ORIGINAL ARTICLE

A robust multiple-locus method for quantitative trait locus analysis of non-normally distributed multiple traits

Z Li^{1,2,3,4}, J Möttönen⁵ and MJ Sillanpää^{1,3,4}

Linear regression-based quantitative trait loci/association mapping methods such as least squares commonly assume normality of residuals. In genetics studies of plants or animals, some quantitative traits may not follow normal distribution because the data include outlying observations or data that are collected from multiple sources, and in such cases the normal regression methods may lose some statistical power to detect quantitative trait loci. In this work, we propose a robust multiple-locus regression approach for analyzing multiple quantitative traits without normality assumption. In our method, the objective function is least absolute deviation (LAD), which corresponds to the assumption of multivariate Laplace distributed residual errors. This distribution has heavier tails than the normal distribution. In addition, we adopt a group LASSO penalty to produce shrinkage estimation of the marker effects and to describe the genetic correlation among phenotypes. Our LAD-LASSO approach is less sensitive to the outliers and is more appropriate for the analysis of data with skewedly distributed phenotypes. Another application of our robust approach is on missing phenotype problem in multiple-trait analysis, where the missing phenotype items can simply be filled with some extreme values, and be treated as outliers. The efficiency of the LAD-LASSO approach is illustrated on both simulated and real data sets.

Heredity (2015) **115**, 556–564; doi:10.1038/hdy.2015.61; published online 15 July 2015

INTRODUCTION

The association mapping has become a powerful tool for detecting quantitative trait loci (QTL) in plant and animal genomes. Standard regression-based methods assume that the quantitative traits are normally distributed. In reality, the sample may contain some outlying phenotypic observations (Estaghirou *et al.*, 2014). Besides, the data might be collected from different environments or populations, which result in quantitative trait following some skewed distributions (for example, with heavier tails than the normal distributions). Directly applying a normal distribution-based QTL/association mapping method on such skewedly distributed data or data with outliers may cause biased estimation of QTL effects and QTL locations (for example, reducing the statistical power to detect true QTLs).

One possible solution for handling non-normality is to perform some transformation to increase the normality in the phenotype data (Goh and Yap, 2009). Examples are logarithm, Box-Cox, arcsine (for proportion data) and rank-based transformations. However, it is generally unclear which transformation method is most appropriate for a particular data set. Alternatively, some statistical approaches have been proposed, which specify the residual errors in their models to follow some skewed distribution such as skew-normal distribution, Student's *t*-distribution or Laplace distribution instead of normal distribution (Fernandes *et al.*, 2007; Wang *et al.*, 2009). Intuitively, such a regression model should be more robust to the non-normally distributed phenotype data and outliers. In this article, we focus on the Laplace residual errors corresponding to the least absolute deviations

(LAD) method (Li *et al.*, 2010), which is regarded as a robust alternative to the classical least squares estimation.

In association studies, multiple-locus methods refer to a class of regression methods that attempts to simultaneously estimate the additive effects of many markers in a single model. As a large number of markers (for example, single-nucleotide polymorphisms (SNPs)) are present in many association mapping data sets, the multiple-locus methods usually involve some variable selection or parameter regularization procedures to reduce the dimensionality of the model. Least absolute shrinkage and selection operator or the LASSO (Tibshirani, 1996) is a popular regularization method, and has been widely used for analyzing high-dimensional data in various applied fields including genetics (Li and Sillanpää, 2012). It is possible to combine the LASSO regularization with the LAD method to produce sparse estimation of genetic effects, which is also resistant to the outliers or non-normality in the phenotype data. In this work, we further generalize the idea of LASSO and LAD to identify the association between markers and multiple correlated quantitative traits or a single trait measured at multiple time points (that is, longitudinal data). Our approach is developed on the basis of Möttönen and Sillanpää (2015). In their work, they focused on variable selection and coefficient estimation of the QTL effects; whereas in our work, we incorporated multiple hypothesis testing to formally judge QTLs, and evaluated the methods on more biologically meaningful examples. Compared with the single-trait method, the multiple-trait analysis with proper modeling of grouping and covariance structure in the data

¹Biocenter Oulu, University of Oulu, Oulu, Finland; ²Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland; ³Department of Mathematical Sciences, University of Oulu, Oulu, Finland; ⁴Department of Biology, University of Oulu, Oulu, Finland and ⁵Department of Social Research, University of Helsinki, Helsinki, Finland
Correspondence: Professor MJ Sillanpää, Departments of Mathematical Sciences and Biology, University of Oulu, P.O. Box 3000, FI-90014 Oulu, Finland.
E-mail: Mikko.Sillanpaa@oulu.fi

Received 19 December 2014; revised 25 May 2015; accepted 26 May 2015; published online 15 July 2015

may have the benefit to investigate the pleiotropy—the same gene influencing multiple traits which are (genetically) correlated.

The structure of this article is as follows. Materials and methods describe the robust multivariate LAD-LASSO model, its computational algorithm and the relevant QTL decision rules. In Results, our method was illustrated with a QTLMAS2009 simulated data set (Coster *et al.*, 2010) and a real sugar beet data set (Würschum *et al.*, 2011). We focus on the comparison between our robust multivariate LAD-LASSO and the multivariate normal LASSO method. Finally, we discuss the strengths and weaknesses of our LAD approach, and point out the future research directions.

MATERIALS AND METHODS

LAD-LASSO for single-trait analysis

A normal multiple linear regression for single-trait association mapping is defined by

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_0^2), \quad (1)$$

where y_i is the phenotype value of individual i ($i = 1, \dots, n$), $\mathbf{x}_i = [1, x_{i1}, \dots, x_{ip}]^T$, x_{ij} is the genotype value of individual i and marker j ($j = 1, \dots, p$) coded as -1 , 0 and 1 for three SNP genotypes AA, AB and BB, respectively, $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_p]^T$, β_0 is an intercept term, β_j is the effect of marker j , and ε_i is the residual error, which is assumed to follow mutually independent normal distribution with zero mean and unknown variance σ_0^2 .

