p values, genomewide threshold and markers selected as fixed effects. The p values are corrected for the number of chromosomes tested either by the Holm correction or by permutation. If the number of permutations (`n.perm`) is greater than 0, then for the Xth iteration an additional file "`filestem.permX`" will be created which contains the test statistics for the permuted datasets. Examples of the two files will be given in the following section.

5. Worked examples

We present three examples of how to use **dlmap** here. First, we provide an introductory example based on simulated backcross data. This is primarily an opportunity to demonstrate the features of the package during the analysis procedure, the documentation accompanying analysis, and the visualization and summary of results. Second, we analyze the Sunco x Tasman wheat doubled haploid population and compare the results to those previously published. This more complex analysis demonstrates the utility of **dlmap** in a real-world QTL mapping study. Third and finally, we analyze an association mapping population from *Arabidopsis thaliana*. This demonstrates the usage of **dlmap** for association analysis, including its ability to accommodate pedigree and population structure effects.

5.1. Simulated backcross population

A simulated data set is included with **dlmap** for a backcross population with 250 lines in a randomized block design. Each line is replicated across four blocks. This example is constructed to illustrate the features of **dlmap** under a relatively simple scenario. In practice such a design should be avoided with a large number of lines, since large heterogeneity may exist within complete blocks. This may lead to a loss of efficiency, even if some is recoved via spatial modeling during analysis.

Chromosome	Location	Effect
1	30 cM	0.76
1	70 cM	0.76
2	30 cM	0.76
2	70 cM	−0.76
3	50 cM	0.76
4	30 cM	0.76
5	0 cM	0.76

Table 2: QTL locations for simulated backcross data.

The genetic map is generated with nine chromosomes of length 100 cM, each containing 11 equally spaced markers. This set-up mirrors Broman and Speed (2002), as do the simulated QTL locations and effects (see Table 2). Because each genetic line is replicated the data is split into two parts, the phenotypic data and the genotype data. The first step is to read in the phenotypic data matrix. Here the identifier linking the genetic and phenotypic data is named `ID` and has values `S1`, . . . , `S250`. The associated genetic data is named `BSdat`.

```
R> summary(BSdat)
```

```
    Backcross

    No. individuals:    250

    No. phenotypes:     2
    Percent phenotyped: 100 100

    No. chromosomes:    9
        Autosomes:      1 2 3 4 5 6 7 8 9

    Total markers:      99
    No. markers:        11 11 11 11 11 11 11 11 11
    Percent genotyped:  100
    Genotypes (%):      AA:48.6  AB:51.4
```

An appropriate base model can be fit with the `asreml` command. In this case we specify a non-zero intercept, and `Block` and `ID` as random effects. Using `ID` as a random effect allows for a per-line polygenic effect unrelated to the genotype data. The `varcomp` component of `summary.asreml` summarizes the model random effects. The `gamma` component refers to variance ratios, and in this case the polygenic random effect has around the same variance as the residual term, but the `Block` effect has a variance more than three times as large. Other components describe the fixed effect estimates for the model.

```
R> basemodel1 <- asreml(phenotype ~ 1, random = ~ Block + ID,
+    data = BSphe, na.method.X = "include")
R> summary(basemodel1)$varcomp
```

```
                   gamma component  std.error   z.ratio constraint
Block!Block.var 3.488196 3.3014714 2.69873125  1.223342   Positive
ID!ID.var       1.007629 0.9536899 0.10737817  8.881599   Positive
R!variance      1.000000 0.9464694 0.04897352 19.326148   Positive
```

In this example the genetic data BSdat already has class cross so we can apply dlcross directly to it and BSphe. The ID phenotype in this object is used to link the phenotypic and genetic data.

```
R> dlin1 <- dlcross(format = "rqtl", genobj = BSdat, pheobj = BSphe,
+    idname = "ID", fixpos = 1)
```

By default, we re-estimate the genetic map in this function. By specifying fixpos = 1, we search over all markers and midpoints of marker intervals for QTL positions. The analysis can be performed with

```
R> dlout1 <- dlmap(dlin1, phename = "phenotype", baseModel = basemodel1,
+    algorithm = "asreml")
```

