Metabolomics: Open Access
ISSN: 2153-0769

Acharjee et al., Metabolomics 2013, 3:3
DOI: 10.4172/2153-0769.1000126

Research Article Open Access

Comparison of Regularized Regression Methods for ~Omics Data

Animesh Acharjee[1,2,4*], Richard Finkers[2,3], Richard GF Visser[2,3] and Chris Maliepaard[2,3]

[1]Graduate School Experimental Plant Sciences, Droevendaalsesteeg 1, Wageningen, The Netherlands
[2]Wageningen UR Plant Breeding, Wageningen University and Research Center, Wageningen, The Netherlands
[3]Centre for BioSystems Genomics, Wageningen, The Netherlands
[4]BASF-CropDesign N.V., 9052 Gent (Zwijnaarde), Belgium

Abstract

Background: In this study, we compare methods that can be used to relate a phenotypic trait of interest to an ~omics data set, where the number or variables outnumbers by far the number of samples.

Methods: We apply univariate regression and different regularized multiple regression methods: ridge regression (RR), LASSO, elastic net (EN), principal components regression (PCR), partial least squares regression (PLS), sparse partial least squares regression (SPLS), support vector regression (SVR) and random forest regression (RF). These regression methods were applied to a data set from a potato mapping population, where we predict potato flesh colour from a metabolomics data set.

Results: We compare the methods in terms of the mean square error of prediction of the trait, goodness of fit of the models, and the selection and ranking of the metabolites. In terms of the prediction error, elastic net performed better than the other methods. Different numbers of variables are selected by the methods that allow variable selection but seven variables were in common between LASSO, EN and SPLS. SPLS performed better than EN with respect to the selection of grouped correlated variables.

Conclusions: Four out of these seven variables selected by LASSO, EN, SPLS were putatively identified as carotenoid derived compounds; since the carotenoid pathway is important for flesh colour of potato, this indicates that meaningful compounds are selected. We developed a web application that can perform all the described methods, and that includes a double cross validation for optimization of the methods and for proper estimation of the prediction error.

Keywords: Metabolomic; Regularized regression; Variable selection; Variable ranking

Introduction

High-throughput technologies like microarrays [1,2], mass spectrometry (e.g. LC-MS, GC-MS) [3,4] and protein chips [5-7] have gained much interest in the biological domain. These techniques allow one to measure thousands of variables (genes, metabolites, proteins) simultaneously. The data generated by these techniques are often denoted as ~omics data [8]. These data sets are generally very large in terms of the number of variables (p) and often small in terms of the number of the biological samples (n). In statistics, this problem is often termed as the "large p and small n problem" (p>>n). In such wide data sets, there will be collinearity due to p>>n [9], but also because of high correlations due to common biological functions (e.g. metabolites in the same pathway).

In many of these ~omics situations, one wants to find functional relationships between a phenotypic traits of interest and the ~omics variables, and often the interest would also be in selecting a smaller subset of the variables that have good prediction of the trait. Even if the prediction is not very strong, the top ranked or selected variables may still be meaningful with respect to having a functional relationship to the phenotypic trait.

In traditional statistical methods, multiple linear regression techniques are used for prediction situations such as outlined above, but due to the high collinearity, these methods cannot be applied. Therefore, we need different approaches: penalization regression methods or machine learning methods.

We wanted to compare the different methods on real data, but we still wanted to be able to infer whether results were biologically meaningful, so we chose a trait for which a fair amount of information is already available, including a possible relationship to underlying metabolites. Therefore, we considered potato tuber flesh colour as the phenotypic trait of interest, the response in our regression and a large metabolomics data set as the set of predictor variables. For tuber flesh colour, there is a well-established relationship to the carotenoid pathway, and especially to beta caroteen [10], therefore, compounds related to this pathway are expected to be observed in a top list or selected set of predictive variables.

We apply a double cross validation scheme to include optimization of any hyperparameters needed in the models and allow estimation of prediction error.

We apply different regression methods: ridge regression (RR) [11], LASSO [12], elastic net (EN) [13], principal component regression (PCR) [14], partial least squares regression (PLS) [15], sparse PLS regression (SPLS) [16], support vector regression (SVR) [17] and random forest regression (RF) [18].

We use univariate regression as a reference and compare the results of univariate regressions with multiple regression methods. We also study the properties of these methods both from a theoretical point of view, as well as their performance in practical situations in terms

***Corresponding author:** Animesh Acharjee, Wageningen UR Plant Breeding, Wageningen University and Research Center, Wageningen, The Netherlands, Tel: +32 9242 3418; Fax: +32 9 2415089; E-mail: animesh.acharjee@gmail.com

Received July 10, 2013; **Accepted** October 01, 2013; **Published** October 10, 2013

Citation: Acharjee A, Finkers R, Visser RGF, Maliepaard C (2013) Comparison of Regularized Regression Methods for ~Omics Data. Metabolomics 3: 126. doi:10.4172/2153-0769.1000126

of variable selection [19], or ranking of the variables, grouping of correlated variables in variable selection [13], and the prediction error. Regarding the grouping of correlated variables, we are interested in finding out whether in the variable selection methods (LASSO, EN, SPLS) variables are selected as a group or not, and we also compare regression coefficients of these correlated variables. In addition to the real data, we used simulated data to study the selection and grouping properties of the methods with respect to highly correlated variables.

So far in literature, comparison studies are usually focused on classification methods instead of regression methods [20], and data used in these studies often were transcriptomics data [21,22]. In the context of regression, Kiers and Smilde [9] did a comparison of various multiple regression methods on simulated data with collinear variables, but their study was mainly focused on prediction and comparison of the regression coefficients when predictor variables are collinear. Menendez et al. [23] reported comparison of stepwise linear regression, LASSO, EN and RR, but did not cover other penalization methods, such as SPLS, PLS, PCR, RF and SVM. We compare these methods (RR, EN, LASSO, SPLS, RF, SVM, PLS, PCR), in terms of mean square error of prediction, goodness of fit, variable selection and the ranking of the variables. In addition, we developed a web application Omics Fusion with all the methods mentioned, including a double cross validation procedure. This website can be accessed from: http://www.plantbreeding.wur.nl/omicsFusion/

Materials and Methods

Plant material

Ninety-one individuals from a diploid mapping population of potato (denoted as CxE) were used in this study. Clone C is a hybrid between *Solanum phureja* and *Solanum tuberosum*. Clone E is the result of a cross between Clone C and *Solanum vernei* [24]. All clones were grown in the field, Wageningen, The Netherlands in 1998. For each genotype, tubers from two plants were collected and representative samples from these tubers of each genotype, were used for phenotypic analysis directly after harvest, and for LC-MS.

