BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 6 2011, pages 815–821
doi:10.1093/bioinformatics/btr006

Genetics and population analysis

Advance Access publication January 7, 2011

Robust linear regression methods in association studies

V. M. Lourenço[1,*], A. M. Pires[2] and M. Kirst[3]

[1]Department of Mathematics, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, [2]Department of Mathematics and CEMAT, Instituto Superior Técnico (TULisbon), 1049-001 Lisboa, Portugal and [3]School of Forest Resources and Conservation, Plant Molecular and Cellular Biology Program, Genetics Institute, University of Florida, PO Box 110410, Gainesville, FL 32611, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: It is well known that data deficiencies, such as coding/rounding errors, outliers or missing values, may lead to misleading results for many statistical methods. Robust statistical methods are designed to accommodate certain types of those deficiencies, allowing for reliable results under various conditions. We analyze the case of statistical tests to detect associations between genomic individual variations (SNP) and quantitative traits when deviations from the normality assumption are observed. We consider the classical analysis of variance tests for the parameters of the appropriate linear model and a robust version of those tests based on M-regression. We then compare their empirical power and level using simulated data with several degrees of contamination.

Results: Data normality is nothing but a mathematical convenience. In practice, experiments usually yield data with non-conforming observations. In the presence of this type of data, classical least squares statistical methods perform poorly, giving biased estimates, raising the number of spurious associations and often failing to detect true ones. We show through a simulation study and a real data example, that the robust methodology can be more powerful and thus more adequate for association studies than the classical approach.

Availability: The code of the robustified version of function `lmekin()` from the R package *kinship* is provided as Supplementary Material.

Contact: vmml@fct.unl.pt

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on September 15, 2010; revised on December 22, 2010; accepted on January 4, 2011

1 INTRODUCTION

Genetic association studies aim to identify genetic polymorphisms that cause phenotypic variation for a trait of interest, or that are in linkage disequilibrium with the causative genetic variant. We focus our analysis on biallelic genetic variants, such as single-nucleotide polymorphisms (SNPs). In this case, the unit of analysis can be regarded as a three-category variable. For instance, a SNP with alleles A (adenine) and G (guanine) has categories AA, AG and GG.

We are interested in using a number of genotyped SNPs in a gene, or region, to detect the genetic factors underlying a quantitative trait of interest that does not follow simple Mendelian patterns of inheritance. The most straightforward and still more favoured approach in association studies, though raising multiple testing problems (Nyholt, 2004), is to perform a single SNP test for every genotyped SNP via regression or analysis of variance (ANOVA) methods [Tao and Boulding (2003), used a linear model to test the association between SNPs in eight candidate genes and age-specific growth rate in the Artic charr; Martínez *et al*. (2007), used mixed linear models to test for association between 57 SNPs from 20 candidate genes and some wood properties in *Pinus taeda*; Weber *et al*. (2007, 2008) used a mixed random effects linear model to test for the association between a collection of SNPs and some Teosinte traits; Moe *et al*. (2009), used a linear model to test for the association between 151 SNPs from 57 candidate genes and several traits of boar). Though the single SNP approach may be considered if we are looking for a single causal variant, it is not very efficient when the SNPs have limited LD with that causal variant, meaning smaller power. Moreover, quantitative traits are usually controlled by several and sometimes many genes. Thus, a joint analysis of SNPs may be more adequate, being much more informative than single-SNP analysis (Jannot *et al*., 2003). However, it also may lose power due to the usually large number of degrees of freedom involved. Ideally, one should make use of the information provided by multiple SNPs, capturing as much of the genetic variance as possible, without raising the degrees of freedom too much (Bureau *et al*., 2005; Chapman and Whittaker, 2008; Kwee *et al*., 2008; Li *et al*., 2009; Wang and Elston, 2006; Xiang *et al*., 2009) and thus not compromising power. Note that the joint analysis of SNPs (multiple-SNP approach) can only be applied to situations where the number of explanatory variables is much smaller than the number of individuals, therefore implying that in a genome-wide association study context a preliminary step of dimension reduction is necessary.

There is an extensive literature on how two specific data problems—LD and population structure (PS)—may affect both the power to detect true associations as well as the number of false positives, therefore distorting the conclusions when testing for association between a quantitative trait and a set of candidate SNPs in a population-based study (Cardon and Palmer, 2003; Freedman *et al*., 2004; Pritchard *et al*., 2000a). We also find in literature many methods for overcoming these problems (Bacanu *et al*., 2002; Carlson *et al*., 2004; Devlin and Roeder, 1999; Li *et al*., 2008; Malo *et al*., 2008; Price *et al*., 2006; Pritchard *et al*., 2000b; Yu *et al*., 2006). Another frequent data problem, which may have the same sort of undesirable effects, is non-normality and/or presence of outliers in the phenotypic data. This problem is far less studied

*To whom correspondence should be addressed.

than LD or PS. For instance, the review paper by Balding, 2006, treats non-normality in one sentence where the only mentioned remedy is a transformation of the original trait values. However, transformation may not be sufficient to solve all the problems caused by non-normality (Pires and Rodrigues, 2007) and frequently raises interpretation issues.

For many real-life datasets, the distribution of the quantitative traits is not normal and often shows heavy tails, which in turn tend to make regular observations look like outliers. This is mainly the reason why non-normality and outlier presence are usually associated. In such scenarios, the classical approach, whose likelihood-based inference leans on the normality assumption and is known to be non-robust to small model deviations (Huber, 1964; Tukey, 1960), may be inappropriate, having low statistical efficiency if the tails are symmetric and large bias if the tails are asymmetric. This leads to tests with unreliable level and low power and to confidence intervals with also unreliable level and large expected interval length. We emphasize the fact that, contrary to some statements in the literature dating back to Box, 1953, the ANOVA F-test is not robust against non-normality (de Haan *et al.*, 2009; Ronchetti, 1987; Schrader and Hettmansperger, 1980). Robust methods are designed to be resistant to influent factors such as outlying observations, non-normality and other model misspecifications (Daszykowski *et al.*, 2007; Huber, 1972; Maronna *et al.*, 2006). Moreover, if the model verifies the classical assumptions, robust methods provide results close to the classical ones. Therefore, the use of robust methods has been advocated for inference in the linear and mixed linear model setup (Copt and Feser, 2006; Copt and Heritier, 2007; Daszykowski *et al.*, 2007; Pires and Rodrigues, 2007). Also, we already see some applications in genetic association studies (Gudbjartsson *et al.*, 2010; Tan *et al.*, 2010; Xu *et al.*, 2010), which show increasing concern with the violation of model assumptions and growing interest in using methods that are capable of coping with them.