To view the positions of the QTL detected, we use the command plot(dlout1) to plot the linkage map. The result is shown in Figure 1. Purple squares indicate the positions estimated for QTL during the localization stage, while flanking markers are indicated in red. Larger rectangles may also be drawn to denote the regions between flanking markers; however these should not be mistaken for QTL support intervals. Finally, we can access the estimated sizes and locations of the QTL by printing a full summary of the results.

```
R> summary(dlout1)
```

```
 Summary of final results:
  Chr   Pos Left Marker Right Marker Effect    SD Z-value p-value
1   1 33.05        D1M3         D1M5  0.638 0.078    8.18       0
2   1 71.27        D1M7         D1M8   0.81 0.081      10       0
3   2 32.04        D2M3         D2M5  0.831 0.073   11.38       0
4   2 70.87        D2M7         D2M9 -0.894 0.073  -12.25       0
5   3 51.79        D3M5         D3M7  0.777 0.063   12.33       0
6   4 25.29        D4M3         D4M5  0.703 0.063   11.16       0
7   5     0        D5M1         D5M2  0.696 0.063   11.05       0
```

The result of the analysis is that two QTL have been detected on chromosome 1, two on chromosome 2, and one each on chromosomes 3, 4 and 5. The position estimates are very close to the true values, and all the QTL size estimates have the correct signs and values at most 0.13 in magnitude from the truth.

More information about how this result was determined is shown below. During analysis, this information is stored in the file <filestem>.det.log. For each iteration of testing and scanning in the detection stage, observed test statistics and adjusted p values are output to the file along with the genomewide threshold for the likelihood test. Then, for each chromosome found to contain significant genetic variation, one marker is selected for inclusion as a fixed effect in future iterations and output to the file. Additional files containing permutation test statistics are created if the number of permutations is greater than zero.