Evaluation of phenotypic traits

Many quality traits were collected for this potato population [24-26]. In this study, we used one well-studied phenotypic trait (potato tuber flesh colour), allowing better to compare methodology and to be able to verify the obtained results. Potato tuber flesh colour was visually scored on a scale from 1 (white) to 9 (dark yellow/orange) in three repeats, consisting of two plants each. Flesh colour scores were averaged over the three repeats.

Data preprocessing

For metabolomics analysis, the exact same material (potato tubers of the same genotypes) was used for Liquid chromatography–time of flight mass spectrometry (LC-QTOF MS) analysis, which resulted in over 16,000 individual mass peaks. Mass peak signals below background were removed, resulting in about 10,000 remaining mass peaks. The next step was to make a selection of these 10,000 peaks based on skewness of the data, and all mass peaks with a skewness score below -2 and above 2 for the progeny and a score below -1 and above 1 for the parental repeats were discarded. The signal intensities of the 1,100 remaining mass peaks were then correlated to the available quality trait data of this population, in order to obtain the most interesting metabolites, i.e. the metabolites linked to quality traits. Significance of these correlations was calculated using Student's t-test. A number of 163 mass peaks with the highest significance ($p<0.0005$) was selected. Before analysis, the metabolite data was ^{10}log transformed for symmetry and then autoscaled. Autoscaled variables have a mean of zero and a variance (and also standard deviation) of one, thereby giving all variables (mass scan numbers) an equal weight in the analysis. LC-MS peaks are characterized by their mass and scan number (mass_scan).

Statistical methods for regression in p >> n situations

Methods used: We compared the prediction, variable selection and ranking of variables. In this section, we first review the regression model in these eight methods. For all methods, values for one or more tuning parameters needed to be chosen. This was done using tenfold cross-validation, as described in the section on criteria for comparison of the methods.

Regression methods: Regression methods are essentially curve-fitting approaches. When there is one response variable and one predictor variable, simple linear regression consists of finding the best straight line relating the response to the predictor variable. In case of multiple predictors, a hyperplane is fitted. The usual criterion, the least squares criterion, minimizes the sum of squared distances between the observed responses and the fitted responses from the regression model [27]. We can represent the least squares criterion as:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Where; y=response vector (here: flesh colour); β=regression coefficients; x=predictor variables (the log intensity values of the mass scans from LC-MS data measured over different samples).

Here, we are describing nine different regression methods which were applied to relate potato tuber flesh colour to the LC-MS data set.

Univariate regression: Univariate regression was used as a reference. We compare the variable selection and ranking of variables in the multivariate regression methods to the results from the univariate regressions of flesh colour to each of the individual LC-MS peaks. Univariate regression with a FDR (False discovery rate) adjustment was done according to the procedure of Benjamini and Hochberg [28].

Penalization or shrinkage methods: Shrinkage methods, also called penalization methods, impose a penalty on the size of the regression coefficients. The penalty term is also called a 'regularization parameter'. We have grouped the methods according to the type of penalty applied to the regression coefficients. The mean square error (MSE) of a regression model can be decomposed into two components: the square of the bias (difference between the estimate and the expectation of a parameter) and the variance. In situations with high collinearity (p >> n), regression models usually have a very large variance, and the MSE will mainly be determined by this large variance. Therefore, in such situations, it can be advantageous (in terms of decreasing the MSE) to accept some bias if it is allows us to decrease the variance by considerable amount [29]. Penalization methods impose a bias by applying a penalty to the regression coefficients.

Continuous penalization methods: In this category of regression methods, shrinkage factors can take any value between zero and infinity. LASSO, RR and EN belong to this category. The value of the shrinkage parameter decides the amount of penalization applied to the regression coefficients. We use tenfold cross validation [20], to choose the optimum penalty value; this will be discussed in detail in the section

about criteria for comparison of the methods

Ridge regression (RR): Ridge regression [11] shrinks the regression coefficients by imposing a penalty on the sum of squares (L2 norm) of the regression coefficients.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

The left part of the term shown above is the usual least squares criterion. In the right part, λ_2 is a shrinkage factor applied to the sum of the squared values of the regression coefficients. The larger the value of λ_2, the heavier the penalty on the regression coefficients, and the more they are shrunk towards zero. In ridge regression, all the regressor variables stay in the model since regression coefficients do not become exactly zero (that would be equivalent to variables dropping out of the regression model). Ridge gives equal weight to absolutely correlated variables in the data set [29].

Lasso: The LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani [12] is another regularization method, but here the penalty is applied to the sum of the absolute values of the regression coefficients, the L1 norm. Mathematically, we can write this in the following way:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} \left| \beta_j \right|$$

Again, the left part of the term is the normal least squares criterion. The right part now is the penalized sum of the absolute values of the regression coefficients. Similar to ridge regression, the shrinkage parameter (λ_1) has to be decided on, and again we use tenfold cross validation for this. Penalizing the absolute values of the regression coefficients has the effect that a number of the estimated coefficients will become exactly zero, which means that some regressors drop out of the regression model so that a LASSO fitted model will consist of fewer variables than the original number of available regressors. In other words, LASSO can implicitly perform variable selection. The number of selected variables is upper limited by the numbers of samples (n). In case of absolutely correlated variables, LASSO just selects one of these variables and ignores the rest in the group [29].

Elastic net (EN): Elastic net [13] is a combination of LASSO and ridge regression. It uses both a ridge penalty (penalty on the sum of the squares of the regression coefficients) and a LASSO penalty (on the sum of the absolute values of the regression coefficients):

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} \left| \beta_j \right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

In elastic net, we optimize both penalty parameters, simultaneously using tenfold cross validation. Variable selection is encouraged by the LASSO penalty (λ_1) and groups of correlated variables get similar regression coefficients. Groups of correlated variables are either in or out of the model [13]. In contrast with LASSO, the number of selected variables is not limited by the number of individuals.

Discrete penalization methods: Partial least squares (PLS) and Principal components regression (PCR) are based on latent variables or components, which are linear combination of the original variables. For both methods, it is essential to select the optimum numbers of latent components for prediction of the response variable. We used tenfold cross validation to choose the optimum number of latent components based on the smallest mean square error of prediction (MSEP) value. The number of latent components can only take discrete values; hence these methods are discrete penalization methods.

Principal components regression (PCR): Principal components regression [30] is a combination of principal components analysis (PCA) and multiple linear regression. First, PCA is done on all original regressors and each component (latent variable) is represented by a linear combination of the original variables. The number of latent variables (components) is chosen by tenfold cross validation and the response is regressed on the selected latent variables. These latent variables in PCA are uncorrelated, and there are fewer latent variables than the number of individuals, therefore solving the collinearity problem. In PCR, the principal components are found by maximization of the variance in the predictors; the covariance of the predictors with the response variable is not taken into account, as is the case in partial least squares regression.