In this article, we propose that Huber M-estimation is used together with adapted Wald-type tests (Section 2) to assess trait/SNP associations. The performance of the proposed approach, is compared with both the classical and two non-parametric methods in terms of type I error rate and power in a simulation study under several contamination settings (Sections 3 and 4). Finally, we present a real data example (Section 5) and discuss the results obtained (Section 6).

2 BACKGROUND STATISTICAL METHODS

We describe the general multiple linear regression model as

$$Y_i=\beta_0+\beta_1 X_{i1}+\ldots+\beta_{p-1}X_{ip-1}+\varepsilon_i,\ \ i=1,\ldots,n \qquad (1)$$

where $n>p$ is the number of observed individuals. This model appropriately rewrites to $Y=X\beta+\varepsilon$, where $Y=(Y_1,\ldots,Y_n)^T$ is the $(n\times 1)$ vector of the response variable, X is the $(n\times p)$ design matrix, $\beta=(\beta_0,\ldots,\beta_{p-1})^T$ are the unknown parameters and $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)^T$ is a vector of non-observable independent errors with expectation $E(\varepsilon)=\mathbf{0}$ and covariance matrix $\text{var}(\varepsilon)=\sigma^2 I_n$.

The least squares (LS) estimate of β is obtained by minimizing the residual sum of squares,

$$\widehat{\beta}_{\text{LS}}=\underset{\beta}{\arg\min}\sum_{i=1}^{n}(Y_i-X_{i\bullet}\beta)^2, \qquad (2)$$

where $X_{i\bullet}=(1,X_{i1},\ldots,X_{ip-1})$. If X has rank $p\leq n$, the result is $\widehat{\beta}_{\text{LS}}=(X^TX)^{-1}X^TY$, with covariance matrix $\widehat{\Sigma}_{\beta_{\text{LS}}}=\sigma^2(X^TX)^{-1}$. In the classical approach, we have $\varepsilon\sim N(\mathbf{0},\sigma^2 I_n)$ and $\widehat{\beta}_{\text{LS}}$ is also the maximum likelihood estimate (MLE) of β.

A general linear hypothesis concerning β is of the form H_0: $H\beta=\mathbf{0}$, where H is a known $q\times p$ matrix with $q\leq p$. The general test for testing this hypothesis is to reject H_0, at the level α, if $F\geq F_\alpha$, where $P_{H_0}(F\geq F_\alpha)=\alpha$ and

$$F=\frac{(Y-X\widehat{\widehat{\beta}})^T(Y-X\widehat{\widehat{\beta}})-(Y-X\widehat{\beta})^T(Y-X\widehat{\beta})}{(Y-X\widehat{\beta})^T(Y-X\widehat{\beta})}. \qquad (3)$$

$\widehat{\beta}$ and $\widehat{\widehat{\beta}}$ are the unrestricted and restricted MLE of β, respectively. We also know that, under H_0, $(n-p)F/q\sim F_{q,n-p}$.

We are interested in the following two testing situations: (i) $H_0:\beta_1=\cdots=\beta_{p-1}=0\equiv H\beta=\mathbf{0}$ where $H=[\mathbf{0}\,I_{p-1}]$; (ii) $\{H_{0k}:\beta_k=0\}_{k=1,\ldots,p-1}\equiv\{H\beta=0\}$ where now H is a $1\times p$ vector of nulls with 1 at position $k+1$. In (i), $q=p-1$, $\widehat{\widehat{\beta}}_1=\cdots=\widehat{\widehat{\beta}}_{p-1}=0$, $\widehat{\widehat{\beta}}_0=\overline{Y}$, thus F rewrites to

$$F=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2}-1=\frac{\text{SST}}{\text{SSE}}-1 \qquad (4)$$

and, under H_0, $(n-p)F/(p-1)\sim F_{p-1,n-p}$. In (ii), for each H_{0k} test, we have $q=1$ and so, again under H_0, $(n-p)F\sim F_{1,n-p}$. In practice, for each biallelic SNPs, there are two dummy variables in the regression model, that is, $q=2$ and $(n-p)F/2\sim F_{2,n-p}$ (under H_0).

As to the robust approach, the normality condition on the error distribution is relaxed to a quasi-normality condition and the estimators are obtained by methods other than ML. There are many robust regression methods in the literature but, since in the context of genetic association studies there are no outliers in the explanatory variables, we can restrict our attention to M-estimators, which have good computational and efficiency properties (Maronna *et al.*, 2006). In the M-regression approach, the estimates are the solutions to the following minimization problem,

$$\widehat{\beta}_R=\underset{\beta}{\arg\min}\sum_{i=1}^{n}\rho\Big(\frac{Y_i-X_{i\bullet}\beta}{\widehat{\sigma}}\Big)=\underset{\beta}{\arg\min}\sum_{i=1}^{n}\rho\Big(\frac{r_i(\beta)}{\widehat{\sigma}}\Big) \qquad (5)$$

where ρ is an appropriate function and $\hat{\sigma}$ is a robust estimate of σ. It is easy to verify that when $\rho(x)=x^2$, we have the LS/ML situation described above. Differentiating (5) for every β_j, and equating to zero we get the p equations

$$\sum_{i=1}^{n}\psi\Big(\frac{r_i(\beta_j)}{\widehat{\sigma}}\Big)X_{ij}=0,\ j=0,\ldots,p-1. \qquad (6)$$

Although (5) and (6) are not always equivalent, (6) is useful in the search of solutions to (5). Moreover, considering the weights

$$W_i=\psi\left(\frac{r_i(\beta_j)}{\widehat{\sigma}}\right)\Big/ r_i(\beta_j),\ \ i=1,\ldots,n \qquad (7)$$

leads to $\widehat{\beta}_R=(X^TWX)^{-1}X^TWY$, where W is a diagonal matrix with elements W_i. This shows that (6) can be solved by iteratively reweighted least squares (IRWLS).