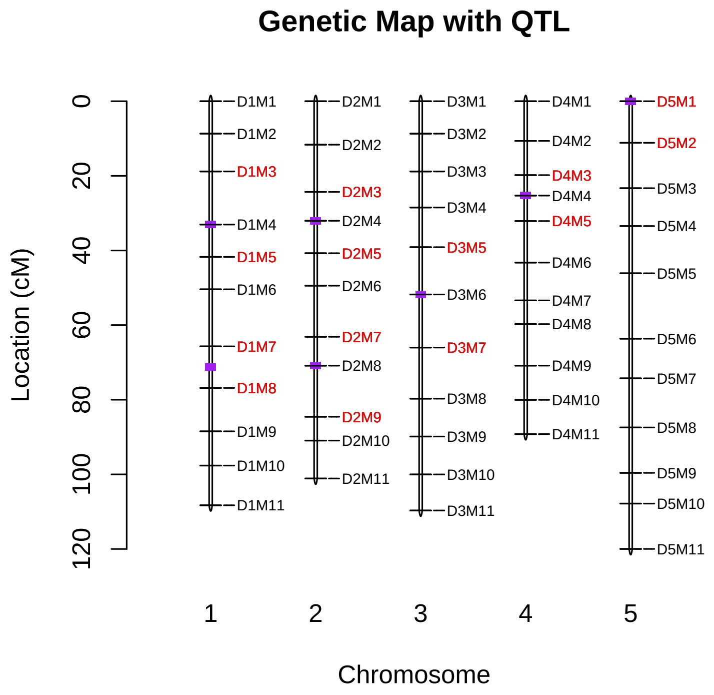

Figure 1: Genetic map with estimated QTL locations shown.

```
***************************************************************
Iteration 1: No. Permutations=0
             1        2        3       4       5      6      7      8      9
Obs   276.3626 123.9277 108.5962 77.6036 87.3012 2.9690 -2e-04 0.0596 -3e-04
P-val   0.0000   0.0000   0.0000  0.0000  0.0000 0.1697  1e+00 1.0000  1e+00
5% Genomewide Threshold: 6.4475
Significant chromosomes for next round of testing/scanning:

Chromosome 1 2 3 4 5
Marker     8 8 6 4 1

***************************************************************
Iteration 2: No. Permutations=0
Chromosomes from previous iteration:
            1        2      3      4 5
Obs   68.7955 101.4466 0.1374 -1e-04 0
P-val  0.0000   0.0000 1.0000  1e+00 1
5% Genomewide Threshold: 5.4119
Significant chromosomes for next round of testing/scanning:

Chromosome 1 2
Marker     4 4
```

```
****************************************************************
Iteration 3: No. Permutations=0
Chromosomes from previous iteration:
      1      2
Obs   0 0.0027
P-val 1 0.9582
5% Genomewide Threshold: 3.8415
```

5.2. Sunco x Tasman doubled haploid wheat population

In the previous example we analyzed a simple simulated dataset in order to demonstrate the typical analysis commands and resulting output from **dlmap**. However, the simulated data does not encompass many of the complexities of real data, which may require modeling more complicated design and environmental effects.In this example we apply **dlmap** to the Sunco x Tasman doubled haploid bread wheat population previously analyzed in Verbyla *et al.* (2007).

The experimental population and phenotypic data collection have been fully described in Verbyla *et al.* (2007), so we will give only a brief overview here. The example is taken from a field experiment conducted in 2000 involving doubled haploid (DH) lines from the Sunco x Tasman mapping population. A randomized complete block design was used for the experiment, with two replicates of each DH line as well as additional plots for parental and commercial lines. In total 183 lines unique genotypes and 456 phenotypes were collected.

Grain samples were collected from most of the plots for milling, with 23% of the samples replicated in the milling process. Samples were randomly assigned to mill days and mill order within mill days. The trait "milling yield" is the focus of this example, and the QTL analysis of this trait hence takes into account both stages (field and milling) of the experimental design. Further details of the base models used for phenotypic analysis of the population can be found in Smith, Lim, and Cullis (2006).

The linkage map used in QTL analysis was originally presented in Lehmensiek, Eckermann, Verbyla, Appels, Sutherland, and Daggard (2005), and contains a mixture of 287 AFLP (amplified fragment length polymorphism), RFLP (restriction fragment length polymorphism) and microsatellite markers after removing those which were co-located with other markers.

To analyze this population, we first construct a `dlcross` object as before from the phenotypic and genetic information. We can see more details of the data from the summary of the `dlcross` objects.

```
R> dlin2 <- dlcross(format = "rqtl", genobj = stmap, pheobj = stpheno,
+    idname = "id", fixpos = 1)
R> summary(dlin2)

This is an object of class dlcross.
Summary of genetic and phenotypic data:

         This is a bc population.
         No. individuals:          183
         No. phenotypic traits:    13
         Percent phenotyped:        100 100 100 100 100 100 100 100 100 100
99.5614 100 100

         No. chromosomes:           21
         Total markers:             287
         Percent genotyped:         92.98246
         No. markers per chr:       9 16 15 13 12 22 12 16 13 19 13 8 18 17 8
5 16 6 31 13 5

There are 183 unique genotypes and 456
unique phenotyped individuals.
```

Compared to the previous example, we have more markers, more linkage groups, and a less even distribution of markers across the genome. Additionally, design data has been recorded to help account for variability across both field and milling stages. These are of particular importance in constructing the base model, as Smith *et al.* (2006) showed that this trait is subject to large amounts of non-genetic sources of variation. In the base model we include fixed terms to identify doubled haploid or commercial lines (`Type`), and mean-centered milling order (`lord`) and row (`lrow`) to capture the natural linear trends occurring in the samples in these two stages. Design components in the field are further accounted for in the random effects (`Rep`, `Range`, `Row`) as well as milling day (`MillDay`) and a polygenic effect for the doubled haploid lines (`id`). The milling design (milling order and day) covariates are further used to model correlation between the observations.

```
R> basemodel2 <- asreml(myield ~ Type + lord + lrow, random = ~ id +
+    Rep + Range:Row + Millday, rcov = ~ Millday:ar1(Millord),
+    data = stpheno, na.method.X = "include")
```