Partial least squares (PLS): Partial least squares (PLS) [15,31,32] is a method to relate a single response variable or a matrix of response variables to a matrix of regressor variables. Here, we are considering only a single trait as the response. PLS is a dimension reduction method like PCA, but it uses a different criterion: maximization of the covariance between the latent variables and the response. As a consequence, usually fewer components are required for prediction as compared to PCR. The optimum number of latent components is chosen by tenfold cross validation. Since the optimum number of latent components is a discrete number, this method is also a discrete penalization method. Like in PCR, latent variables in PLS are also uncorrelated.

A hybrid penalization method: In this section, we consider a method in which two different types of penalties (continuous and discrete) are applied simultaneously.

Sparse partial least square (SPLS): SPLS [16] is a combination of two different penalties. The continuous penalty is a LASSO penalty and discrete penalization is achieved by PLS. Variable selection is achieved by LASSO, dimension reduction by PLS. The respective hyperparameters, i.e. the number of PLS components and the size of the LASSO penalty are optimized simultaneously by tenfold cross validation. As in normal PLS, each of the latent components is a linear combination of the original variables.

Machine learning methods: The goal of machine learning is to build a computer system that can adapt and learn from experience [33]. Machine learning methods can handle data which are not normally distributed, whereas the methods mentioned above assume normality. Machine learning methods can also handle nonlinear relationships between response and predictor variables.

Support vector machine (SVM): The support vector machine (SVM) [17] was originally developed in a classification [34] context and maximizes predictive accuracy, while avoiding overfitting [29,35] to the data. Two parameters, such as epsilon (insensitive zone) and regularization parameter "C", are optimized [17]. However, the methodology can also be used in a regression model [35]. Mathematically, given the input data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to find a function which will fit the following equation:

$$f(x) = wx + b,$$

Where w is a weight vector and b is a constant.

The goal of support vector regression (SVR) is to find a function

f(x) that has at most ε deviation [35] from the actually obtained targets (response) for all the predictors, and at the same time, minimizes the distance between predicted and target values. SVR does not encourage grouping or variable selection.

Random Forest (RF): A random forest [18] is a collection of unpruned decision trees [29], usually developed for a classification purpose, but this method can be applied in a regression context as well [36]. A random forest model is typically made up of hundreds of decision trees. Each decision tree is built from bootstrap samples of the data set. That is, some samples will be included more than once in the bootstrap sample, and others will not appear at all. Generally, about two thirds of the samples will be included in this training dataset, and one third will be left out (called the out-of-bag samples or OOB samples). In RF regression, the prediction error is calculated as the average prediction error over OOB predictions. Variable importance [18] can be quantified in random forest regression. Variables used which decrease the prediction error obtain a higher variable importance. Two parameters have to be chosen in RF regression: the number of candidate variables (mtry) to choose from at any split in the regression trees, and the number of trees (ntree). The number of variables to choose from was optimized by cross validation. The number of trees was fixed at 500 trees.

Criteria for comparison of the methods

Double cross validation: All methods above require input values for one or more hyperparameters (e.g. the number of components in PCR and PLS, the penalty parameter lambda in ridge regression and LASSO, etc.), and the values for these hyperparameters were optimized using cross validation. Using a single cross validation to estimate both the hyperparameters and the prediction error will result in an overly optimistic estimate of the error rate value [37]. Hence, a double cross validation scheme was used [20,38,39]. We used tenfold double cross validation [29] for choosing optimum values for the hyperparameters, and to estimate prediction error. First, tenfold cross validation is performed and one tenth portion of the data is left out for estimation of the prediction error, this portion is called the outer test set. The remaining nine tenth portions is the outer training set. Another tenfold cross validation uses nine tenth portions of the outer training data set, which then are called the inner training sets and one tenth portions, which are called the inner test sets. The inner cross validation loop is performed to optimize hyperparameters such as the number of principal components or PLS components, or the amount of shrinkage in ridge, lasso, elastic net. The outer loop cross validation is used to quantify properly the predictive value of the model on independent data. The hyperparameters are chosen which give the lowest MSEP values on the inner test data. We run this procedure 100 times, each with different tenfold divisions, and in each division, prediction was done and then averaged over the results from 100 runs to obtain results in Table 1. The same divisions were used for all regression methods.

Mean squared error of prediction (MSEP): The mean squared error of prediction (MSEP) is frequently used to assess the performance of regressions [40,41]. MSEP of a regression can be estimated by predicting the test data set and comparing the predicted response with the observed response of the test set samples. Often, a (large enough) independent test set is not available. In such situations, the MSEP has to be estimated from the test data in cross-validation. An estimate of the MSEP is obtained by averaging the squared prediction errors of the outer test samples. Mathematically, we can write

$$\mathbf{MSEP} = (1/n)\sum_{i=1}^{n} (y_i - y_{predicted})^2$$

Where y and $y_{predicted}$ are the observed and predicted response values for the i th test sample, respectively. We calculated and compared the MSEP on outer test sets for all the regression methods to evaluate the different methods. We consider the lowest MSEP to correspond to a better predictive model.

Variable selection or ranking: Variable selection is defined as selecting subsets of variables that together have predictive power. LASSO, SPLS and EN are variable selection methods, as they select a subset of the predictor variables. For the variable selection methods, we investigated the numbers of variables and the identity of the variables which were selected by those methods. For the methods that do not include variable selection, we can still rank the variables, according to their estimated regression coefficients or variable importance measures. In case of RR, PLS, PCR, RF, all the variables remain in the regression model. In case of SVM, we do not perform variable ranking or variable selection, as we cannot estimate regression coefficients. We compare the ranking between these different methods and we compare the ranks in the ranking methods with the selection of variables in the variable selection methods.

Goodness of fit (R^2): Goodness of fit (R^2) of statistical models is used to describe how well the predictions fit a set of observations. It is a measure for the proportion of variability in a data set that is accounted for by the statistical model. In our analysis, we use R^2 values to compare the methods. R^2 is calculated as the square of the Pearson correlation between observed and fitted values for training and test data set, and is converted to a percentage. The usual R^2 from a linear regression is just a measure of goodness-of-fit of the data at hand (training data), and not for future predictions (test data). We calculated R^2 values both for training and for the cross-validation test data. It is important to realize that the R^2 on the cross-validation test data is a prediction R^2, not just a goodness-of-fit for the data at hand; it refers to future predictions on independent samples.