Choosing a robust estimator within the class of M-estimators is not always an easy task. We have considered the ρ function proposed by Huber (1964) since it is known that this function leads to efficient estimators under general conditions and provides a unique solution to (6):

$$\rho(x)=\begin{cases}x^2/2, & \text{if } |x|\leq b\\ b(|x|-b/2), & \text{if } |x|>b\end{cases}. \qquad (8)$$

Other choices are available in the literature, e.g. Tukey's biweight, but do not guarantee a unique solution to (6) and one needs good initial estimates of the parameters to assure the convergence of the corresponding algorithm to the optimal solution. Using Huber's ρ, the resulting M-estimator of β is efficient (for both normal and non-normal data) and it is robust against outliers in the response variable, which is precisely the situation we may find in practice. Library MASS in R has a model-fitting function `rlm()` in the conditions described above. By default, it uses ρ-Huber with tuning constant $b=1.345$ and both β and the scale parameter σ are estimated by the IRWLS procedure with initial estimates of β and σ given by the LS estimate and the rescaled MAD, respectively.

As to the robust tests for the general linear hypothesis, taking $\gamma = H\beta$, we considered the robust Wald-type statistic

$$T_W = \frac{\widehat{\gamma}^T \widehat{\Sigma}_{\gamma}^{-1} \widehat{\gamma}}{q} \underset{H_0}{\sim} F_{q,n-p} \tag{9}$$

where $\widehat{\Sigma}_{\gamma}$ is an estimate of the covariance matrix of γ. Since $\widehat{\Sigma}_{\beta_R} = \widehat{\upsilon}(X^T X)^{-1}$, with $\widehat{\upsilon} = \widehat{\sigma}^2 \dfrac{\mathrm{ave}_i(\psi(r_i/\widehat{\sigma})^2)}{(\mathrm{ave}_i(\psi'(r_i/\widehat{\sigma})))^2} \times \dfrac{n}{n-p}$, it follows that $\widehat{\Sigma}_{\gamma} = \widehat{\upsilon} H (X^T X)^{-1} H^T$.

The two non-parametric methods selected for comparison were rank transform (RT; Conover and Iman, 1981), which was recently used in genetics (de Haan *et al*., 2009) and a Wilcoxon based (WIL; McKean *et al*., 2009), used in QTL analysis (Zou *et al*., 2003).

3 SIMULATION STUDY

3.1 The simulation model

We simulate N biallelic genes on one pair of chromosomes of an F2 population. We start by simulating the genotype for the first gene. Afterwards, each new gene genotype is simulated based on the previous gene genotype, assuming no crossover interference and a recombination fraction r between both genes, randomly taken from the uniform distribution, $U(0,0.5)$ (see Liu, 1997, for further details on the simulation of an F2 population genotype).

If a quantitative trait is assumed to be controlled by a number N of genes, without epistatic interactions, it can be simulated by

$$y_j = \mu + \sum_{i=1}^{N} \left(a_i x_{(2i-1)j} + d_i x_{(2i)j} \right) + \varepsilon_j, \tag{10}$$

where y_j is the trait value for the j-th individual in the population, $(x_{(2i-1)j}, x_{(2i)j})$ are the dummy variables associated to the additive and dominance effects for the N genes, respectively, (a_i, d_i) are the additive and dominance effects for each gene, μ is the overall mean for the trait and ε_j is the random error for the j-th individual, $j = 1,\ldots,n$. The dummy variables are coded $(x_{(2i-1)j}, x_{(2i)j}) = (1,0)$, $(x_{(2i-1)j}, x_{(2i)j}) = (0,1)$, and $(x_{(2i-1)j}, x_{(2i)j}) = (-1,0)$, for genotypes *AA*, *Aa* and *aa*, respectively. In order to use (10), we need to specify some parameters underlying the simulation model. These are as follows:

Heritability (*broad sense*): heritability is the proportion of phenotypic variation in a population that is attributable to genetic variation among individuals. It therefore quantifies the importance of the genetic effects to the trait value and is defined as the ratio of genotypic to phenotypic variance

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_e^2} \tag{11}$$

where σ_G^2 and σ_e^2 are the variance components associated with the genetic effects and the residual error, which may include undetected genetic effects, environmental effects and the random effects. From Equation (11) and for a certain value of H^2 ($\neq 0,1$) we have the relations

$$\sigma_G^2 = \frac{H^2 \sigma_e^2}{1-H^2} \quad \text{and} \quad \sigma_e^2 = \frac{(1-H^2)\sigma_G^2}{H^2} \quad . \tag{12}$$

Genetic, additive and dominance variances: the genetic variance can be decomposed as $\sigma_G^2 = \sigma_A^2 + \sigma_D^2$, where σ_A^2 and σ_D^2 are the additive and dominance variances, respectively. Under model (10), the additive and dominance variances between genes are given by (Wu *et al*., 2007)

$$\begin{aligned} \sigma_A^2 &= \tfrac{1}{2} \textstyle\sum_{i=1}^{N} a_i^2 + \tfrac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} (1-2r_{ij}) a_i a_j \\ \sigma_D^2 &= \tfrac{1}{4} \textstyle\sum_{i=1}^{N} d_i^2 + \tfrac{1}{4} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} (1-2r_{ij})^2 d_i d_j \end{aligned} \tag{13}$$

where r_{ij} represents the recombination fraction between genes i and j. From the genotype data simulation, the values of $r_1,\ldots,r_{N-1}$ coincide with $r_{12}, r_{23}, \ldots, r_{(N-1)N}$. We also have $r_{ii} = 0 \ \forall_i$. In order to obtain the remaining values of the recombination fractions between genes, and since recombination fractions are not additive, $r_1,\ldots,r_{N-1}$ can be converted to map distances $d_1^*,\ldots,d_{N-1}^*$ via Kosambi's or Haldane's map function. Map distances are additive so, with the values of $d_1^*,\ldots,d_{N-1}^*$, the distances between all genes can be calculated, and through the inverse process the recombination fractions between all genes can thus be obtained. The relative importance of the additive and dominance variances can be quantified by their ratio and so we may consider, for simulation purposes, $\sigma_D^2/\sigma_A^2 = t$ and write

$$\sigma_A^2 = \sigma_G^2 \frac{1}{t+1} \text{ and } \sigma_D^2 = \sigma_G^2 \frac{t}{t+1}. \tag{14}$$