According to normal residual errors in equation (1), the classical ordinal least squares (OLS) seek the minimization of an l_2 objective function: the sum of squared error function $\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$. The OLS estimates might not be optimal for the non-normally distributed phenotype data. A robust alternative is the LAD estimates, which minimize the sum of absolute error or l_1 objective function $\sum_{i=1}^n |y_i - \mathbf{x}_i^T \boldsymbol{\beta}|$. The LAD estimates are in line with assuming the residual errors ε_i of model (1) to follow an i.i.d. Laplace distribution (Li *et al.*, 2010):

$$\text{La}(\varepsilon_i | 0, \sigma_0) = \frac{1}{2\sigma_0} \exp\left(-\frac{|\varepsilon_i|}{\sigma_0}\right).$$

A multiple-locus method for association mapping aims at estimating and testing for the effects of a huge number of SNPs simultaneously. In such a case, standard OLS or LAD methods often fail to provide such accurate estimates due to the limited sample size in the data. An improved estimation can be obtained by using regularization—adding a certain penalty function of SNP effects β_j to the sum of squared error or sum of absolute error function. We consider the LASSO, or the l_1 norm regularized method, defined for the sum of squared error case:

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left[\frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j| \right] \quad (2)$$

and for the sum of absolute error case:

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left[\frac{1}{n} \sum_{i=1}^n |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| + \lambda \sum_{j=1}^p |\beta_j| \right]. \quad (3)$$

We name (2) and (3) as normal LASSO and LAD-LASSO, respectively. In both (2) and (3), the inclusion of l_1 norm penalty produces a sparse solution by shrinking the effects of many unimportant markers to zero. Note that no penalty is added to the intercept term β_0 . The tuning parameter $\lambda (> 0)$ determines the amount of shrinkage and degree of sparsity, an optimal value of which can typically be chosen by some model selection criteria such as cross validation (Hastie *et al.*, 2009).

The normal LASSO estimates of (2) can be obtained by algorithms such as LARS (Efron *et al.*, 2004) and coordinate descent method (Friedman *et al.*, 2010). In LAD-LASSO (3), as both the sum of absolute error function and penalty term is in l_1 norm, it is possible to combine these two components into one. Let $(y_i^*, \mathbf{x}_i^*) = (y_i, \mathbf{x}_i)$, if $i = 1, \dots, n$, and $(y_i^*, \mathbf{x}_i^*) = (0, n\lambda \mathbf{e}_{-n})$, if $i = n+1, \dots,$

$n+p$, where $\mathbf{e}_i$ is a $1 \times p$ standard unit vector with the i th element to be 1, and others to be 0. Now the LAD-LASSO problem (3) can be re-defined as

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \frac{1}{n+p} \sum_{i=1}^{n+p} |y_i^* - \mathbf{x}_i^{*T} \boldsymbol{\beta}|. \quad (4)$$

Obviously, equation (4) becomes a standard LAD problem, which implies that any estimation algorithm developed for LAD or quantile regression (for example, Koenker and d'Orey, 1987; 1994) can be directly used for computing LAD-LASSO solutions.

Group LAD-LASSO for multiple-trait analysis

Now we turn to the multivariate regression for mapping multiple correlated traits, defined by

$$\mathbf{y}_i = \mathbf{B}^T \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \stackrel{\text{i.i.d.}}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_0), \quad (5)$$

where $\mathbf{y}_i = [y_{i1}, \dots, y_{im}]^T$ is an $m \times 1$ observation vector of the i th individual on traits $k = 1, \dots, m$, $\mathbf{x}_i$ is a vector of genotype data, which is defined in the same way as in equation (1), $\mathbf{B}$ is a $(p+1) \times m$ matrix consisting of effect for the j th SNP on m th trait, and $\boldsymbol{\varepsilon}_i$ is a vector of residual errors, following a multivariate normal distribution with zero mean vector and covariance matrix $\boldsymbol{\Sigma}_0$. Like in the univariate regression (1), the OLS estimates for (5) minimize an l_2 objective function: $\sum_{i=1}^n (\mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i)^T (\mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i) = \sum_{i=1}^n \|\mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i\|^2$, where

$$\|\mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i\| = \sqrt{\sum_{k=1}^m (y_{ik} - \mathbf{B}_{\bullet k}^T \mathbf{x}_i)^2}, \quad \text{with locus effects of each trait}$$

$\mathbf{B}_{\bullet k} = (B_{0k}, B_{1k}, \dots, B_{pk})^T$. In fact, the least squares estimates of the multivariate regression are given by assuming the covariance matrix $\boldsymbol{\Sigma}_0$ to be an identity matrix, which is equivalent to the separate OLS estimates of each single trait, so that the possible covariance structure among multiple traits is not considered in the estimation stage. However, the covariance structure might be taken into account in the later hypothesis testing stage (Knott and Haley, 2000).

On the other hand, the robust LAD estimates for the multivariate regression minimize an l_1 objective function:

$$\sum_{i=1}^n \|\mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i\|.$$

The multivariate LAD corresponds to a multivariate Laplace residual error assumption: $\boldsymbol{\varepsilon}_i \stackrel{\text{i.i.d.}}{\sim} \text{MVLA}(\mathbf{0}, \boldsymbol{\Sigma}_0)$, with the scale matrix $\boldsymbol{\Sigma}_0$ (which is proportional to the covariance matrix) to be an identity matrix just like in the least squares case. According to Gómez *et al.*, (1998), the density function of the multivariate Laplace distribution is defined by $\text{MVLA}(\boldsymbol{\varepsilon}_i | \mathbf{0}, \boldsymbol{\Sigma}_0) = C |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} [\boldsymbol{\varepsilon}_i^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\varepsilon}_i]^{\frac{1}{2}}\right\}$, where $C = \frac{m \Gamma(\frac{m}{2})}{\pi^{\frac{m}{2}} \Gamma(1+m) 2^{1+m}}$.

In addition, the l_1 and l_2 multivariate regressions are extended to the following penalized regression forms:

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \left[\frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i\|^2 + \lambda \sum_{j=1}^p \|\mathbf{B}_{\bullet j}\| \right], \quad (6)$$

and

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \left[\frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i\| + \lambda \sum_{j=1}^p \|\mathbf{B}_{\bullet j}\| \right], \quad (7)$$

respectively.