After running the analysis with **asreml**, we compare the results of the analysis using **dlmap** to those in the original 2007 paper, where analysis was performed using R/**wgaim**. In Table 3 we present the chromosome for detected QTL using both packages, the position of the QTL and the Z statistic for the effect sizes. Note that the output from **wgaim** is in the form of marker intervals, while **dlmap** can potentially estimate QTL positions anywhere along the genome. For this analysis, thus, we restrict the set of possible positions to markers and the midpoints of intervals for better comparison with the results from **wgaim**.

```
R> dlout2 <- dlmap(dlin2, phename = "myield", baseModel = basemodel2,
+    algorithm = "asreml")
```

Verbyla *et al.* (2007) detected nine QTL in this population, on chromosomes 1B, 2B, 4A, 4B, 4D, 5A, 6B and 7D. Using **dlmap** we detect all but the QTL on chromosome 4A; however, we note that this chromosome has an adjusted p value of 0.09 in the initial detection stage and is hence of marginal significance. This disparity in results may be resolved by noting that **dlmap** provides more rigorous control of type I error than does **wgaim**.

5.3. Association population

Our final example considers an association mapping population to highlight the behavior of **dlmap** in a study with a) high-dimensional marker data and b) population structure and

	wgaim		**dlmap**	
Chromosome	Position (cM)	Z statistic	Position (cM)	Z statistic
1B	(0, 9)	2.57	11.0	2.48
1B	(90.9, 239.5)	−4.50	162.4	−4.36
2B	(54.8, 59.6)	9.31	51.9	12.06
4A	(10.3, 11.4)	2.49		
4B	(0, 12)	−4.82	4.9	−6.67
4D	(0, 1.8)	3.90	9.6	3.98
5A	(95.1, 102.1)	−3.79	91.7	−5.58
6B	(8.9, 9.4)	−5.45	9.1	−7.01
7D	(86.6, 94)	4.01	65.5	4.73

Table 3: QTL locations previously detected in Sunco x Tasman population compared to those using **dlmap**.

kinship effects. This example is based on the association mapping population of 95 lines of *Arabidopsis thaliana* described in Zhao *et al.* (2007). The lines formed a structured sample and were originally used to demonstrate the confounding effect of structure on association mapping in traits related to flowering time. Phenotypic data from the study was already adjusted for non-genetic effects.

While measurements were replicated for the original data, Zhao *et al.* (2007) found that the variance across replicates was very small compared to the variance across accessions. They report that including it in the analyses did not seem to affect results, and hence their analysis was performed using accession means. For best comparison with their published results we do the same using their publically available data.

We focus on a single trait from the study to demonstrate the potential of **dlmap** to detect associated loci in the presence of confounding genetic effects. We analyze the trait "vernalization response under short days", denoted by ± V (SD) in Zhao *et al.* (2007) and VSD in the following. This is computed as the ratio of short days without vernalization (SD) to short days with 5-week vernalization (SDV). Figure 2 depicts the distribution of the trait across accessions.

```
R> dlassoc3 <- dlcross(format = "other", genobj = gen3,
+    pheobj = ph3, mapobj = mapobj, idname = "acc", estmap = FALSE)
R> summary(dlassoc3)
```

```
This is an object of class dlcross.
Summary of genetic and phenotypic data:

         This is an association mapping population.

         No. individuals:         89
         No. phenotypic traits:   23
         Percent phenotyped:      100 100 98.8764 96.62921 97.75281
78.65169 86.51685 83.14607 89.88764 100 100 100 100 100 100 100 100
100 100 95.50562 98.8764 96.62921 82.02247
```