Additive and dominance effects: Assuming the relative sizes of the effects to be given by $k_i = a_i/a_1 = d_i/d_1$, $i = 1,\ldots,N$, the additive and dominance effects can be obtained from (13) with a little algebra:

$$\begin{cases} a_1 = \sqrt{\sigma_A^2 / \left(\frac{1}{2} \sum_{i=1}^{N} k_i^2 + \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} (1-2r_{ij}) k_i k_j \right)} \\ a_i = a_1 k_i, \quad i = 2,\ldots,N \\ d_1 = \sqrt{\sigma_D^2 / \left(\frac{1}{4} \sum_{i=1}^{N} k_i^2 + \frac{1}{4} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} (1-2r_{ij})^2 k_i k_j \right)} \\ d_i = d_1 k_i, \quad i = 2,\ldots,N \end{cases} \quad . \tag{15}$$

For a fixed value of H and $\sigma_G^2 = 1$, and using Equation (12), we have $\varepsilon_j = z\sigma_e = z\sqrt{(1-H^2)/H^2}$ for each individual $j = 1,\ldots,n$, where z is a random observation from the $N(0,1)$ distribution. Finally, for a fixed value of t, the additive and dominance variances can be calculated from (14) and from there all the additive and dominance effects (15). The quantitative trait is then simulated for the n individuals.

3.2 Simulation settings

The variation of quantitative traits is only shortly explained by genetic factors and thus we considered a trait heritability of 30%, i.e. $H^2 = 0.3$. Other parameters fixed were $n = 500$, $\mu = 50$ and $t = 0.6$. Additionally, for the purpose of the simulation, we considered independent SNPs, that is $r_{ij} = 0.5 \ \forall_{i \neq j}$. This assures that the SNPs are in Hardy–Weinberg equilibrium (HWE) and that there is no LD between pairs of SNPs. Also, we considered the relative additive and dominance effects of the genes $k_i \sim U(0.8,1)$, $i = 2,\ldots,N$ so that in this way every gene in the model would have an important, though unequal, contribution to the trait value. As to the number of SNPs in the model, we took $N = 2,3,4,5$ and 10. A percentage, 2, 5 and 10%, of data contamination (outliers) was also considered. The contamination was generated from a normal distribution $N(\mu_c, \sigma_c^2)$, where σ_c was obtained from a uniform distribution $U(1,5)$ and μ_c from $U(80,90)$, $U(60,70)$ and $U(55,60)$, corresponding, respectively, to *gross*, *intermediate* or *smooth* contamination. Association tests were run 10 000 times.

In order to compare the modelling approaches, one needs first to investigate how they control the family wise error rate (FWER) at a pre-specified level. On the other hand, a better performance of one over another means POWER. Hence, we have the following two testing settings:

Under the null hypothesis: Under the null hypothesis, we looked at three distinct situations: (i) traits were simulated independently from the SNPs genotypes out of a $N(50,1)$ distribution; (ii) traits were simulated independently from the SNPs genotypes out of a $N(50,1)$ and a percentage of outliers (2, 5 and 10%) was then introduced in the trait values; (iii) a percentage of the traits (98, 95 and 90%) were simulated independently from the SNPs genotypes, that is under H_0, but using the normal contaminant distributions, whereas the remaining traits (2, 5 and 10%) were generated from model (10), that is under H_1.

Under the alternative hypothesis: Under the alternative hypothesis two situations were considered: (i) the traits were simulated according to model (10); (ii) the traits were simulated from model (10) and a percentage (2, 5 and 10%) of outliers was introduced.

Table 1. FWER (in %) obtained from simulation under H_0 with a percentage of contaminated traits—(iii) smooth contamination

Smooth (%)	*Model*	2 SNP	3 SNP	4 SNP	5 SNP	10 SNP
2	Classical	5	5	5	5	5
	Robust	5	5	5	5	5
5	Classical	6	6	6	6	5
	Robust	5	5	5	5	5
10	Classical	8	7	7	7	6
	Robust	5	5	5	5	5

4 SIMULATION RESULTS

We adjusted in R the additive model described in (1) via the usual linear regression (`lm()`) and the robust linear regression (`rlm()`) with the M-Huber estimator, in the two simulation settings described. Function `lm()` was also used for the rank transform method, i.e. the original quantitative traits were replaced by their ranks. As to the Wilcoxon methodology, we used the R function `wwest()` from McKean *et al.* (2009). We then tested for association using the ANOVA table for the usual linear regression model and rank transform method, and robust Wald-type tests for the robust model as described in Section 2. Tests for Wilcoxon were obtained from function `wwest()` also via Wald-type tests. SNPs were always tested in a multiple regression framework. We must also underline that due to the computational effort of the `wwest()` function, simulations for the Wilcoxon approach ran only 1000 times instead of the 10 000.

We concluded that under the null without contamination, the distribution of the P-values is, in all cases, approximately $U(0,1)$ and thus all methods are comparable, allowing for the use of the same multiple testing corrections. We used Bonferroni correction in order to control the FWER at the 5% level.

In the (ii) under H_0, there were no surprises, i.e. all methods were able to control the FWER at the 5% level, whatever the contamination, gross, intermediate or smooth. However, in the (iii) situation under H_0, we noticed that in the smooth contamination setting, the FWER of the classic approach surpasses 5% most of the times, while the robust approach always keeps it around that threshold (Table 1). This tendency of the classic approach towards inflated type I error rates was also observed in the intermediate contamination setting and accompanied by the Wilcoxon approach also in the smooth and intermediate contamination settings (see results in the Supplementary Material). As to the rank transform approach, it only failed to control for the FWER at the desired level once.

Under H_1 (Table 2 and Supplementary Material), all methods show very good power to detect association between the SNPs and the quantitative trait, in the following order: Cls>WIL>Rob>RT and with relatively small differences. Even using Bonferroni correction, we had over 99% power in all methods, up to 5 SNPs, and still over 76% power when we took the model with 10 SNPs (0% contamination). With the introduction of contamination, as expected, the power of all methods decreases as the contamination level increases. However, the robust method shows much higher power to detect associations than the classic method and higher power than the rank transform approach, being neck to neck with the Wilcoxon approach. At the worst scenario, 10 SNPs in association with the trait and 10% gross outliers, the robust method, with rank transform and Wilcoxon close behind, has a power over 52% to detect those associations, while the usual model stays under 1%. Even in the smooth contamination case, the power of the classical method is only 23.5% versus over 48% for all the other approaches. Moreover, we must stress out that there are relevant power losses even in cases where the residual deviations from normality are not evident. See, for example, in Table 2 the 10 SNP model with 2% smooth contamination, where there is 13% power loss from the classic approach relative to the robust one but whose residuals do not look different from normal (see Q-Q plots in the Supplementary Material). If we now analyse Table 3 and the correspondent table from the Supplememtary Material, we see that the robust method detects in general more SNPs than the classical and rank transform procedures. Although that difference may not look substantial in the smooth contamination setting, it is quite evident in the 10 SNP simulation for 5 and 10% gross contamination. If compared to the Wilcoxon approach, in the 10 SNP simulation Wilcoxon comes off better than the robust approach but with a maximum difference of only 0.08.