Here, $\lambda \sum_{j=1}^p \|\mathbf{B}_{\bullet j}\| = \lambda \sum_{j=1}^p \sqrt{\sum_{k=1}^m B_{jk}^2}$ is a group LASSO penalty (Yuan and Lin,

2006). Such a penalty is added to the effect parameters of each loci, but not to the intercept $\mathbf{B}_0$. By using the group LASSO penalty, selecting a marker is equivalent as selecting a vector of marker's effects on each trait instead of selecting a single parameter, which differs from standard LASSO penalty in equations (2) and (3). Thus, the group LASSO penalty takes the dependencies among multiple traits explained by genetic effects into account, although in the original l_1 and l_2 objective functions, the multiple traits are in fact assumed to be independent from each other. We name equations (6) and (7) as

multivariate group LASSO and multivariate group LAD-LASSO, respectively (abbreviated as MG-LASSO and MGLAD-LASSO).

Like in the univariate case, it is possible to combine the two components of MGLAD-LASSO into one. Let $(y_i^*, x_i^*) = (y_i, x_i)$ if $i = 1, \dots, n$, and $(y_i^*, x_i^*) = (0, n\lambda e_{i-n})$, if $i = n+1, \dots, n+p$. Equation (7) can now be rewritten as

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \frac{1}{n+p} \sum_{i=1}^{n+p} \|y_i^* - \mathbf{B}^T x_i^*\|, \quad (8)$$

which is a standard multivariate LAD problem. To solve equation (8), we used a fixed point algorithm introduced in Oja (2010), which is implemented by the R-package 'MNM' (Nordhausen and Oja, 2011). More details are explained in Appendix A.

The solution of MG-LASSO can be computed by a coordinate descent algorithm, implemented in the R-package 'glmnet' (Friedman *et al.*, 2010; <http://cran.rproject.org/web/packages/glmnet/glmnet.pdf>).

QTL decision rules

By using the l_1 penalized regression, the effects of unimportant markers are shrunk towards zero, and only the relatively important markers are left in the model. However, in practice, it is very difficult to choose an 'optimal' tuning parameter λ leading to the selection of a true set of QTLs. Use of standard criteria such as cross validation for determining λ may result in a large number of false-positive signals (Li and Sillanpää, 2012). Therefore, some QTL decision rules are needed to further filter out the false-positive markers. For both MG-LASSO and MGLAD-LASSO, we considered two decision-making procedures, including a stability selection method (Meinshausen and Bühlmann, 2010) and a multiple-split test method (Meinshausen *et al.*, 2009) to formally judge the QTLs.

In stability selection, a subsample of the data is randomly chosen including half the number of individuals. The LASSO analysis is performed on this subsample of the data, and the set of markers with non-zero effects are being recorded. This procedure is repeated for many times (say 100 times). Then the stability selection probability of each marker being selected is calculated, and used to judge the support for QTLs. Meinshausen and Bühlmann (2010) suggested a decision rule on the basis of stability selection probabilities to control the expected number of false positives.

In multiple-split test, the data are randomly divided into two parts, preferably with equivalent sample size. The MGLAD-LASSO/MG-LASSO is implemented on the first part of the data to choose a subset of markers. In the second half of the data, the standard normal multivariate LAD regression (or least squares) without penalty are used on those selected markers, hypothesis testing is performed using unpenalized effect estimates, and the corresponding P -values for each selected marker are calculated. Such a procedure is repeated many times, and the P -values for each marker from multiple runs are combined into one and used for judging QTLs. The multiple-split test can control family-wise error, which should be more conservative than a method controlling false discovery rate such as the stability selection. We have earlier used the stability selection and the multiple-split test in the univariate case for the normal LASSO in Li *et al.* (2014).

More details about stability selection and multiple-split test method are shown in Appendix B and Appendix C, respectively. Some R codes for implementing these methods are provided as Supplementary Information.

RESULTS

We consider (i) a simulated longitudinal QTLMAS2009 data set (Coster *et al.*, 2010) and (ii) a real sugar beet data set (Würschum *et al.*, 2011; Würschum and Kraft, 2013) for evaluating our proposed LAD-LASSO methods. The two data sets are available from their original publications: Coster *et al.* (2010) and Würschum and Kraft (2013), respectively.

Analysis and results: QTLMAS simulated data

The simulated data set includes 453 SNP markers distributed over five chromosomes of 1M each from 2025 individuals with a certain population structure and the growth traits following logistic curves

$y(t) = \phi_1 / \{1 + \exp[(\phi_2 - t)/\phi_3]\}$ measured over five time points (0, 132, 265, 397 and 530 days). Measurements at each single time point can be thought as a separate 'quantitative trait', and therefore a multiple-trait method can be used for analyzing such a longitudinal data set. In total, 18 additive QTLs were simulated, with six contributing to each of the three parameters ϕ_1 (asymptotic yield), ϕ_2 (inflection point) and ϕ_3 (slope of the curve) of a growth curve. Besides, some i.i.d. Gaussian errors were also simulated to these three trait parameters. A subsample of 1000 individuals is used in our simulation analyses. The same subsample of individuals was also analyzed in Heuven and Janss (2010) and Li and Sillanpää (2013) by other approaches. In addition, for a group of markers with pairwise correlation coefficients higher than 0.9, only one marker was picked up to represent the whole group of the highly correlated markers to reduce the collinearity of the genotype data, and, therefore, there are only 393 markers used in the LASSO analysis.

Scenario 1: original QTLMAS data. Here, the MGLAD-LASSO and MG-LASSO methods were applied on the original simulated QTLMAS2009 data set. Note that neither MGLAD-LASSO nor MG-LASSO takes the serial correlation simulated in the QTLMAS2009 data set into account, and may not be optimal for analyzing such a longitudinal data set. In contrast, a varying coefficient method might be more appropriate for modeling the longitudinal data (Li and Sillanpää, 2013). Nevertheless, this data set can still be used to compare the performances of the two group LASSO methods. Both the methods were used together with the stability selection and multiple-split test to judge QTLs. In stability selection, we judge a marker as a putative QTL, if its stability selection probability is larger than a threshold, to guarantee the average number of false selected markers to be no greater than 1 (see Appendix B for more details). In multiple-split test, a marker is declared as significant if its P -value is < 0.05 . We claim that a true simulated QTL is successfully identified if a significant marker is located within 10 cM of this simulated QTL, and otherwise the marker was judged as a false positive. Note that the same QTL identification rule will also be used in Scenarios 2 and 3 (see below). Table 1 reports the detected putative QTLs. Both MGLAD-LASSO and MG-LASSO analyses correctly detected four or five QTLs, indicating that their performances were similar on the normally distributed traits. Comparing the QTL decision rules, the stability selection and multiple-split testing seemed to detect slightly different sets of QTLs, and none of them produced any false positive.