Method	Training data (R2) %	Training data (sd)	MSEP	MSEP (sd)	Test data (R2)%
RR	48.1	1.979	1.30	0.031	36.1
LASSO	61.8	5.242	1.24	0.033	41.4
EN	65.4	3.804	1.21	0.035	44.1
PCR	61.8	7.180	1.29	0.032	37.2
PLS	60.0	8.927	1.32	0.037	35.9
SPLS	36.5	11.593	1.31	0.044	36. 36.536.5 36.5
RF	25.0	4.653	1.27	0.032	40.5
SVM	79.7	6.713	1.37	0.032	30.8

Table 1: Comparison of the eight multivariate regression methods based on R^2 for training data, standard deviation of R^2 (sd) for training data, MSEP, sd of MSEP and R^2 for test data set. R^2 for training data, MSEP and R^2 for test data are the mean values of the double cross validation scheme for 10 different divisions with 100 runs and then averaged.

Omics fusion web application

Omics Fusion is a web-based application written in Java EE 6 and Struts 2 and runs on a glassfish v3 application server. SQLite v3 (http://www.sqlite.org) is used as the back end database management system. Standardized excel sheets are used to upload data to Omics Fusion. The end user can select one or several of the described methods for data analysis. An Oracle Grid Engine 6.2u5-1 cluster (http://www.oracle.com) is used to execute the R based (http://cran.r-project.org/) script in parallel. The end user is notified by email upon completion of the analysis. Results are summarized within the web-based interface.

Simulation study

In order to compare the methods with respect to their properties and performance when effects and characteristics of predictors and response are exactly known, we performed a small simulation study in which a population of 100 individuals was generated, with a response y that depended on only 12 out of 1000 variables, no error added (y=x1+x2+...+x12). The effect of these 12 variables was large (together completely determining y), if combined, but each predictor by itself had just an effect the size of roughly 0.2 standard deviations of the response, simple linear regression R^2 per true predictor was around 0.2, as well. From these 12 variables, two at a time (six pairs of two) had pairwise stronger and less strong positive correlations, with correlations 1.0, 0.95, 0.9, 0.85, 0.8 and 0.5 for the six pairs. We were interested in whether or not the true predictors were selected by the methods that perform variable selection; in the rankings of the true predictors in the methods that do not perform variables selection; in the size of the estimated regression coefficients (close to the true effect or not), and in the effect of the correlations among the six pairs of true predictors.

Results

Univariate regression

Univariate regression analysis without FDR correction resulted in 29 significant variables ($p<0.05$). MSEP and R^2 (training) values were calculated (Supplementary Table S1). Univariate regression followed by an FDR adjustment according to Benjamini and Hochberg [28] resulted in still 23 significant variables at an FDR threshold of 0.05. Variable 294_0182 had the highest R^2 (training) of 22.6% and lowest MSEP value of 1.92.

Comparison of the regression methods

We used a double cross-validation scheme for comparison among the methods in 100 runs with different divisions of the data. Differences in MSEP values between different methods were rather small (Table 1). EN had the lowest MSEP value, lower than RR, PLS, SPLS, LASSO, SVM and RF (Table 1). The standard deviation of the MSEP values was about 0.03 for all the methods except SPLS (Table 1). Although SVM showed the highest R^2 value for the training data sets, it did not perform well for prediction on the test data.

Metabolite identification

In the metabolomics analysis, out of 163 mass scans, only a few were putatively identified as being important as they are ranked and also selected by variable section methods. Mass scans 396_1508, 193_1508, 373_1301 and 557_1301 were identified as important variables and were putatively identified as carotenoid derived compounds. More specifically, scan number 1508 putatively identified as 4,7-megastigmadiene-3,9-diol-glucoside and scan number 1301 as 2,3-dihydroxy-4-megastigmen-9-one-glucoside, which are non-volatile glucosides of carotenoid-derived volatile metabolites. From the literature, the relationship between potato flesh colour and the carotenoid pathway is well established [10], so these results makes sense even when the models are not explaining huge parts of the phenotypic variation. Additionally, combining the metabolite data with gene expression profiling for this population (CxE) resulted in the observation that the gene beta-carotene hydroxylase (Bch), the most important gene responsible for flesh colour in potato [10], is also highly correlated with these two metabolites [42].

Comparison of standardized regression coefficients

We ranked the variables based on the standardized regression coefficients in all regression methods and then compared them to the ranks in univariate regression (Figure 1). In Figure 1, variables are in x-axis, whereas standardized regression coefficients are plotted in y-axis. Some of the variables get a zero value for the regression coefficient for those methods that also do variable selection (LASSO, SPLS, EN). Variables that were selected in these methods mostly correspond to variables that also had the largest regression coefficients in the univariate regressions, for example: mass_scan 294_0182 gets the largest negative regression coefficient for both the variable selection (LASSO, EN, SPLS) and the ranking methods (PCR, PLS, RR and RF), and also in univariate regression (Supplementary Table S2). We also compared the ranking of the LASSO selected variables (24 mass_scans) in the ranking methods (PCR, PLS, RR, RF) and the variable selection methods (EN, SPLS) (Table 2).

Pearson correlation coefficients of the twenty-four variables selected by LASSO and flesh colour were visualized in a heat map (Figure 2). There were high correlations among some of the selected variables, for example: between mass_scans 396_1508 and 193_1508 with a correlation coefficient of 0.86; between 373_1301 and 557_1301 with a correlation coefficient of 0.85; between 396_1508 and 373_1301 with a correlation coefficient of 0.65; between 396_1508 and 557_1301 with a correlation coefficient of 0.60.

Comparison of standardized regression coefficients based on variable selection

We compared EN, LASSO and SPLS in terms of the numbers of selected variables. EN, LASSO and SPLS select 17, 24 and 10 variables, respectively. For the pairwise comparison between EN and LASSO, 17 variables are in common. Between LASSO and SPLS, seven variables are in common, the same set of seven are also in common between EN

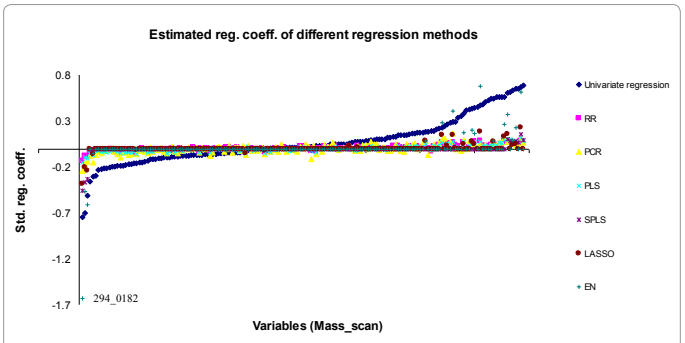


Figure 1: Standardized regression coefficients in the different regression methods. The order of the mass scans on the x-axis is based on the regression coefficients from the univariate regressions. Variable 294_0182 has the highest negative regression coefficient in all the methods shown.