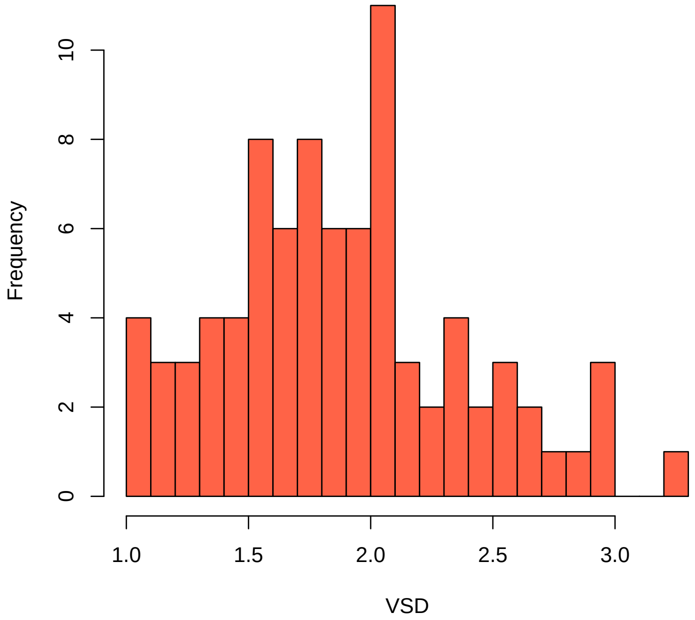

Figure 2: Distribution of VSD trait over 89 accessions.

```
No. chromosomes:          5
Total markers:            5419
Percent genotyped:        100
No. markers per chr:      1441 870 913 1038 1157

There are 89 unique genotypes and 89 unique phenotyped individuals.
```

After computing the kinship matrix, we found that it was not positive semi-definite. Upon closer examination we found five groups of accessions who shared more than 0.99 of the same haplotypes (Knox-10, Pna-10 and RRS-10 from the USA; Got-22 and Got-7 from Germany; Tamm-2 and Tamm-27 from Finland; Bil-5 and Bil-7 from Sweden; Lov-1 and Lov-5 from Sweden). Removing six of these individuals (Knox-10, Pna-10, Got-22, Tamm-2, Bil-5 and Lov-1) produced a kinship matrix which was positive definite. Hence the remainder of the analysis proceeds with the reduced set of 89 accessions rather than the full set of 95.

In their analysis, Zhao *et al.* (2007) compared results from a naive analysis which includes no terms to account for kinship or population structure with models including one or both of these terms. They concluded that including kinship estimates (**K/K***) along with population structure assignments (**P/Q**) works well, although in many cases the kinship estimate only is sufficient to correct for population structure. We focus on the analysis including both kinship and population structure terms here, since it is straightforward to eliminate one or the other. Population structure assignments were originally computed with **STRUCTURE** v2.0 (Falush,

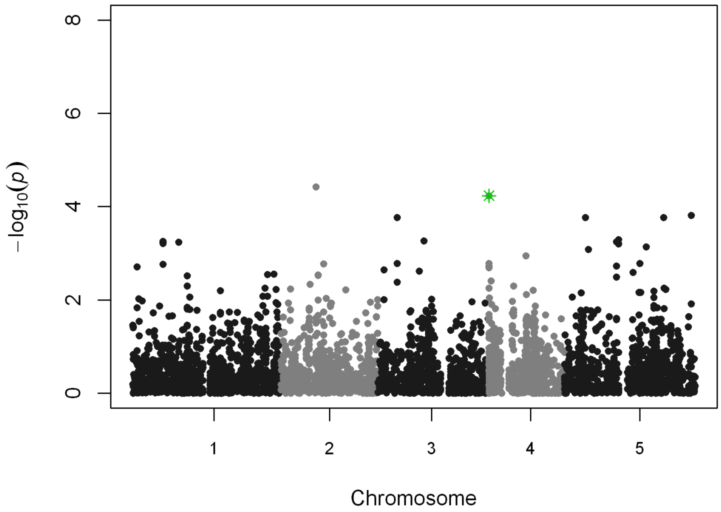

Figure 3: Manhattan plot for original QK* analysis of 5419 SNPs. SNP detected in **dlmap** analysis shown by star (green).