Table 2. POWER (in %) obtained from simulation under H_1 without and with a percentage of grossly (G) and smoothly (S) contaminated traits

Contamination (%)	*Model*	2 SNP	3 SNP	4 SNP	5 SNP	10 SNP
None	Classical	100	100	99.9	99.7	81.2
	Robust	100	100	99.9	99.7	78.2
2 (G)	Classical	62.6	41.1	26.6	15.2	3.4
	Robust	100	100	99.9	99.3	75.0
5 (G)	Classical	25.1	12.4	8.5	4.7	1.2
	Robust	100	100	99.9	98.6	61.3
10 (G)	Classical	11.9	5.6	4.0	2.3	0.8
	Robust	100	99.7	99.1	92.6	52.6
2 (S)	Classical	100	99.9	99.3	95.9	62.2
	Robust	100	100	99.9	99.5	75.5
5 (S)	Classical	99.6	96.8	93.1	82.9	37.1
	Robust	100	99.9	99.8	98.5	63.3
10 (S)	Classical	95.6	82.8	75.2	55.4	23.5
	Robust	100	99.8	99.2	93.4	55.3

Truncated minimum power observed to detect every SNP in the simulation model.

Table 3. Truncated, average number of SNPs detected in the simulation under H_1 without and with a percentage of grossly (G) and smoothly (S) contaminated traits

Contamination (%)	*Model*	2 SNP	3 SNP	4 SNP	5 SNP	10 SNP
None	Classical	2	3	3.99	4.99	9.01
	Robust	2	3	3.99	4.99	8.80
2% (G)	Classical	1.22	1.04	0.84	0.61	0.17
	Robust	2	3	3.99	4.98	8.39
5% (G)	Classical	0.45	0.28	0.18	0.12	0.03
	Robust	2	3	3.99	4.96	7.76
10% (G)	Classical	0.20	0.10	0.06	0.04	0.01
	Robust	2	2.99	3.97	4.85	6.30
2% (S)	Classical	1.99	2.99	3.98	4.89	7.21
	Robust	2	3	3.99	4.99	8.42
5% (S)	Classical	1.99	2.95	3.78	4.41	4.94
	Robust	2	2.99	3.99	4.97	7.89
10% (S)	Classical	1.94	2.65	3.09	3.25	2.65
	Robust	2	2.99	3.98	4.87	6.61

Table 4 and the correspondent table from the Supplementary Material show the results for the 10 SNP simulation scenario. Note that the global robust test, as well as the rank transform and Wilcoxon global tests, keep a 100% power in all simulation settings, whereas the classical power falls down to 15.21% at the 10% gross contamination setting.

5 EXAMPLE

As an example of application, we downloaded the data of Weber *et al.* (2008) (see Zhao *et al.*, 2006, and *www.panzea.org*), with respect to the quantitative trait FERL (length of the female and hermaphroditic portions of the basal-most ear on the lateral branch), getting information on 61 SNPs from 22

Table 4. POWER (%) obtained from simulation under H_1, without contamination and with a percentage of contaminated traits for the global test of association in the 10 SNP simulation model

Contamination (%)	Model	Gross	Intermediate	Smooth
0	Classical	100	100	100
	Robust	100	100	100
2	Classical	76.71	100	100
	Robust	100	100	100
5	Classical	32.48	96.54	100
	Robust	100	100	100
10	Classical	15.21	77.24	99.69
	Robust	100	100	100

candidate genes, chosen based on their possible effects on the trait under study, given their known mutant phenotype in maize or other plants.

In that paper, Weber studied the association between these SNPs and this Teosinte (maize's wild ancestor) trait, among others, by adjusting the mixed linear model (as presented by Yu *et al.*, 2006),

$$y=Pv+S\alpha+Iu+e, \qquad (16)$$

where y is the vector of phenotypic values, v is a vector of fixed effects regarding population structure (inferred via PowerMarker, Liu and Muse, 2005), α is the fixed effect for the candidate SNPs, u is a vector of random effects relative to recent coancestry, e a vector of residuals, P is a matrix of the 10 significant principal components [as suggested by Price *et al.* (2006) and discussed by Zhao *et al.* (2007)], S a vector of the SNPs genotypes and I an identity matrix. The structure assumed for the variances is as follows: $\text{var}(u)=2KV_g$ and $\text{var}(e)=IV_R$, where K is the Kinship matrix, which quantifies the proportion of shared alleles, $V_g=\theta$ is the genetic variance and $V_R=\sigma^2$ is the residual variance. The pertinence of model (16) in this study is justified by the systematic sources for spurious associations found under the simple model,

$$y=S\alpha+e, \qquad (17)$$

non-uniformity of the P-values and high type I error rates (Fig. 1) observed while testing 498 randomly genotyped SNPs one at a time (Weber *et al.*, 2008). The adequacy of the full model to test under the null hypothesis is also clear, with only ≈5.4% by chance significant associations in both the classical and the robust methodologies.

Here, we configured SNP as a numerical covariate. However, when we are not under the null, the codification of SNP as factor is more adequate in order to capture eventual dominance effects, unless of course the effects in the model are only additive, which may not always be the case. In these circumstances, we will therein consider SNP coded as factor. We started by rewriting model (16) as

$$Y^*=X^*\beta+\varepsilon^* \qquad (18)$$

where $Y^*=y$, $X^*=Pv+S\alpha$ and $\varepsilon^*=Iu+e$, with $\text{var}(\varepsilon^*)=\sigma^2V$, $V=K\frac{\theta}{\sigma^2}+I=K\theta_0+I$ (i.e. we consider $\theta_0=\theta/\sigma^2$). When V is known, and is positive semi-definite, taking its Cholesky decomposition, leads to

$$Y=X\beta+\varepsilon \qquad (19)$$

where $Y=\text{chol}(V)^{-1}Y^*$, $X=\text{chol}(V)^{-1}X^*$, $\varepsilon=\text{chol}(V)^{-1}\varepsilon^*$ and now $\text{var}(\varepsilon)=\sigma^2I$. The ML estimates of β and σ^2 may then easily be obtained.