LASSO selected (mass_scan)	PLS	PCR	RR	RF	EN	SPLS	Univariate
294_0182	1	1	1	1	1	1	1
187_0439	6	5	2	24	4	3	18
557_1301	2	9	3	5	3	4	4
225_0983	7	3	6	3	5	2	2
622_0122	14	35	4	71	2	0	20
373_1301	3	15	8	9	7	10	9
157_0111	11	2	11	87	6	0	32
1396_0110	18	10	9	88	8	0	38
393_1416	5	6	5	7	9	0	10
843_0095	24	13	13	16	15	0	51
239_0200	23	79	20	93	11	0	23
207_0918	30	32	25	12	16	0	14
499_1599	17	4	14	106	17	0	30
192_0142	13	11	12	35	12	0	26
193_1508	10	75	18	19	10	7	6
738_0133	15	41	7	10	13	0	22
491_0148	26	24	35	161	0	0	144
396_1508	9	36	21	17	14	9	7
323_0837	56	76	53	149	0	0	110
719_0105	20	88	17	90	0	0	35
589_0148	28	33	33	108	0	0	163
303_1245	61	65	43	77	0	0	101
225_1296	21	43	19	62	0	0	64
484_1597	72	155	29	13	0	0	31

Table 2: Ranking of the twenty-four variables (mass_scans) selected by LASSO is shown in the first column. The other columns show the ranks of these twenty-four variables in PLS, PCR, RR, RF, EN, SPLS and univariate regression analysis. In RF ranking was done based on increase in MSE of the OOB samples after permutation of the variable. Out of these twenty-four variables, some are not selected by EN (17 variables selected) and SPLS (10 variables selected) and these are marked as 0. Variables in bold font are also selected in EN and SPLS.

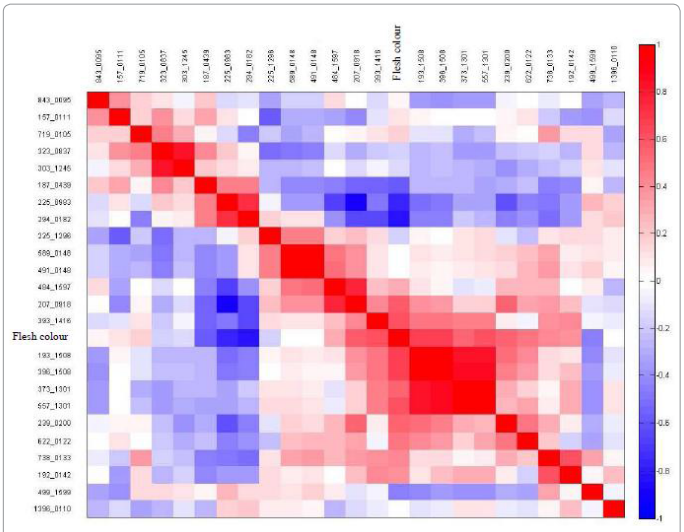


Figure 2: Correlation heat map of 24 variables selected with LASSO. Positive correlations among variables are shown in red, negative correlations are shown in blue. Variables 396_1508, 193_1508, 373_1301 and 557_1301 are grouped together based on high positive correlations.

and SPLS, so these seven are in common between all three methods (Supplementary Figure S1).

Omics fusion

The interface of Omics Fusion contains two major parts: data submission and results visualization. Data submission allows end users to start a new analysis in four distinctive steps: provide user details, upload of excel sheets with the data, select analysis methods and start analysis after final confirmation. An unique token will be sent *via* email to the user. The Omics Fusion analysis for the data set described in this manuscript takes 38 minutes using 20 cores in parallel. Users are notified upon completion of the analysis *via* e-mail. The results of each analysis can be obtained after entering the unique token. The results of all selected methods will be summarized in a table, a snapshot of which is shown (Figure 3). An overall mean rank is calculated for each of the predictor variables and the resulting table is ordered accordingly. The rank for the individual methods can be obtained by hovering over the coefficients. To quickly scan the results, the background of each coefficient is color coded blue (top ranks) to white (lowest ranks). Each predictor variable is hyperlinked and can be used to show the response variable vs. predict or variable for easy interpretation of the results (Supplementary Figure S2). This tool can be found: URL: http://www.plantbreeding.wur.nl/omicsFusion/