Stephens, and Pritchard 2007) and were included as supporting information in an earlier study of the same accessions (Nordborg *et al.* 2005).

```
R> basemodel3 <- asreml(fixed = VSD ~ Q1.8 + Q2.8 + Q3.8 +
+    Q4.8 + Q5.8 + Q6.8 + Q7.8, na.method.X = "include",
+    data = ph3)
R> res.assoc <- dlmap(dlassoc3, phename = "VSD", baseModel = basemodel3,
+    pedigree = kstar, algorithm = "asreml", alpha = 0.15)
R> summary(res.assoc)
```

```
 Summary of final results:
  Chr  Marker Effect    SD Z-value p-value
1   4 C4M3231  0.394 0.128    3.08  0.0021
```

In our analysis, we include fixed effect terms for the probability obtained from **STRUCTURE** that each accession derived from one of eight populations (`Q1.8, ..., Q7.8`), and a random polygenic effect with covariance matrix proportional to the kinship matrix (**K***) to match the models fit in Zhao *et al.* (2007). Note that either a pedigree or kinship matrix can be input via the argument `pedigree`; in both cases the inverse relationship matrix (either expected or empirical) will then be included in the model. In contrast to the experimental cross we do not input positions for the markers, only the chromosome to which they belong; additionally, we do not need to specify the arguments `step` or `fixpos`. For association mapping there is no localization stage - just detection, which performs forward selection on markers while accounting for background variation on the chromosome.

We allow for a less stringent genomewide significance threshold of 0.15 due to the small sample size. In the original analysis associations were detected for unadjusted p values < 0.001, which is a less conservative genomewide significance threshold. From the above results, we see that one significant association is detected on Chromosome 4, at the 3231st marker overall, which corresponds to base pair 153069.

In Zhao *et al.* (2007), they report the most promising associations as those which are significant across multiple methods. While we do not expect our results to match up exactly due to the differences in preliminary processing, it is satisfying to see that the SNP detected using **dlmap** is also significant in the original analyses (available at `doi:10.371/journal.pgen.0030004/sd003`). In particular, for the original **QK*** analysis this SNP has a p value of $5.82 \cdot 10^{-5}$. In Figure 3, we have plotted the $-\log_{10}(p)$ values for the 5419 single marker tests from the original analysis using the approach including both population structure (**Q**) and kinship (**K***) effects. For this Manhattan plot we use the `qqman` function found at Turner and Bush (2011). The SNP detected in our analysis is highlighted on Chr 4, where it has the strongest association with the trait on the chromosome.

6. Summary

In this paper we have described a tool with wide-ranging applicability in genetic studies. Mixed modeling software is becoming available for both experimental crosses (e.g., VSN International 2011) and association mapping studies (e.g., Bradbury *et al.* 2007). Such approaches allow analyses to account for complex environmental, pedigree and design effects as well as genetic influences. The flexible framework of R/**dlmap** allows it to be used for mapping in both types of studies.

We have outlined the theory behind and usage of **dlmap** as well as demonstrating its use in a variety of examples. Unlike software packages such as **TASSEL** (Bradbury *et al.* 2007) and **GenStat** (VSN International 2011), which test single markers for genetic effects, **dlmap** takes a novel approach of using all markers at once to detect and then localize QTL. R/**wgaim** (Verbyla *et al.* 2007) also implements a genomewide interval mapping approach, but is applicable only to backcrosses, doubled haploids and recombinant inbred lines. The extension of **dlmap** to F2 populations allows modeling of dominance (as well as additive) QTL effects, while the extension to association mapping populations broadens its appeal for general genetic studies. Currently, **dlmap** cannot be used in the highly computational detection of epistatic (genetic interaction) or genetic x environment interaction effects, but this is an active area of research for us.

Two-stage approaches are common in QTL analysis and can be quite efficient if properly done (Smith, Cullis, and Thompson 2005; Möhring and Piepho 2009). Adjusted means are computed for each genotype, and then used in a weighted mixed model genetic analysis. These can be implemented using **dlmap** in a straightforward manner by modifying the input data and base model. Indeed, the third example (association mapping) uses a two-stage approach, although it will be less than fully efficient since no weights are used. Altering the computation of the kinship matrix may also improve efficiency. We have attempted to replicate the construction of the kinship matrix in Zhao *et al.* (2007) for consistency, but other approaches to calculating genetic relationship matrices may resolve this issue (Maenhout, Baets, and Haesaert 2009). Alternately, it is possible to modify the kinship matrix to be positive semi-definite by forming a linear combination of kinship and pedigree matrix (VanRaden 2008), but we lack a pedigree in this example.