Usually, the matrix V is unknown and θ_0 is estimated by ML, restricted maximum likelihood (REML) or some other method. We should note that Weber used the SAS Proc Mixed routine with REML, and we used ML from the R function `lmekin()` in package *kinship*, which may justify the slight differences between both results. Once an estimate of V, $\widehat{V}$, has been obtained, it is used in (19) so that the estimates of β and σ^2 can be calculated.

Inference for the fixed-effects terms, i.e. the tests $H_0: H\beta=\mathbf{0}$ of $q\leq p$ fixed-effects, were conducted by the usual F statistic with degrees of freedom estimated by the Satterthwaite approximation (as in Weber *et al.*, 2008).

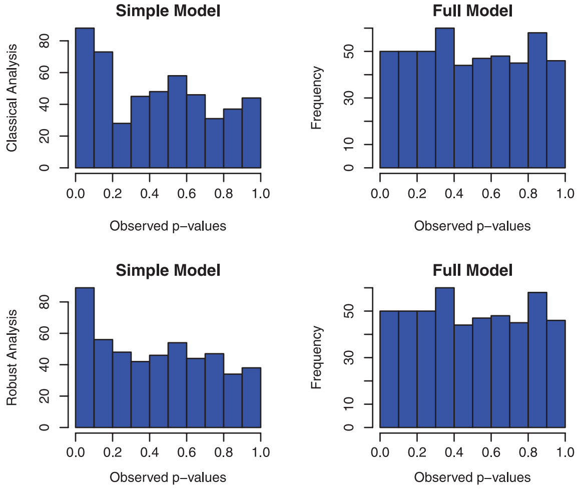

Fig. 1. Histograms of the 498 P-values for both Simple and Full models, in the classical and robust analysis.

Table 5. Significant SNPs detected at the 5% level and with FDR under 10% in the single SNP analysis

	SNP	R^2	$2a/\sigma_P^2$	d/a	$2a$	d	P-value	q-value
	zagl1.1	0.021	0.176	−2.55	1.90	−2.43	0.0057	0.0212
C	PZD00073.5	0.015	0.839	−1.01	9.05	−4.57	0.0069	0.0212
L	PZD00006.1	0.006	0.301	−0.11	3.26	−0.17	0.0083	0.0212
A	PZD00022.3	0.021	0.436	0.87	4.72	2.06	0.0102	0.0212
S	PZD00073.8	0.017	0.934	−0.94	10.20	−4.82	0.0112	0.0212
S	te1.3	0.010	0.422	1.01	4.55	2.29	0.0208	0.0312
I	ba1.9	0.027	0.984	−0.86	10.60	−4.57	0.0267	0.0344
C	PZB00049.7	0.016	0.094	5.08	1.01	2.56	0.0336	0.0378
	zagl1.6	0.008	0.109	3.07	1.18	1.82	0.0499	0.0499
	zagl1.1	0.020	0.139	−3.15	1.50	−2.37	0.0067	0.0169
R	PZD00073.5	0.014	0.834	−1.03	9.00	−4.65	0.0090	0.0169
O	PZD00006.1	0.003	0.279	0.06	3.03	0.09	0.0083	0.0169
B	PZD00022.3	0.019	0.430	0.61	4.66	1.42	0.0092	0.0169
U	PZD00073.8	0.014	0.936	−1.05	10.22	−5.37	0.0106	0.0169
S	te1.3	0.004	0.273	0.96	2.95	1.41	0.0442	0.0201
T	ba1.9	0.028	0.985	−0.79	10.61	−4.21	0.0151	0.0450
	PZD00049.7	0.013	0.052	8.17	0.56	2.72	0.0450	0.0450

5.1 Single SNP analysis

We performed the classical analysis described above for each of the 61 candidate SNPs with routine `lmekin()` from the R package *kinship*. The robust analysis was performed with a robustified version of this instruction (Supplementary Material). This single SNP analysis allowed for the detection of nine associations in the classic analysis and eight in the robust analysis at the 5% level (i.e. $P\leq 0.05$), all associations being significant after correction for multiple testing via false discovery rate (FDR) ($q\leq 0.1$), Table 5. In both analysis, we have the six SNPs identified by Weber *et al.* (2008). Moreover, with the exception of zagl1.6, all other detected SNPs are common between approaches. Also, (i) from the identified SNPs, SNPs PZD00073.5 and PZD00073.8 are not independent since they are in high LD ($r^2=0.6770$); all other observed pairwise LD is low ($r^2\leq 0.02$)—we used `LD()` instruction from the R package *genetics* for the calculations; (ii) only SNP PZD00006.1 shows an additive mode of inheritance; zagl1.1, zagl1.6 and PZB00049.7 show overdominance and the remaining SNPs show partial or complete

Table 6. Multiple SNP analysis results of the SNPs declared significant in the single SNP analysis

SNP	Classic		Robust		Classic[a]	
	p-value	q-value	p-value	q-value	p-value	q-value
zagl1.1	0.1510	0.1699	0.1656	0.2208	0.4125	0.4366
PZD00073.5	0.1075	0.1612	0.2432	0.2780	0.0607	0.1092
PZD00073.8	0.5895	0.5895	0.6252	0.6252	0.2994	0.3849
PZD00006.1	**0.0008**	**0.0076**	**0.00 004**	**0.0003**	**0.0004**	**0.0032**
PZD00022.3	**0.0042**	**0.0193**	**0.0007**	**0.0029**	**0.0040**	**0.0178**
te1.3	0.1356	0.1699	0.0910	0.1456	0.4366	0.4366
ba1.9	0.0138	**0.0413**	**0.0055**	**0.0147**	0.0090	**0.0272**
zagl1.6	0.0327	**0.0736**	–	–	0.1849	0.2774
PZB00049.7	0.0535	**0.0962**	0.0243	**0.0487**	0.0537	0.1092

[a] After removing outliers.
P-values are raw.

dominance; (iii) from the R^2 values, we acknowledge that all individual effects are small, ranging from 0.8% to 2.7% in the classic analysis and from 0.3% to 2.8% in the robust analysis, therefore explaining only a small fraction of the phenotypic variation. This could be because the marker assayed is not the causative site but is in LD with the causative site giving an underestimate R^2 of the real effect, the trait may have low heritability or the associations may be due to alleles of small effect.