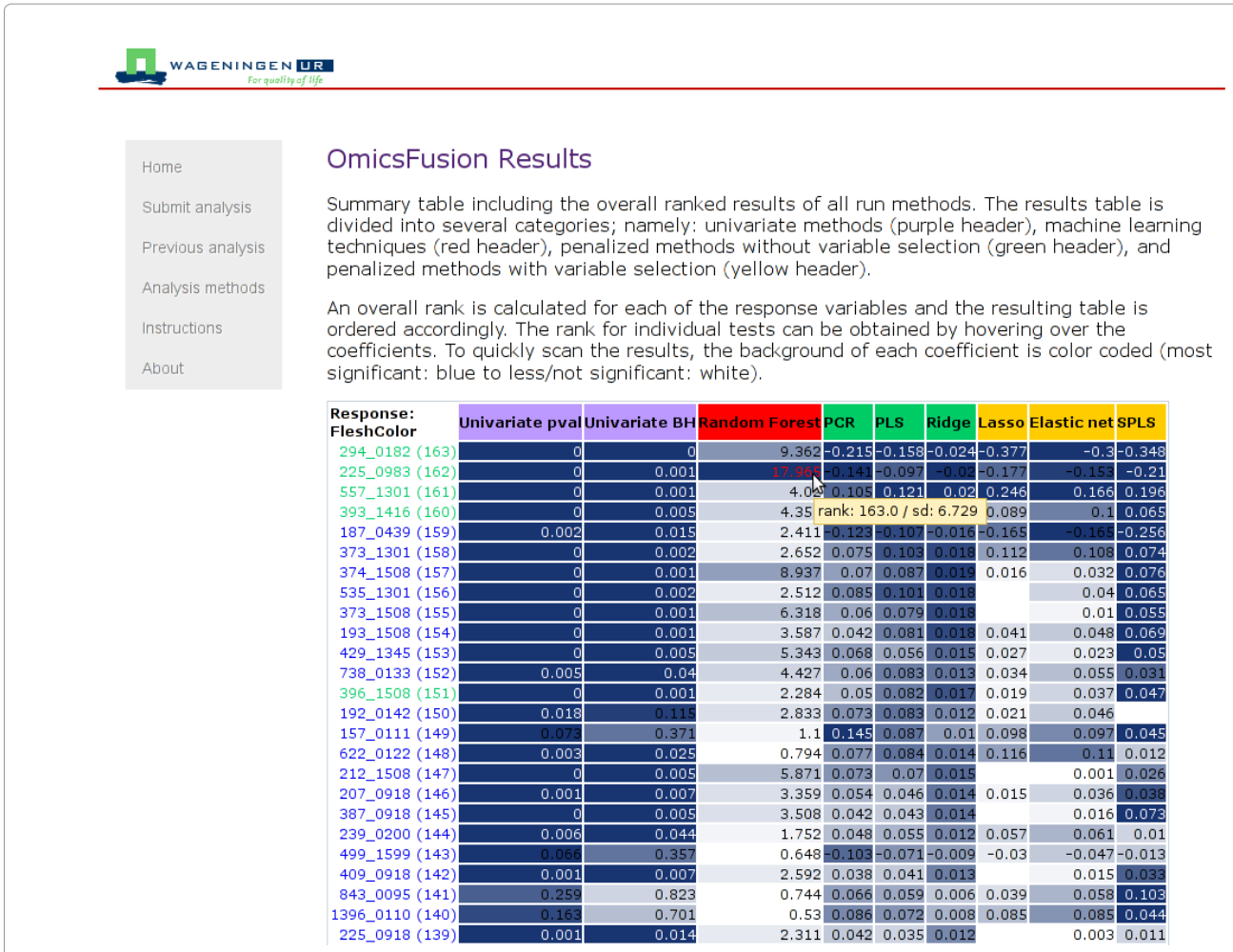
Simulation results

In the simulated data, the variable selection methods LASSO, SPLS and elastic net were often able to select the twelve true predictors, while only few or hardly any of the noise variables were selected. If noise variables were selected, they had much lower regression coefficients than the true predictors, and regression coefficients of the true predictors were often close to the true effect (generally a little bit underestimated, shrunken). In some instances, especially LASSO, but also SPLS and elastic net selected only one of two highly correlated true predictors, in which case, usually this selected predictor absorbed the effect of the one that was not selected, so that the regression coefficient of the selected predictor was estimated as almost twice its true effect. For the methods that do not perform variable selection, usually most of the twelve true predictors were ranked on top and present in the top 12. Occasionally, one to four true predictors were not present in the top 12, but then very often, although not always, they still had high ranks in comparison with the noise variables, for example ranking 14 or 17. In contrast with the variable selection methods (LASSO, elastic net, SPLS), for ridge regression, PCR and PLS the regression coefficients of the true predictors were strongly shrunken (because of the presence of all the noise variables in the model). The two predictors that had a correlation of 1 had, as expected, exactly equal regression coefficients in these methods; with lower correlations, the estimated regression coefficients were still very similar, but more dissimilar as the correlation decreased.

Conclusions

We compared nine regression methods based on MSEP, R^2 on the training and on the test set, variable selection and variable ranking. The range of R^2 values for the training set across different methods is from 48.1% to 79.7%. As expected, R^2 is always lower on the test than on the training data, except for RF. RF includes an internal cross validation on the training data already, using the out-of-bag (OOB) samples. In addition, we used a double cross validation scheme for RF and the other multivariate methods. As a consequence the RF model is actually based on fewer samples than the other methods, and therefore, the R^2 value might be lower.

In the case of other methods, there is a difference between the R^2 for the training and the R^2 for the test set. Taking the R^2 for the training



Response: FleshColor	Univariate pval	Univariate BH	Random Forest	PCR	PLS	Ridge	Lasso	Elastic net	SPLS
294_0182 (163)	0	0	9.362	-0.215	-0.158	-0.024	-0.377	-0.3	-0.348
225_0983 (162)	0	0.001	17.966	-0.141	-0.097	-0.02	-0.177	-0.153	-0.21
557_1301 (161)	0	0.001	4.04	0.105	0.121	0.02	0.246	0.166	0.196
393_1416 (160)	0	0.005	4.35				0.089	0.1	0.065
187_0439 (159)	0.002	0.015	2.411	-0.123	-0.107	-0.016	-0.165	-0.165	-0.256
373_1301 (158)	0	0.002	2.652	0.075	0.103	0.018	0.112	0.108	0.074
374_1508 (157)	0	0.001	8.937	0.07	0.087	0.019	0.016	0.032	0.076
535_1301 (156)	0	0.002	2.512	0.085	0.101	0.018		0.04	0.065
373_1508 (155)	0	0.001	6.318	0.06	0.079	0.018		0.01	0.055
193_1508 (154)	0	0.001	3.587	0.042	0.081	0.018	0.041	0.048	0.069
429_1345 (153)	0	0.005	5.343	0.068	0.056	0.015	0.027	0.023	0.05
738_0133 (152)	0.005	0.04	4.427	0.06	0.083	0.013	0.034	0.055	0.031
396_1508 (151)	0	0.001	2.284	0.05	0.082	0.017	0.019	0.037	0.047
192_0142 (150)	0.018	0.115	2.833	0.073	0.083	0.012	0.021	0.046	
157_0111 (149)	0.073	0.371	1.1	0.145	0.087	0.01	0.098	0.097	0.045
622_0122 (148)	0.003	0.025	0.794	0.077	0.084	0.014	0.116	0.11	0.012
212_1508 (147)	0	0.005	5.871	0.073	0.07	0.015		0.001	0.026
207_0918 (146)	0.001	0.007	3.359	0.054	0.046	0.014	0.015	0.036	0.038
387_0918 (145)	0	0.005	3.508	0.042	0.043	0.014		0.016	0.073
239_0200 (144)	0.006	0.044	1.752	0.048	0.055	0.012	0.057	0.061	0.01
499_1599 (143)	0.066	0.357	0.648	-0.103	-0.071	-0.009	-0.03	-0.047	-0.013
409_0918 (142)	0.001	0.007	2.592	0.038	0.041	0.013		0.015	0.033
843_0095 (141)	0.259	0.823	0.744	0.066	0.059	0.006	0.039	0.058	0.103
1396_0110 (140)	0.163	0.701	0.53	0.086	0.072	0.008	0.085	0.085	0.044
225_0918 (139)	0.001	0.014	2.311	0.042	0.035	0.012		0.003	0.011

Figure 3: Summary of the OmicsFusion analysis. The overall rank is calculated for each of the predictor variables and the resulting table is ordered accordingly. The rank for the individual methods can be obtained by hovering over the coefficients. To quickly scan the results, the background of each coefficient is color coded blue (top ranks) to white (lowest ranks).