Scenario 2: missing data. Missing phenotype data is an important issue in multiple traits or longitudinal studies. A common situation is that the missing values only occur at a small fraction of traits or time points for one individual. In this case, simply deleting these individuals with few missing values is a pity thing to do, and may result in losing essential information from the data. Here, we investigate how to use MGLAD-LASSO to simply handle the missing data problem in multivariate data. In the QTLMAS2009 phenotype data, 5% of individuals (corresponds to 50 individuals) were randomly chosen, and for each individual, the items were specified as missing values at one or two randomly selected time points. We assume that these simulated data points are missing completely at random. Two different approaches were used to impute once the missing data before the LASSO analysis. First, we applied a quadratic polynomial function to 'predict' the values of the missing data points based on known observations of each individual. Alternatively, we replaced all the missing points by an extreme value (say 1000), and the missing data can then be interpreted as outliers. Note that the original scale of the

Table 1 The results from the analysis of original simulated QTLMAS data, including the detected significant markers, their stability selection probabilities or *P*-values (in brackets) and the closest true simulated QTLs

MG-LASSO		MGLAD-LASSO		Simulated QTLs
Stability selection	Multi-split test	Stability selection	Multi-split test	
35 (1)	36 (2.3×10^{-9})	35 (1)	37 (0)	(36, 37)
37 (1)	37 (4.9×10^{-23})	37 (1)		
—	—	—	81 (0)	(77, 78)
98 (1)	98 (4.9×10^{-5})	98 (0.86)	98 (0.0003)	(98, 99)
99 (0.89)	99 (0.0002)			
174 (0.98)	—	174 (0.77)	—	(174, 175)
—	222 (0.001)	—	—	(216, 217)
—	—	270 (0.69)	—	(276, 277)
—	—	—	291 (0.0008)	(293, 294)
338 (0.83)	338 (0.002)	—	338 (0.0118)	(336, 337)
439 (0.78)	—	—	—	(432, 433)
0	0	0	0	Number of false positives

Abbreviations: LAD, least absolute deviation; MG, multivariate group; QTL, quantitative trait locus.

Note that the true simulated QTLs are usually located between their adjacent markers. A true simulated QTL is claimed to be successfully identified if a significant marker is located within 10 cM of this simulated QTL, and, otherwise, the marker is judged as a false positive.

data is from 2 to 40. The simulation was replicated 50 times (in each replication, 50 individuals were randomly selected to simulate some missing items). The performance of MGLAD-LASSO and MG-LASSO were evaluated by the following two quantities calculated over the 50 replications: (i) the frequency of each QTL being correctly identified which measures the statistical power, (ii) the average number of false positive markers measuring the ability of controlling the type I error. Results are shown on Tables 2 and 3 for the two imputed data sets, respectively.

When the missing data were imputed by using quadratic polynomial functions (Table 2), both the MGLAD-LASSO and MG-LASSO methods were able to detect the same set of QTLs as in the complete data case (cf. Table 1), and well control the false positives. This is not surprising, because the phenotype trajectories were originally simulated following monotonic logistic curves and the quadratic polynomial functions can provide accurate prediction to those missing values.

When the missing items were filled by an extreme value (Table 3), MG-LASSO failed to correctly detect any QTL, because it is not robust to the outliers. However, MGLAD-LASSO was still able to find those QTLs with high statistical power, indicating that those artificial outliers do not affect the MGLAD-LASSO analysis. Therefore, in addition to some conventional phenotype imputation methods (Guo and Nelson, 2008), simply filling the missing data with extreme values may also serve as a reasonable approach for pre-processing the phenotype data before the MGLAD-LASSO analysis.

Scenario 3: non-normally distributed data. On the basis of the simulated QTLMAS2009 marker information, we re-simulated five correlated quantitative traits, which follow a non-normal distribution. In detail, we simulated five QTLs exactly at markers 35, 98, 216, 291 and 338. The additive effects of these five QTLs were simulated from a normal distribution with mean 1 and variance 0.25 on each of the five

traits. The intercept (population mean) effects were simulated from a normal distribution with mean 10 and variance 1. The residual errors were simulated from a multivariate Student's *t*-distribution $MVT(0, v, \Sigma_0)$, with degree of freedom $v=1$ (in this case, the multivariate Student's *t*-distribution becomes a multivariate Cauchy distribution) and scale matrix

$$\Sigma_0 = 5 \cdot \begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{bmatrix}.$$

This simulation was again repeated 50 times and the average performances of the two methods are shown in Table 4. The MGLAD-LASSO method was able to detect all the five simulated QTLs with higher frequencies than the MG-LASSO, indicating that MGLAD-LASSO is more robust against the mis-specification of the model likelihood. However, MGLAD-LASSO and MG-LASSO performed rather similarly in terms of controlling the false positives. To better evaluate the ability of the MGLAD-LASSO/MG-LASSO methods on controlling false positives, we further simulated a null data set with only the intercept term but no QTL effects (repeated 50 times), and the residual errors following the same multivariate Student's *t*-distribution as above. Again, all the approaches produce low number of false positives (Table 5). Our observations are in line with some earlier published simulation studies to evaluate the performance of a univariate Bayesian robust method (Yang *et al.*, 2009). When comparing the QTL decision rules, the conservative multiple-split test outperformed the stability selection by giving less false positives.

Analysis and results: sugar beet data

The study is based on 924 diploid elite sugar beet inbred lines, with each line having two replications at one to seven locations in the year 2008. For simplicity, only one replication at each line was selected, and, therefore, there are exactly 924 individuals distributed over 13 locations being used in our analysis. Originally, six quantitative traits were under evaluation including three yield-related traits: white sugar yield (t ha^{-1}), sugar content (%), root yield (t ha^{-1}); and three quality-related traits: potassium content (K, mM), sodium content (Na, mM) and α -amino nitrogen content (N, mM). Würschum *et al.* (2011) performed a single-trait association mapping separately on each of the six traits. In our analysis, we performed the multiple-trait analysis simultaneously on the three yield-related traits including white sugar yield, sugar content and root yield. The histograms of the three yield-related traits are shown in Figure 1. The distribution of sugar content clearly departs from the normal distribution. Furthermore, 677 SNP markers were genotyped. These SNPs were randomly distributed across the sugar beet genome with an average marker distance of 1 cM.