We additionally performed the Shapiro–Francia (SF) normality test on both the residuals from the classic and robust approaches and observed that they all failed the normality assumption showing heavy tails, specially on the left side of the distribution. Although it is not a pre-requisite in the robust analysis, residual normality is one of the classical assumptions. This violation indicates that either the classical analysis normality assumption we have made is not realistic or the model adjusted is not good. Either way, one should take care with false association detection (possibly the case of SNP zagl1.6 detected in the classical methodology but not in the robust) and possible reduction in power.

5.2 Multiple SNP analysis

We further investigated the previous results in a joint analysis of the SNPs detected above. This multiple SNP analysis (Table 6), after correcting for multiple testing with the conservative Bonferroni correction (bold *P*-values), left us with only two significant SNPs in the classic analysis (PZD0006.1 and PZD00022.3) and three significant SNPs in the robust analysis (PZD0006.1, PZD00022.3 and *ba*1.9). If we consider the FDR at level 10% (bold *q*-values), as in Weber *et al.* (2008), then both methods detect SNPs PZD00006.1, PZD00022.3, ba1.9 and PZB00049.7. The classic analysis additionally detects SNP zagl1.6. In order to evaluate the models adequacy, we plotted the conditional residuals plots (Fig. 2), and performed the SF test of normality, verifying, again the non-normality of both the classical ($p \simeq 0.034$) and the robust residuals ($p \simeq 0.003$) at the 5% level.

The plots in Figure 2 also show heavier tails on the left of the distributions (normal QQ-plots) and the presence of possible outlying observations (conditional residuals plots).

To conclude, we removed the outlying observations identified by the robust analysis, namely, TAMex0344/0719/0775/0802/0805/0807/0821/1534, and re-run the classical analysis without them (Table 6,*). We now had $p \simeq 0.9325$ for the SF normality test. Also, we observed a reduction in the number of SNPs detected by the classical methodology from 5 to 3, where zagl1.6 no longer appears as a significant association, reinforcing the idea that it was really a false positive. It might be expected though, that the classical analysis without the outlying observations would produce the same results as the robust analysis with all observations. However, SNP PZB00049.7 now misses significance by merely 0.0092. This happens because the robust analysis gave the outlying observations weights between 0.36 and 0.55, whereas the classical approach without the outliers actually corresponds to giving zero weight to the outliers and one to the remaining observations.

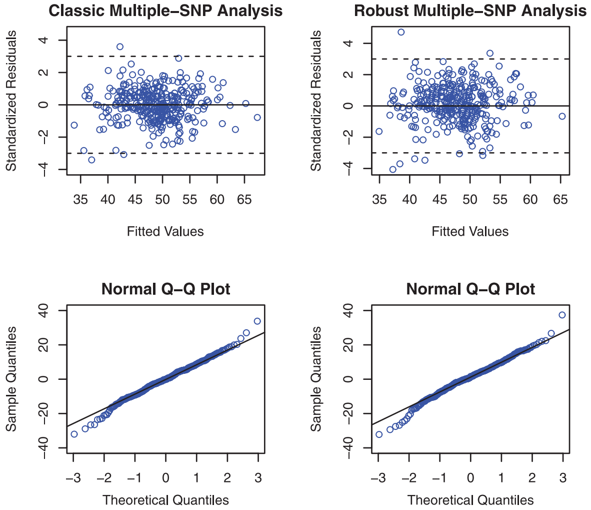

Fig. 2. Conditional residual plots and QQ-normal plots from the multiple SNP analysis.

We have seen in the simulation study that, when testing under the alternative hypothesis with no contamination, the classical methodology can be in some cases slightly more powerful than the robust. However, we argue that removing outliers to achieve normality is not acceptable since they may contain some information on the model. This is also sustained in the simulation contamination setting, where we acknowledged the robust approach to be actually much more powerful than the classic approach. We, therefore, have reasons to believe that the SNPs identified in the robust analysis should be the ones further investigated for association with FERL. Moreover, SNP PZB00049.7 that missed significance in the final classical analysis may well be a false negative.

6 DISCUSSION

In Sections 3 and 4, we compared the performance of classic, robust and non-parametric methodologies in association studies in a particular simulation frame. We showed that under the null, without contamination, the four methods considered have control of the FWER at the desired level. We acknowledged not only the tendency of both the classic and Wilcoxon approaches towards inflated type I error rates, but also that the robust approach proposed (Huber M-regression plus Wald-type tests to assess association) is not as sensitive to outlier contamination as the classical approach and is more powerful than the rank transform approach to detect SNP/trait associations. Despite the fact that the Wilcoxon approach kept close to the robust methodology in terms of power, its tendency to inflated type I error rates and computational issues indicate that for an association study involving a small number of independent SNPs to be tested (as in the simulation study), the robust multiple SNP linear model is preferable over the remaining approaches. In Section 5, we applied the classic and robust methodologies to a published real dataset. This dataset was not in the conditions of the simulation study and was therefore treated accordingly (Yu *et al.*, 2006). Results showed the presence of outliers and therefore the non-normality of residuals. Having these outliers removed and re-running the classical analysis, still the robust methodology proved to be more adequate.

It is clear that the disregard of the non-normality in the classical approach may lead to the use of suboptimal estimators and hence inaccurate conclusions, i.e. spurious associations and/or false negatives. Furthermore, it is well known that the ordinary LS estimator is quite sensitive to outliers and long-tailed distributions and that it has poor efficiency relative to many robust estimators when the errors are not normally distributed. We also argue that the removal of outlying observations on statistical grounds alone is not advisable: (i) with classical methods the 'true outliers' are not always visible due to masking and swamping effects; (ii) it may be easy to identify *gross* outliers but it is not easy to separate those from *mild* outliers and these from *regular data*; (iii) outliers may still contain some relevant information. Plugging-in robust estimators to classical approaches reveals to be a proper way of addressing the problem.