set as a criterion for evaluation of the methods to be used in regression of ~omics data is of limited use, because it only refers to the fit of the data at hand and would be too optimistic for prediction purposes. Therefore, we need instead to have a look at the MSEP. This is why we performed a double cross validation. EN has the lowest MSEP value, which means that EN finds a better predictive model than other methods like RR, LASSO, SVM, PCR, PLS, RF and SPLS. RR, PCR and PLS have similar MSEP values. The lowest univariate MSEP was 1.92 for the variable 294_0182, but for the other methods the MSEP values are lower, which suggests that these methods, which use more than single variables, are predicting better than the best univariate predictor. None of the prediction R^2 values are higher than 50%. For most phenotypic traits, very high R^2 values are not expected, since the phenotype is not just determined by biology, but also by variation in environmental conditions, interactions of genotypes with the environment, measurement and observation error in both the response and the predictors. Moreover, metabolites as quantified by LC-MS do not necessarily comprise the most important components (for example if they are volatile compounds or if they are proteins).

Comparing the variable selection methods, we see that LASSO selected 24 variables and EN selects 17, a subset of those selected by LASSO. The average number of selected variables over 100 runs for EN, LASSO and SPLS were 31.3, 16.1 and 26.7 with standard deviations 11.9, 5.4 and 23.2, respectively. SPLS showed a high variability in the number of selected variables, whereas LASSO had the lowest variability across 100 runs. Regarding consistency of the selected seven variables (Figure 4) across 100 runs, we observed that variable 294_0182 was selected the highest number of times, almost always, in the different folds, whereas 396_1508 was selected the smallest number of times in EN, LASSO and SPLS.

Variables 294_0182, 187_0439 and 225_0983 are consistent in terms of the sign and size of standardized regression coefficients, ranking and size of standardized regression coefficients across methods (Supplementary Figure S3 and S4).

According to Tibshirani [12], LASSO tries to select only one variable from a set of correlated variables, but in our analysis, we find that a group of correlated variables was selected, for example: mass_scans 396_1508 and 193_1508 have a correlation coefficient of 0.86; 373_1301 and 557_1301 have a correlation coefficient of 0.85 (Figure 2). The simulation results showed that when there are two absolutely correlated variables (correlation coefficient of 1), LASSO usually picks only one of the two variables, and in that case, the regression coefficient of the selected variable was close to double the simulated regression

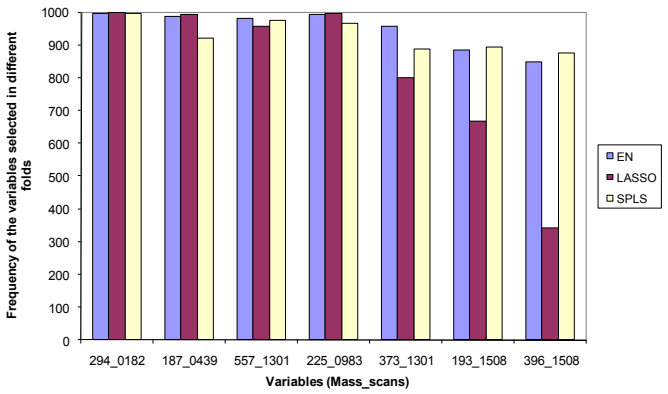


Figure 4: Frequency of the seven variables 294_0182, 187_0439, 557_1301, 225_0983, 373_1301, 193_1508 and 396_1508 over different folds in EN, LASSO and SPLS are shown. Variable 294_0182 shows highest number of times selected in all the methods where as 396_1508 selected lowest numbers of times.

coefficient. For correlations lower than one this effect that one variable absorbs the regression coefficient of its correlated partner also happens, but usually both are still selected. Therefore, in a situation with strongly correlated variables, it is possible that all or a subset of the correlated variables are selected together.

Regarding the grouping of correlated variables: five variables, 373_1508, 396_1508, 374_1508, 193_1508 and 212_1508, are correlated with different correlation coefficients. Among these, 373_1508 and 374_1508 had a correlation coefficient of 0.97 (highest) and 193_1508 and 212_1508, a correlation coefficient of 0.74 (lowest). EN selects 396_1508 and 193_1508 with a correlation coefficient of 0.86, whereas SPLS selects 373_1508, 396_1508, 374_1508 and 193_1508.

The variables, 535_1301, 373_1301 and 557_1301, are also correlated with each other, with the highest correlation coefficient (0.94) between 535_1301 and 373_1301 and the lowest correlation coefficient (0.84) between 373_1301 and 557_1301. EN selects only two (373_1301 and 557_1301), whereas SPLS selects three of them.

SPLS performs better for selecting groups of correlated variables, when compared to EN in the sense of selecting a larger number of correlated variables (simulation results, not shown).

If we evaluate the ranking methods (RR, PCR, PLS and RF), we see that mass_scan 294_182 showed the highest absolute standardized regression coefficient in the different methods used and also the highest variable importance in RF. LASSO selected 24 variables, which were ranked in decreasing order of the absolute standardized regression coefficients (Table 2). Within these 24 variables, the top 18 from PLS, top 12 from PCR, top 18 from RR, top 12 from RF, top 17 from EN, top 7 from SPLS and top 12 from univariate regressions were included. Variables like 557_1301 and 225_0983 obtained high ranks in all methods. Standardized regression coefficients of PLS, PCR and RR were more or less similar (Supplementary Figure S3). The standardized regression coefficients of RR and PLS are more similar than the regression coefficients of PCR. The correlation coefficient of standardized regression coefficients between PCR and PLS is 0.85, between RR and PCR is 0.85, between RR and PLS 0.95. These results confirm the observation of Hastie et al. [29], in saying that "PLS, PCR and RR tend to behave similarly. Variable selection methods rather than non-selection methods here performed better in terms of the MSEP. This could be due to the fact that the variables which are not associated with the trait (noise variables) get regression coefficients with the value zero, so that they effectively drop out of the regression model. In the simulation studies where only 12 out of 1000 variables had a true relationship with the response, usually all twelve true predictors were selected, while hardly any of the 'noise' variables were selected. The estimated regression coefficients of the true predictors were close to the true effect, which would explain the good performance in prediction. For noise variables that were occasionally selected, the regression coefficients were much lower than those of the set of 12 true predictive variables. For the regularized regression methods that do not perform variable selection, but where all variables remain in the model, it is expected that true predictors will rank higher than noise variables (as observed in the simulation results), but due to the non-zero regression coefficients of large numbers of regression coefficients, the predictions are expected to be less good.