The pre-processing of the data was done as follows. First, the phenotypes of white sugar yield, sugar content and root yield were standardized to have zero mean and unit variance, so that they are roughly on the same scale. Second, for the genotype data, a marker was excluded from the analysis if its minor allele frequency was <0.05 or the missing data rate was >0.1 . For each of the remaining markers, as the missing genotype rates were generally very low, the missing genotypes were simply imputed by the mean of known genotypes at the same position. Furthermore, like in the QTLMAS simulated data example, only one marker was kept in data to represent a group of highly correlated markers (with pairwise correlation >0.9). After these steps, a number of 481 markers were used in the asso-

Table 2 In the QTLMAS phenotype data, 50 individuals were simulated with one or two missing items

Simulated QTLs	MG-LASSO		MGLAD-LASSO	
	Stability selection	Multi-split test	Stability selection	Multi-split test
(36, 37)	1.00	1.00	1.00	1.00
(77, 78)	0.00	0.04	0.00	1.00
(98, 99)	1.00	1.00	1.00	1.00
(174, 175)	1.00	0.00	1.00	0.00
(216, 217)	0.00	0.64	0.04	0.14
(276, 277)	0.04	0.00	1.00	1.00
(293, 294)	0.00	0.20	0.00	0.00
(336, 337)	1.00	0.94	0.00	0.84
(432, 433)	1.00	0.00	0.00	0.00
Average number of false positives	0.00	0.30	0.72	0.00

Abbreviations: LAD, least absolute deviation; MG, multivariate group; QTL, quantitative trait locus.

For each individual, the missing values were predicted by a quadratic function based on the known observations. The results comprising the frequency of each simulated QTL has been correctly detected by MG-LASSO/MGLAD-LASSO analyses over 50 replications, and the average number of false positives. Note that the true simulated QTLs are usually located between their adjacent markers. A true simulated QTL is claimed to be successfully identified if a significant marker is located within 10 cM of this simulated QTL, and, otherwise, the marker is judged as a false positive.

Table 3 In the QTLMAS phenotype data, 50 individuals were simulated with one or two missing items

Simulated QTLs	MG-LASSO		MGLAD-LASSO	
	Stability selection	Multi-split test	Stability selection	Multi-split test
(36, 37)	0.02	0.00	1.00	1.00
(77, 78)	0.02	0.00	0.00	1.00
(98, 99)	0.00	0.00	1.00	1.00
(174, 175)	0.00	0.00	1.00	0.00
(276, 277)	0.04	0.04	0.74	0.00
(293, 294)	0.00	0.00	0.00	1.00
(336, 337)	0.04	0.00	0.00	0.58
Average number of false positives	0.84	0.44	1.22	0.22

Abbreviations: LAD, least absolute deviation; MG, multivariate group; QTL, quantitative trait locus.

Those missing data were simply filled by an extreme value before the statistical analysis. The results comprising the frequency of each simulated QTL has been correctly detected by MG-LASSO/MGLAD-LASSO analyses over 50 replications, and the average number of false positives. Note that the true simulated QTLs are usually located between their adjacent markers. A true simulated QTL is claimed to be successfully identified if a significant marker is located within 10 cM of this simulated QTL, and, otherwise, the marker is judged as a false positive.

ciation analysis. In the multivariate LASSO analyses, the 13 locations were considered as extra covariates, and they were just treated in the same way as the markers and were subjected to variable selection.

The results are summarized in Table 6. The MG-LASSO with either QTL decision rules detected markers 192, 259, 289, 349, 418 and 553 as putative QTLs. Most of the putative QTLs have also been found by a linear mixed model-based association mapping method in Würschum *et al.* (2011), except the one at marker 289. On the other hand, the MGLAD-LASSO was able to detect putative QTLs including

Table 4 The analysis of replicated data of the non-normally distributed phenotypes simulated based on the QTLMAS marker information

Simulated QTLs	MG-LASSO		MGLAD-LASSO	
	Stability selection	Multi-split test	Stability selection	Multi-split test
35	0.74	0.56	1.00	0.98
98	0.14	0.18	0.58	0.78
216	0.94	0.62	1.00	0.96
291	0.98	0.60	1.00	1.00
338	0.92	0.76	1.00	1.00
Average number of false positives	0.64	0	1.38	0

Abbreviations: LAD, least absolute deviation; MG, multivariate group; QTL, quantitative trait locus.

The results comprising the frequency of each simulated QTL has been correctly detected by MG-LASSO/MGLAD-LASSO analyses over 50 replications and the average false positives.

markers 27, 162, 184, 259, 349 and 411. Except the markers 27 and 162, all the others were also reported in Würschum (2011).

Interestingly, the MG-LASSO detected 9–13 location covariates as significant in addition to those putative QTLs, but MGLAD-LASSO only identified three location variables. This may indicate that the non-normality of the trait distributions under study was mainly caused by those location factors. The MG-LASSO needs to include the locations into the model to weaken the non-normality in the phenotype data. To illustrate this point, we used the following single-trait mixed-effect model to fit each of the three traits (in the original scale) separately:

$$y_{ij} = \mu + l_j + e_{ij}, \quad (9)$$

where y_{ij} is the phenotype of i th sugar beet line at the j th location, l_j is the effect of the j th location and e_{ij} is the residual term. Figure 2 shows the distribution of the estimated residuals $\hat{e}_{ij} = y_{ij} - \hat{\mu} - \hat{l}_j$. Clearly, after the correction of location effects, residuals become more normally distributed compared with the original phenotypes. On the other hand, MGLAD-LASSO does not need to include location variables because it is robust against skewedly distributed data. Thus, the MGLAD-LASSO method will be particularly suitable for data sets collected from multiple sources like the sugar beet data, but without having any information on which group/location/environment an individual belongs to.