Mixed modeling is a computationally intensive endeavour, and most software packages are built on proprietary software for this task (e.g., SAS Institute Inc. 2004; VSN International 2011). To increase the usability of **dlmap**, it has been implemented using two different mixed model software packages in the free software R (R Development Core Team 2012). R/**nlme** is freely available from CRAN, and as of February 2011, R/**asreml** is freely available for all academic users and users from third-world countries.

The examples presented here rely on the mixed modeling software **asreml** rather than the freely available **nlme**. This is primarily due to the greater flexibility in modeling achievable with **asreml**. While **nlme** can be used to fit the basic DLMapping algorithm, it cannot be done in conjunction with modeling random effects or covariance structure, cannot handle more than 200 genetic markers, and only allows for a single phenotypic observation per genotype. As such it is more useful for basic analyses or two-stage approaches wherein first phenotypic modeling is performed, and then the genetic analysis on a simplified response. While our third example utilizes a two-stage approach by applying genetic analysis to means of lines, **nlme** still cannot be used due to the inclusion of the kinship matrix in the genetic analysis.

Acknowledgments

We thank the editor and two anonymous reviewers for their helpful comments with the manuscript and the software code.

References

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). "**TASSEL**: Software for Association Mapping of Complex Traits in Diverse Samples." *Bioinformatics*, **23**, 2633–2635.

Broman KW, Speed TP (2002). "A Model Selection Approach for the Identification of Quantitative Trait Loci in Experimental Crosses." *Journal of the Royal Statistical Society B*, **64**, 641–656.

Broman KW, Wu H, Sen S, Churchill GA (2003). "R/**qtl**: QTL Mapping in Experimental Crosses." *Bioinformatics*, **19**, 889–890.

Churchill G, Doerge R (1994). "Empirical Threshold Values for Quantitative Trait Mapping." *Genetics*, **138**, 963–971.

Falush D, Stephens M, Pritchard JK (2007). "Inference of Population Structure Using Multilocus Genotype Data: Dominant Markers and Null Alleles." *Molecular Ecology Notes*, **7**, 895–908.

Gilmour AR (2007). "Mixed Model Regression Mapping for QTL Detection in Experimental Crosses." *Computational Statistics & Data Analysis*, **51**, 3749–3764.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009). ***ASReml** User Guide*, release 3.0 edition.

Huang B, George A (2009). "Look Before You Leap: A New Approach to Mapping QTL." *Theoretical and Applied Genetics*, **119**, 899–911.

Jansen RC (1993). "Interval Mapping of Multiple Quantitative Trait Loci." *Genetics*, **135**, 205–211.

Kao CH, Zeng ZB, Teasdale RD (1999). "Multiple Interval Mapping for Quantitative Trait Loci." *Genetics*, **152**, 1203–1216.

Lander ES, Botstein D (1989). "Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps." *Genetics*, **121**, 185–199.

Lee SH, Van der Werf JHJ (2006). "Simultaneous Fine Mapping of Multiple Closely Linked Quantitative Trait Loci Using Combined Linkage Disequilibrium and Linkage with a General Pedigree." *Genetics*, **173**, 2329–2337.

Lehmensiek A, Eckermann PJ, Verbyla AP, Appels R, Sutherland MW, Daggard GE (2005). "Curation of Wheat Maps to Improve Map Accuracy and QTL Detection." *Australian Journal of Agricultural Research*, **56**, 1347–1354.

Li H, Ribaut JM, Li Z, Wang J (2008). "Inclusive Composite Interval Mapping (ICIM) for Digenic Epistasis of Quantitative Traits in Biparental Populations." *Theoretical and Applied Genetics*, **116**, 243–260.

Maenhout S, Baets BD, Haesaert G (2009). "Marker-Based Estimation of the Coefficient of Coancestry in Hybrid Breeding Programmes." *Theoretical and Applied Genetics*, **118**, 1181–1192.