We implemented all methods used here in an intuitive web-based interface Omics Fusion which offers non-statisticians to easily analyze their own data using the statistical approaches described in this manuscript. In addition, Omics Fusion allows end users to analyze data with more than one approach and summarizes the results of each method in a table which is easy to interpret. So, as an end user, this application serves as a web based omics analysis tool, which produces results in terms of ranking and selection among the ~omics variables for prediction of a phenotypic trait of interest, using a variety of different methods, and it provides an easy summary of the most important results. The results table can be exported to Excel for further analysis and visualization.

In this paper, we have applied regression methods relating a phenotypic trait of interest as the response with a metabolomics data set, but the same methodology can be used in prediction of quantitative variables from other ~omics data sets as transcriptomics or proteomics data, where also the numbers of samples (n) is usually much smaller than the number of variables (p). In addition, these prediction methods can also be applied in the context of genomic selection [43], where prediction of phenotype is done from large data sets of molecular markers [44].

Acknowledgements

The authors are grateful to Prof. Cajo ter Braak, Biometris, Wageningen UR, for his helpful discussions and valuable input.

Author Disclosure Statement

The authors declare no conflicting financial interests.

References

1. Brazma A, Vilo J (2000) Gene expression data analysis. FEBS Lett 480: 17-24.
2. Gaasterland T, Bekiranov S (2000) Making the most of microarray data. Nat Genet 24: 204-206.
3. Fiehn O (2002) Metabolomics — The link between genotypes and phenotypes. Plant Mol Biol 48: 155-171.
4. Dunn WB, Bailey NJC, Johnson HE (2005) Measuring the metabolome: Current analytical technologies. Analyst 130: 606-625.
5. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198-207.
6. Patterson SD, Aebersold RH (2003) Proteomics: The first decade and beyond. Nat Genet 33: 311-323.
7. Zhu H, Bilgin M, Snyder M (2003) Proteomics. Ann Rev Biochem 72: 783-812.
8. Joyce AR, Palsson B (2006) The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol 7: 198-210.

9. Kiers H, Smilde A (2007) A comparison of various methods for multivariate regression with highly collinear variable. Stat Method Appl 16: 193-228.

10. Brown CR, Kim TS, Ganga Z, Haynes K, De Jong D, et al. (2006) Segregation of total carotenoid in high level potato germplasm and its relationship to beta-carotene hydroxylase polymorphism. Amer J Potato Res 83: 365-372.

11. Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for non-orthogonal problems. Technometrics 12: 55-67.

12. Tibshirani R (1996) Regression shrinkage and selection *via* the LASSO. J Royal Stat Soc 58: 267-288.

13. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Statist Soc B 67: 301-320.

14. Massy WF (1965) Principal components regression in exploratory statistical research. J Amer Stat Assoc 60: 234-256

15. Wold H (1975) Soft modeling by latent variables; the nonlinear iterative partial least squares approach. Persp Prob Stat, Papers in Honour of M. S.Bartlett, Gani J (Ed.), Academic Press, London, UK.

16. Chun H, Keles S (2009) Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics 182: 79-90.

17. Vapnik VN (1995) The Nature of Statistical Learning Theory. Springer, New York, USA.

18. Breiman L (2001) Random forests. Machine Learning 45: 5-32.

19. Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23: 2507-2517.

20. Hendriks MM, Smit S, Akkermans WL, Reijmers TH, Eilers PH, et al. (2007) How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. Proteomics 7: 3672-3680.

21. Bovelstad HM, Nygard S, Størvold HL, Aldrin M, Borgan O, et al. (2007) Predicting survival from microarray data - a comparative study. Bioinformatics 23: 2080-2087.

22. Bovelstad HM, Nygard S, Borgan O (2009) Survival prediction from clinico-genomic models -A comparative study. BMC Bioinformatics 10: 413.

23. Menendez P, Eilers P, Tikunov Y, Bovy A, Eeuwijk, FV (2012) Penalized regression techniques for modeling relationships between metabolites and tomato taste attributes. Euphytica 183: 379-387.

24. Celis-Gamboa C, Struik PC, Jacobsen E, Visser RGF (2003) Temporal dynamics of tuber formation and related processes in a crossing population of potato (*Solanum tuberosum*). Ann Appl Biol 143:175-186.

25. Celis-Gamboa BC (2002) The life cycle of the potato (*Solanum tuberosum* L.): from crop physiology to genetics. Ph.D. Thesis, Wageningen University, Wageningen, The Netherlands.

26. Werij JS, Kloosterman B, Celis-Gamboa C, de Vos CH, America T, et al. (2007). Unravelling enzymatic discoloration in potato through a combined approach of candidate genes, QTL, and expression analysis. Theor Appl Genet 115: 245-252.

27. Montgomery CD, Peck EA (1992) Introduction to linear regression analysis. Wiley, New York, USA.

28. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57: 289-300.

29. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: Data mining, inference, and prediction. (2nd Edn), Springer, New York, USA.

30. Jolliffe IT (1982) A note on the use of principal components in regression. J Royal Stat Soc 31: 300-303.

31. Geladi P, Kowlaski B (1986) Partial least square regression: A tutorial. Analytica Chimica Acta 35: 1-17.

32. Hoskuldson A (1988) PLS regression methods. J Chemometrics 2: 211-228.

33. Robert A Wilson, Frank C Keil (2001) The MIT Encyclopedia of the Cognitive Sciences, MIT Press, USA.

34. Demiriz A, Bennett KP,Breneman CM, Embrechts MJ (2001) Support vector machine regression in chemometrics. Comp Sci Stat 33: 289-296.

35. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge Universtity Press, UK.

36. Segal MR (2004) Machine learning benchmarks and random forest regression. Technical Report. Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco.

37. Smit S, Van Breemen MJ, Hoefsloot HC, Smilde AK, Aerts JM, et al. (2007). Assessing the statistical validity of proteomics based biomarkers. Analytica Chimica Acta 592: 210-217.

38. Stone JR (1974) Cross-validatory choice and assessment of statistical predictions. J Royal Stat Soc 36: 111-147.

39. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7: 91.

40. Stallard BR, Garcia MJ, Kaushik S (1996) Near-IR reflectance spectroscopy for the determination of motor oil contamination in sandy loam. Appl Spectroscopy 50: 334-338.

41. Mevik BH, Cederkvis HR (2004) Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J Chemometrics 18: 422-429.

42. Acharjee A, Kloosterman B, de Vos RCH, Werij JS, Bachem CWB, et al. (2011) Data integration and network reconstruction with omics data using Random Forest regression in potato. Analytica Chimica Acta 705: 56-63.

43. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using Genome-Wide dense marker maps. Genetics 157: 1819-1829.

44. Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: A comparison of models. Crop Sci 52:146-160.