DISCUSSION

This paper introduces a multivariate robust regression approach for QTL mapping of multiple traits. The method is developed on the basis of the well-known LASSO (Tibshirani, 1996) and group LASSO methods (Yuan and Lin, 2006) for shrinkage estimation of marker effects, and stability selection (Meinshausen and Bühlmann, 2010) and/or multiple-split test (Meinshausen *et al.*, 2009) for formally judging QTLs. Consequently, the method can be used to do simultaneous estimation and testing of thousands of markers. Different from the standard MG LASSO (MG-LASSO), which assumes the residual terms to follow a multivariate normal distribution, our LAD LASSO (MGLAD-LASSO) assumes the residuals to follow a multivariate Laplace distribution, which has heavier tails compared with the normal distribution. Because of this, MGLAD-LASSO should be a more appropriate method for a data set with phenotypic outliers or a data set with skewedly distributed traits. Our simulation studies clearly

Table 5 Simulation study under the null model based on the QTLMAS marker information: the average number of wrongly selected QTLs (false positives) over 50 replications

Null simulation	MG-LASSO		MGLAD-LASSO	
	Stability selection	Multi-split test	Stability selection	Multi-split test
Average number of false positives	0.26	0.02	0.82	0

Abbreviations: LAD, least absolute deviation; MG, multivariate group; QTL, quantitative trait locus.

Figure 1 The histogram of three yield-related traits including white sugar yield (WSY), sugar content (SC) and root yield (RY) in the sugar beet data set.

Table 6 The results from the analysis of the three yield-related traits: white sugar yield, sugar content and root yield of the sugar beet data, including the detected QTLs, their stability selection probabilities or *P*-values (in brackets)

	MG-LASSO		MGLAD-LASSO		Reported in literature
	Stability selection	Multi-split test	Stability selection	Multi-split test	
Marker ID	—	—	27 (0.99)	—	—
	—	—	162 (0.94)	—	—
	—	—	184 (1)	—	Würschum <i>et al.</i> (2011)
	192 (0.76)	192 (0.0007)	—	—	Würschum <i>et al.</i> (2011)
	259 (0.93)	259 (1×10^{-6})	—	259 (0.0070)	Würschum <i>et al.</i> (2011)
	289 (0.73)	289 (8×10^{-4})	—	—	—
	—	349 (8×10^{-5})	349 (0.79)	349 (0.0047)	Würschum <i>et al.</i> (2011)
	—	—	411 (0.73)	—	Würschum <i>et al.</i> (2011)
	418 (0.68)	—	—	—	Würschum <i>et al.</i> (2011)
	—	553 (0.0113)	—	—	Würschum <i>et al.</i> (2011)
Number of locations being detected	9	13	3	3	

Abbreviations: LAD, least absolute deviation; MG, multivariate group; QTL, quantitative trait locus.

Figure 2 The histogram of the estimated residuals of the three yield-related traits including white sugar yield (WSY), sugar content (SC) and root yield (RY) in the sugar beet data set, after correcting the location effects by a linear mixed model.

demonstrate that MGLAD-LASSO is superior to the MG-LASSO in terms of correctly detecting more QTLs for such kind of data sets. A drawback of the MGLAD-LASSO method is its high computational cost. For a data set with 924 markers, MGLAD-LASSO implemented by R-MNM package costs several minutes for a single run, but MG-LASSO implemented by the R-glmnet package only costs several seconds.

Nevertheless, MGLAD-LASSO is a suitable tool for analyzing some normally distributed traits as well. An interesting application of MGLAD-LASSO is on multiple traits with missing phenotype data. As shown in our simulation studies, it is possible to use a very simple strategy to handle such a missing data problem in which all the missing items are filled by some extreme values and treated as outliers. The MGLAD-LASSO method is resistant to those extreme values filled for missing items, and it provides reasonable parameter estimation and testing based on all the available data. Our suggestive strategy for handling missing phenotype data by using MGLAD-LASSO is connected to a missing data technique named the maximum-likelihood estimation using observed data likelihood (Baraldi and Enders, 2010), where they proposed to build a specific-likelihood function and do the inference on the basis of all the available data without treating any missing data.

A relevant research topic connected to association mapping of using multiple-locus models is genomic selection, which aims to estimate and predict the genomic breeding values based on a dense marker set instead of identifying any specific quantitative trait locus (Meuwissen *et al.*, 2001). Previous studies demonstrate that LASSO can also be used for genomic selection (Li and Sillanpää, 2012; Xu *et al.*, 2014). With simulation studies, Estaghirou *et al.* (2014) showed that

performance of standard genomic selection methods might be significantly affected when the data have outlying phenotypic observations. In addition, a multiple-trait genomic selection method might be superior to a single-trait genomic selection method in some circumstances (Yi and Jannink, 2012; Hayashi and Iwata, 2013). All these evidences indicate that our MGLAD-LASSO approach may potentially be an appropriate tool for genomic selection as well, especially when the outliers or missing items exist in the phenotype data. Further investigation is needed to measure the usability of the MGLAD-LASSO in genomic selection.

Our current MGLAD-LASSO method still has some room for improvement from the following three perspectives. First, the current fixed point algorithm for finding MGLAD-LASSO solutions is easy to derive, and should work for moderate-sized data sets with thousands of markers and individuals. However, it may not be efficient for genome-wide data sets with very large number of markers and individuals which have become more and more common recently. For improvement, more efficient algorithms are needed to be developed for MGLAD-LASSO such as the coordinate descent algorithm, which has been available for single-trait LAD-LASSO (Wu and Lange, 2008). Second, the current MGLAD-LASSO method assumes the model residual terms to follow a symmetric Laplace distribution. However, in reality, the distributions of some quantitative traits may show non-symmetric pattern, such as 'sugar content' in the sugar beet data set (see Figure 1). Quantile regression (Koenker, 2005), a generalization of the LAD regression, may provide better description of such non-symmetric distributed data. The quantile regression with LASSO penalty has been available for analyzing single response/trait (Li *et al.*, 2010), and potentially its multivariate version can be

developed. Third, our current method describes the correlation among traits in the marker effects part by using group LASSO penalty, but assumes the residuals to be independent. It has been shown that the modeling of residual covariance in the standard multivariate LASSO regression may have some advantages (Rothman et al., 2010). In principle, it is possible to include the residual covariance in our MGLAD-LASSO model as well, but the computation needs to be done by more sophisticated algorithms. This can be taken as one of our future research direction.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the research funding from Biocenter Oulu and Doctoral Program in Mathematics and Statistics in the University of Helsinki. We thank the three anonymous referees as well as the editor for their valuable comments on the manuscript.