Möhring J, Piepho HP (2009). "Comparison of Weighting in Two-Stage Analysis of Plant Breeding Trials." *Crop Science*, **49**, 1977–1988.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005). "The Pattern of Polymorphism in *Arabidopsis thaliana*." *PLoS Biology*, **3**, e196.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Development Core Team (2011). ***nlme**: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-102, URL `http://CRAN.R-project.org/package=nlme`.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

SAS Institute Inc (2004). *SAS 9.1.3 Help and Documentation*. SAS Institute Inc., Cary, NC. URL `http://www.sas.com/`.

Self SG, Liang KY (1987). "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions." *Journal of the American Statistical Association*, **82**, 605–610.

Smith AB, Cullis BR, Thompson R (2005). "The Analysis of Crop Cultivar Breeding and Evaluation Trials: An Overview of Current Mixed Model Approaches." *The Journal of Agricultural Science*, **143**, 449–462.

Smith AB, Lim P, Cullis BR (2006). "The Design and Analysis of Multi-Phase Plant Breeding Experiments." *The Journal of Agricultural Science*, **144**(05), 393–409.

Soller M, Brody T (1976). "On the Power of Experimental Designs for the Detection of Linkage between Marker Loci and Quantitative Loci in Crosses between Inbred Lines." *Theoretica*, **47**, 35–39.

Taylor J, Verbyla A (2011). "R Package **wgaim**: QTL Analysis in Bi-Parental Populations Using Linear Mixed Models." *Journal of Statistical Software*, **40**(7), 1–18. URL `http://www.jstatsoft.org/v40/i07/`.

Turner S, Bush W (2011). "Getting Genetics Done: Annotated Manhattan Plots and QQ Plots." URL `http://gettinggeneticsdone.blogspot.com.au/2011/04/annotated-manhattan-plots-and-qq-plots.html`.

Van Ooijen JW, Maliepaard C (1996). "**MapQTL** Version 3.0: Software for the Calculation of QTL Positions on Genetic Maps." In *Plant Genome IV Conference*. San Diego.

VanRaden PM (2008). "Efficient Methods to Compute Genomic Predictions." *Journal of Dairy Science*, **91**, 4414–4423.

Verbyla A, Cullis BR, Thompson R (2007). "The Analysis of QTL by Simultaneous Use of the Full Linkage Map." *Theoretical and Applied Genetics*, **116**, 95–111.

VSN International (2011). ***GenStat** for Windows 14th Edition.* Hemel Hempstead, UK.

Wang DL, Zhu J, Li ZK, Paterson AH (1999). "Mapping QTLs with Epistatic Effects and QTL x Environment Interactions by Mixed Linear Model Approaches." *Theoretical and Applied Genetics*, **99**, 1255–1264.

Wang S, Basten CJ, Zeng ZB (2005). ***Windows QTL Cartographer 2.5***. Raleigh.

Whittaker CJ, Curnow RN, Haley CS, Thompson R (1995). "Using Marker-Maps in Marker-Assisted Selection." *Genetical Research*, **66**, 255–265.

Zeng ZB (1993). "Theoretical Basis for Separation of Multiple Linked Gene Effects in Mapping Quantitative Trait Loci." *Proceedings of the National Academy of Sciences USA*, **90**, 10972–10976.

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007). "An *Arabidopsis* Example of Association Mapping in Structured Samples." *PLoS Genetics*, **3**, e4.

Affiliation:

B. Emma Huang, Rohan Shah, Andrew W. George
Division of Mathematics, Informatics and Statistics and
Food Futures National Research Flagship

Commonwealth Scientific and Industrial Research Organisation (CSIRO)
Queensland Ecosciences Precinct
41 Boggo Road
Dutton Park QLD 4102, Australia
E-mail: Emma.Huang@csiro.au, Rohan.Shah@csiro.au, Andrew.George@csiro.au
URL: http://www.csiro.au/people/Emma.Huang.html

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 50, Issue 6 *Submitted:* 2010-12-08
August 2012 *Accepted:* 2012-05-08