Singh *et al. BMC Genomics* (2016) 17:324
DOI 10.1186/s12864-016-2651-0

BMC Genomics

METHODOLOGY ARTICLE Open Access

Dissimilarity based Partial Least Squares (DPLS) for genomic prediction from SNPs

Priyanka Singh[1,2], Jasper Engel[2], Jeroen Jansen[2], Jorn de Haan[1] and Lutgarde Maria Celina Buydens[2*]

Abstract

Background: Genomic prediction (GP) allows breeders to select plants and animals based on their breeding potential for desirable traits, without lengthy and expensive field trials or progeny testing. We have proposed to use Dissimilarity-based Partial Least Squares (DPLS) for GP. As a case study, we use the DPLS approach to predict Bacterial wilt (BW) in tomatoes using SNPs as predictors. The DPLS approach was compared with the Genomic Best-Linear Unbiased Prediction (GBLUP) and single-SNP regression with SNP as a fixed effect to assess the performance of DPLS.

Results: Eight genomic distance measures were used to quantify relationships between the tomato accessions from the SNPs. Subsequently, each of these distance measures was used to predict the BW using the DPLS prediction model. The DPLS model was found to be robust to the choice of distance measures; similar prediction performances were obtained for each distance measure. DPLS greatly outperformed the single-SNP regression approach, showing that BW is a comprehensive trait dependent on several loci. Next, the performance of the DPLS model was compared to that of GBLUP. Although GBLUP and DPLS are conceptually very different, the prediction quality (PQ) measured by DPLS models were similar to the prediction statistics obtained from GBLUP. A considerable advantage of DPLS is that the genotype-phenotype relationship can easily be visualized in a 2-D scatter plot. This so-called score-plot provides breeders an insight to select candidates for their future breeding program.

Conclusions: DPLS is a highly appropriate method for GP. The model prediction performance was similar to the GBLUP and far better than the single-SNP approach. The proposed method can be used in combination with a wide range of genomic dissimilarity measures and genotype representations such as allele-count, haplotypes or allele-intensity values. Additionally, the data can be insightfully visualized by the DPLS model, allowing for selection of desirable candidates from the breeding experiments. In this study, we have assessed the DPLS performance on a single trait.

Keywords: Bacterial wilt, Genomic prediction, Phenotype prediction, Genetic distance, Dissimilarity based Partial Least Squares

Background

Genome wide association studies (GWAS) have been widely applied in human, plant and animal studies to identify genetic variants associated with complex traits [1–3]. In GWAS, the association between SNPs and a complex trait is usually analyzed by testing each marker individually. This requires a large number of significance tests. Because of this, a stringent *p-value* is generally used to select significant SNPs to reduce the number of false positive SNPs. At the same time, however, many real associated variants may be missed. The success of GWAS relies on the underlying trait architecture [4], heritability [5–7], effective population size and environmental factors [8, 9]. There is a general consensus that complex traits are controlled by many quantitative trait loci (QTL) with small effects. Typically, single marker GWAS analyses approaches are only able to capture QTLs with large effects and miss QTLs with small effects [1]. In addition, these significantly identified SNPs account for only a small fraction of the variation of the complex traits. Instead of GWAS where the target is to identify SNPs associated with a complex trait, one can also use

* Correspondence: l.buydens@science.ru.nl
[2]Radboud University Nijmegen, Institute for Molecules and Materials, Nijmegen, The Netherlands
Full list of author information is available at the end of the article

SNPs to predict complex traits by fitting all SNPs simultaneously to select individuals. Prediction of a quantitative trait using all SNPs is referred as genomic prediction (GP) [10]. Breeders can use predicted trait values to select candidates for their future breeding programs, termed as genomic selection (GS) [11]. It has been shown that GP provides a cost-effective and time-efficient tool for breeders to predict traits, which may be difficult and expensive to measure directly, are limited to sex or only observable later in life. GP has been successfully applied in selection of breeding candidates in plants [12–14] and animals [15–17]. The approach has also been applied in humans to predict disease risk and many complex traits [1, 18]. Several studies have shown that the prediction accuracy of complex traits may be improved by using all SNPs simultaneously [1, 9, 12–15]. For example, Yang et al. [1] have shown that fitting all SNPs simultaneously leads to approximately ten-fold increase in the predictive ability of human height compared to the individual SNP.