- Alexander DH, Lange K (2011). Stability selection for genome-wide association. *Genet Epidemiol* **35**: 722–728.
- Baraldi AN, Enders CK (2010). An introduction to modern missing data analyses. *J Educ Psychol* **48**: 5–37.
- Bühlmann P, Kalisch M, Meier L (2014). High-dimensional statistics with a view toward applications in biology. *Annu Rev Stat Appl* **1**: 255–278.
- Coster A, Bastiaansen JWM, Calus MPL, Maliepaard C, Bink MCAM (2010). QTLMAS 2009: simulated dataset. *BMC Proc* **4**: S3.
- Estaghirou SBO, Ogutu JO, Piepho HP (2014). Influence of outliers on accuracy estimation in genomic prediction in plant breeding. *G3 (Bethesda)* **4**: 2317–2328.
- Efron B, Hastie T, Johnstone I, Tibshirani R (2004). Least angle regression. *Ann Stat* **32**: 407–451.
- Fernandes E, Pacheco A, Penha-Gonçalves C (2007). Mapping of quantitative trait loci using the skew-normal distribution. *J Zhejiang Univ Sci B* **8**: 792–801.
- Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**: 1.
- Goh L, Yap VB (2009). Effects of normalization on quantitative traits in association test. *BMC Bioinformatics* **10**: 415.
- Gómez E, Gómez-Villagas MA, Marín JM (1998). A multivariate generalization of the power exponential family of distributions. *Commun Stat Theory Methods* **27**: 589–600.
- Guo Z, Nelson JC (2008). Multiple-trait quantitative trait locus mapping with incomplete phenotypic data. *BMC Genet* **9**: 82.
- Hastie T, Tibshirani R, Friedman JH (2009). *The Elements of Statistical Learning*. Springer: New York, NY, USA.

APPENDIX

A. Multivariate LAD computation and hypothesis testing

We discuss about the estimation and hypothesis testing issues for the standard multivariate LAD problem without penalty: $\hat{\beta} = \arg \min_{\mathbf{B}} \frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i\|$. According to Oja (2010), the multivariate LAD estimates $\hat{\mathbf{B}}$ can be calculated by a fixed point algorithm with repeatedly updating the residuals $\mathbf{\varepsilon}_i$ ($i=1, \dots, n$) and regression coefficients $\mathbf{B}$ in the following two steps until convergence:

$$(\text{step1}) \mathbf{\varepsilon}_i \leftarrow \mathbf{y}_i - \mathbf{B}^T \mathbf{x}_i,$$

$$(\text{step2}) \mathbf{B} \leftarrow \mathbf{B} + \left[\sum_{i=1}^n (\|\mathbf{\varepsilon}_i\|^{-1} \mathbf{x}_i \mathbf{x}_i^T) \right]^{-1} \sum_{i=1}^n [x_i U(\mathbf{\varepsilon}_i)^T],$$

where $U(\mathbf{\varepsilon}_i) = \|\mathbf{\varepsilon}_i\|^{-1} \mathbf{\varepsilon}_i$, if $\mathbf{\varepsilon}_i \neq \mathbf{0}$ and $U(\mathbf{\varepsilon}_i) = \mathbf{0}$ if $\mathbf{\varepsilon}_i = \mathbf{0}$. In practice,

- Hayashi T, Iwata H (2013). A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformatics* **14**: 34.
- Heuven HCM, Janss LLG (2010). Bayesian multi-QTL mapping for growth curve parameters. *BMC Proc* **4**: S12.
- Li Q, Xi R, Lin N (2010). Bayesian regularized quantile regression. *Bayesian Anal* **5**: 533–556.
- Li Z, Sillanpää MJ (2012). Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theor Appl Genet* **125**: 419–435.
- Li Z, Sillanpää MJ (2013). A Bayesian nonparametric approach for mapping dynamic quantitative traits. *Genetics* **194**: 997–1016.
- Li Z, Hällingback HR, Abrahamsson S, Fries A, Gull BA, Sillanpää MJ, García-Gil MR (2014). Functional multi-locus QTL mapping of temporal trends in Scots pine wood traits. *G3 (Bethesda)* **4**: 2365–2379.
- Knott SA, Haley CS (2000). Multitrait least squares for quantitative trait loci detection. *Genetics* **156**: 899–911.
- Koenker R (2005). *Quantile Regression*. Cambridge University Press: New York, NY, USA.
- Koenker R, d'Orey V (1987). Computing regression quantiles. *Appl Stat* **36**: 383–393.
- Koenker R, d'Orey V (1994). Computing regression quantiles. *Appl Stat* **43**: 410–414.
- Meinshausen N, Bühlmann P (2010). Stability selection. *J R Stat Soc B* **72**: 417–473.
- Meinshausen N, Meier L, Bühlmann P (2009). P-values for high-dimensional regression. *J Am Stat Assoc* **104**: 1671–1681.
- Meuwissen TH, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.
- Möttönen J, Sillanpää MJ (2015). Robust variable selection and coefficient estimation in multivariate multiple regression using LAD-lasso. Accepted for publication in *Modern Multivariate and Robust Methods - Festschrift in Honour of Hannu Oja*. Springer.
- Nordhausen K, Oja H (2011). Multivariate L_1 methods: The Package MNM. *J Stat Softw* **43**: 1–28.
- Oja H (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Springer: New York, NY, USA.
- Rothman AJ, Levina E, Zhu J (2010). Sparse multivariate regression with covariance estimation. *J Comput Graph Stat* **19**: 947–962.
- Scheiner SM (2001). *MANOVA: multiple response variables and multispecies interactions. Design and Analysis of Ecological Experiments*, 2nd edn. Oxford University Press: Oxford, UK.
- Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc B* **58**: 267–288.
- Wang X, Piao Z, Wang B, Yang R, Luo Z (2009). Robust Bayesian mapping of quantitative trait loci using Student-t distribution for residual. *Theor Appl Genet* **118**: 609–617.
- Wu T, Lange K (2008). Coordinate descent algorithms for Lasso penalized regression. *Ann Appl Stat* **2**: 224–244.
- Würschum T, Maurer HP, Kraft T, Janssen G, Nilsson C, Reif JC (2011). Genome-wide association mapping of agronomic traits in sugar beet. *Theor Appl Genet* **123**: 1121–1131.
- Würschum T, Kraft T (2013). Cross-validation in association mapping and its relevance for the estimation of QTL parameters of complex traits. *Heredity* **112**: 463–468.
- Yang R, Wang X, Li J, Deng H (2009). Bayesian robust analysis for genetic architecture of quantitative traits. *Bioinformatics* **25**: 1033–1039.
- Yi J, Jannink JL (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* **192**: 1513–1522.
- Yuan M, Lin Y (2006). Model selection and estimation in regression with group variables. *J R Stat Soc B* **68**: 49–67.
- Xu S, Zhu D, Zhang Q (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc Natl Acad Sci USA* **111**: 12456–12461.

such an algorithm can be implemented by the function *mv.l1lm()* in the R-package 'MNM'.