Since most genetic association studies in plants and animals are on economically important traits, too many false positive associations can incur in time and money losses. It is, therefore, compelling that a good compromise is achieved between true and spurious associations. We believe that this work enlightens a new pathway in achieving that goal.

Funding: Fundação para a Ciência e a Tecnologia (FCT), Portugal (EUCANET/ERA-PG/0002/2006); pluri-annual FCT program (Center for Mathematics and its Applications, Lisbon, Portugal).

Conflict of interest: none declared.

REFERENCES

Bacanu,S.A. *et al.* (2002) Association studies for quantitative traits in structured populations. *Genet. Epidemiol.*, **22**, 78–93.

Balding,D.J. (2006) A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.,* **7**, 781–791.

Bureau,A. *et al.* (2005) Identifying SNP predictive of phenotype using random forests. *Genet. Epidemiol.*, **28**, 171–182.

Box,G.E.P. (1953) Non-normality and tests on variances. *Biometrika*, **40**, 318–335.

Cardon,L.R. and Palmer,L.J. (2003) Population stratification and spurious allelic association. *Lancet*, **361**, 598–604.

Carlson,C.S. *et al.* (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.

Chapman,J.M. and Whittaker,J. (2008) Analysis of multiple SNPs in candidate gene or region. *Genet. Epidemiol.*, **32**, 560–566.

Conover,W.J. and Iman,R.L. (1981) Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.,* **35**, 121–129.

Copt,S. and Feser,V. (2006) High-breakdown inference for mixed linear models. *J. Am. Stat. Assoc.*, **101**, 292–300.

Copt,S. and Heritier,S. (2007) Robust alternatives to the F-Test in mixed linear models based on MM-estimates. *Biometrics*, **63**, 1045–1052.

Daszykowski,M. *et al.* (2007) Robust statistics in data analysis - a review, basic concepts. *Chemometr. Intell. Lab.*, **85**, 203–219.

de Haan,J.R. *et al.* (2009) Robust ANOVA for microarray data. *Chemometr. Intell. Lab.*, **98**, 38–44.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Freedman,M. *et al.* (2004) Assessing the impact of population stratification on genetic association studies. *Nat. Genet.*, **36**, 388–393.

Gudbjartsson,D.F. *et al.* (2010) Association of variants at UMOD with chronic kidney disease and kidney stones - role of age and comorbid diseases. *PLoS Genet.,* **6**, e1001039.

Huber,P.J. (1964) Robust estimation of a location parameter. *Ann. Math. Stat.*, **35**, 73–101.

Huber,P.J. (1972) Robust statistics: a review. *Ann. Math. Stat.*, **43**, 1041–1067.

Jannot,A.S. *et al.* (2003) Improved use of SNP information to detect the role of genes. *Genet. Epidemiol.*, **25**, 158–167.

Kwee,L.C. *et al.* (2008) A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, **82**, 386–397.

Li,M. *et al.* (2008) A semiparametric test to detect associations between quantitative traits and candidate genes in structured populations. *Bioinformatics*, **24**, 2356–2362.

Li,M. *et al.* (2009) ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics*, **25**, 497–503.

Liu,B.H. (1997) *Statistical Genomics.* CRC Press, Florida.

Liu,K. and Muse,S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, **21**, 2128–2129.

Malo,N. *et al.* (2008) Accommodating linkage disequilibrium in genetic association analysis via ridge regression. *Am. J. Hum. Genet.*, **82**, 375–385.

Maronna,R.A. *et al.* (2006) *Robust Statistics.* Wiley, Chichester.

Martínez,S. *et al.* (2007) Association genetics in *Pinus taeda* L.I. wood property traits. *Genetics*, **175**, 399–409.

McKean,J.W. *et al.* (2009) Computational rank-based statistics. *Wiley Interdiscipl. Rev. Comput. Stat.* **1**, 132–140.

Moe,M. *et al.* (2009) Association between SNPs within candidate genes and compounds related to boar taint and reproduction. *BMC Genet.*, **10**, 32.

Nyholt,D.R. (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage desiquilibrium with each other. *Am. J. Hum. Genet.*, **74**, 765–769.

Pires,A.M. and Rodrigues,I.M. (2007) Multiple linear regression with some correlated errors: classical and robust methods. *Stat. Med.*, **26**, 2901–2918.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Pritchard,J.K. *et al.* (2000a) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard,J.K. *et al.* (2000b) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170–181.

Ronchetti,E. (1987) Robust C(α)-type tests for linear models. *Indian J. Stat. Ser. A*, **49**, 1–16.

Schrader,R.M. and Hettmansperger,T.P. (1980) Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, **67**, 93–101.

Tan,S. *et al.* (2010) Large effects on body mass index and insulin resistance of fat mass and obesity associated gene (FTO) variants in patients with polycystic ovary syndrome (PCOS). *BMC Med. Genet.*, **11**, 1–9.

Tao,W.J. and Boulding,E.G. (2003) Association between single nucleotide polymorphisms in candidate gene and growth rate in the Artic Charr (*Salvelinus alpinus*). *Heredity*, **91**, 60–69.

Tukey,J.W. (1960) A survey of sampling from contaminated distributions. In Olkin,I. *et al.* (eds) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Stanford, CA, pp. 448–485.

Wang,T. and Elston,R. (2006) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.*, **80**, 353–360.

Weber,A.L. *et al.* (2007) Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp. *parviglumis*). *Genetics*, **177**, 2349–2359.

Weber,A.L. *et al.* (2008) The genetic architecture of complex traits in teosinte (*Zea mays* ssp. *parviglumis*): new evidence from association mapping. *Genetics*, **180**, 1221–1232.

Wu,R. *et al.* (2007) *Statistical Genetics of Quantitative Traits: Linkage, Maps and QTL.* Springer, New York.

Xiang,Z. *et al.* (2009) Efficient algorithm for genome-wide association study. *ACM Trans. Knowl. Discov. Data*, **3**, 4.

Xu,C. *et al.* (2010) Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. *Mol. Cancer*, **9**, 1–12.

Yu,J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Zhao,K. *et al.* (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet.* **3**, e4.

Zhao,W. *et al.* (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.*, **34**, D752–D757.

Zou,F. *et al.* (2003) Rank-based statistical methodologies for quantitative trait locus mapping. *Genetics* **165**, 1599–1605.