Furthermore, we are interested in testing the null hypothesis: $\mathbf{B}_{j\bullet} = \mathbf{0}$, for each marker $j=1, \dots, p$. A score test statistic Q^2 for marker j can be derived from the following partition model

$$\mathbf{y}_i = \mathbf{B}_{j\bullet}^T \mathbf{x}_{ij} + \mathbf{B}_{(-j)\bullet}^T \mathbf{x}_{i(-j)} + \varepsilon_i, \varepsilon_i \stackrel{i.i.d.}{\sim} \text{MVN}(\mathbf{0}, \Sigma_0),$$

where $\mathbf{x}_{i(-j)}$ and $\mathbf{B}_{(-j)\bullet}$ represent the covariates and regression coefficients of all the markers except the marker j , respectively. The score test statistic can be calculated as

$$Q^2 = n \bullet \text{trace}(P_{\mathbf{X}_j} P_U),$$

where

$$\begin{aligned} P_{\hat{X}_2} &= \hat{X}_2(\hat{X}_2^T \hat{X}_2)^{-1} \hat{X}_2^T, \\ \hat{X}_2 &= (I_n - P_{X_1})X_2, \\ P_{X_1} &= X_1(X_1^T X_1)^{-1} X_1^T, \\ X_1 &= [x_{1(-j)}, \dots, x_{n(-j)}]^T, \\ X_2 &= [x_{1j}, \dots, x_{nj}]^T, \\ P_{\hat{U}} &= \hat{U}(\hat{U}^T \hat{U})^{-1} \hat{U}^T, \\ \hat{U} &= [\hat{u}_1, \dots, \hat{u}_2]^T \text{ and } \hat{u}_i = U(y_i - B_{(-j)}^T x_{i(-j)}). \end{aligned}$$

Under the null hypothesis, the score test statistic Q^2 has an approximate chi-square distribution with m (number of traits) degrees of freedom (Oja, 2010). The score test statistic and corresponding P -value can be calculated, for example, by the function `anova.mvllm()` in the R-package 'MNM'.

B. Stability selection

Stability selection is a subsampling-based method for quantifying the uncertainties of the estimates calculated by simultaneous parameter estimation and variable selection algorithms in linear regression such as LASSO. The selection procedure is as follows:

- i. In each of the K (we use $K=100$) replicates, a random subsample of size $n/2$ were drawn from the whole data.
- ii. MG-LASSO (or MGLAD-LASSO) was performed on the subsample of data with a fixed value of the tuning parameter λ , guaranteeing that approximately q markers are selected in each run.
- iii. The frequency $\hat{\pi}_j$ of each marker j ($j=1, \dots, p$) being selected over K replicates was calculated. We name $\hat{\pi}_j$ as the stability selection probability (SSP) of marker j .
- iv. A marker is claimed as a putative QTL, if its SSP is larger or equivalent to a threshold π_{thres} (see below).

In step (ii), the tuning parameter λ was pre-determined by performing MG-LASSO/MGLAD-LASSO on a small number of random subsamples of size $n/2$ searching a λ so that about q markers are selected for each subsample, and finally averaging these values of λ . A similar strategy for determining λ has been previously used in Alexander and Lange (2011).

According to Bühlmann *et al.* (2014), the threshold for QTL decisions can be calculated as

$$\pi_{\text{thres}} = \frac{1}{2} + \frac{q^2}{2p \cdot \text{efp}} \quad (q^2 \leq p \cdot \text{efp}),$$

where efp is the maximum number of false positives, which can be tolerated in the stability selection procedure. For the QTLMAS simulated data with $p=393$ markers, we specify $\text{efp}=1$ and $q=13$, so that $\pi_{\text{thres}} \approx 0.71$. Based on our empirical experiments, the optimal value of q is reached when $0.6 \leq \pi_{\text{thres}} \leq 0.8$. Setting π_{thres} to be too large or too small may reduce the chance to detect true QTLs.

For the sugar beet data with $p=493$, we then have $\text{efp}=1$ and $q=15$ result in $\pi_{\text{thres}} \approx 0.73$.

C. Multiple-split test

Similar to stability selection, the multiple-split test also adopts a subsampling strategy to construct uncertainty measures for marker effects estimated using any variable selection approaches. The multiple-split test aims to calculate a test statistic for each marker and identify putative QTLs by hypothesis testing, differing from stability selection which detects QTLs based on selection probabilities. Below is a summary of the multiple-split test procedure:

- i. In each of the K (we use $K=100$) replicates, the data were randomly divided into two parts with roughly equivalent number of individuals.
- ii. MG-LASSO (or MGLAD-LASSO) was performed on the first half of the data with a fixed value of the tuning parameter, guaranteeing that approximately q markers were selected in each run.
- iii. Normal least squares or LAD method was used on the second half of the data, with only the selected q (we use $q=25$ for both QTLMAS and sugar beet data sets) markers included. For those selected markers, test statistics and adjusted P -values after Bonferroni correction were also calculated. The score test statistic introduced in Appendix A was used for MGLAD-LASSO, and the Pillai's Trace test (Scheiner, 2001; in R, it can be calculated by the function `anova()`) was used for MG-LASSO. For the unselected markers, the P -values were just set to 1.
- iv. For each marker included in the analysis, the P -values calculated over K replicates were combined into one by using the procedure introduced in Meinshausen *et al.* (2009). A marker is claimed as a putative QTL, if its P -value is < 0.05 .

Supplementary Information accompanies this paper on Heredity website (<http://www.nature.com/hdy>)