The accurate prediction of a complex trait can be extremely challenging, as the trait may be affected by multiple loci that interact. Another major challenge is the fact that the number of SNPs in GP studies greatly exceeds the number of samples leading to the so-called 'large-p - small-n' problem. Because of this, many traditional statistical approaches are not applicable to such data. Deriving an accurate prediction of complex traits by the high-density SNPs, while at the same time taking into account possible interactions between multiple loci, requires powerful feature reduction methods. A variety of methods such as Bayesian regression [19–21], Genomic Best-Linear Unbiased Prediction (GBLUP) [19, 22, 23], kernel regression [24] and dimension reduction methods [25] have been developed and applied in GP.

Most of the suggested statistical models differ in their assumptions of the distribution of the SNPs effect. For example, the GBLUP model assumes that the SNPs effect size is drawn from a common Gaussian distribution and the variances of SNPs effect are equal. This unrealistic assumption of GBLUP corresponds to use of a single random effect term in the model, which is a severe and unnecessary limitation. Whereas the Bayesian methods [21, 26] assume that the variance of SNPs effect differs among loci with most of the SNPs having a zero to low effect and only a few having moderate to large effect. Several studies have shown that Bayesian regression outperformed GBLUP for the prediction of traits with few QTLs with large effect [4, 27, 28].

Multiple Linear Regression (MLR) is an often used quantitative technique for prediction from predictors [29, 30]. However, MLR can be applied for prediction when the number of independent variables does not significantly exceed the number of observations and no significant collinearity between predictors exists [31]. Considering the characteristics of genomic data, MLR is not directly applicable for GP. Partial Least Squares (PLS) [32, 33] may overcome these issues for high-dimensional and collinear data by combining the principles of Principal Component Analysis (PCA) and MLR. It has been successfully applied in metabolomics for analysis of high-dimensional chromatography, and mass spectrometry data [34, 35]. PLS tries to extract latent variables (LVs) that combine SNPs to optimally predict a dependent variable such as a complex trait, taking into account their mutual correlation. However, PLS cannot be directly used for SNPs, as these are generally discrete (often represented as 0, 1, 2 for bi-allelic SNPs) while conventional PLS has been developed for the analysis of continuous data.

Therefore, we propose to use Dissimilarity-based Partial Least Squares (DPLS) to predict one or multiple traits from a large set of SNPs. In DPLS, measurements of the dissimilarity between the accessions (instead of the raw SNPs) are used for prediction. Because of this, DPLS may also be used for GP, when the method is used in combination with a suitable measure of the genomic distance between genotype accessions. Note that, during the dissimilarity calculation (between accessions), SNPs information is lost, which means effect of SNPs on the traits cannot be directly calculated from DPLS. However, there is a pseudo-sample technique proposed in literature to extract variables interaction effect from DPLS model [36]. Unlikely PLS, which uses PCA-like technique to extract LVs, DPLS takes the advantages of Multi-Dimensional Scaling (MDS)-like technique to extract LVs to predict complex trait. Both PCA and MDS techniques are widely used for dimension reduction purposes. MDS uses a distance matrix and is often recommended to analyze distance matrices. MDS minimize the dimensions, while preserving actual distance between data points. The DPLS combines features of MDS and PLS in order to perform GP. Several measures have been developed and proposed in literature to calculate dissimilarity between genomic accessions from SNPs [37, 38]. As a case study, we have explored and compared eight of such widely used genomic distance measures (Table 1) in combination with DPLS to predict bacterial wilt (BW) in tomato. In this study we have used SNPs as predictors to predict BW. BW is a complex trait caused by bacteria *Ralstonia solanacearum* and is considered as one of the most destructive diseases for a wide range of crops, including tomato [39]. In this analysis we have not accounted for any environmental factor in the prediction model. We have focused this study on genotype effects, by comparing accessions grown in the same controlled greenhouse environment.

The prediction quality (PQ) of the DPLS model was measured in term of R2 estimated from observed and predicted trait values in a cross validation (CV) setup. Furthermore, we have compared the prediction performance

Table 1 The selected dissimilarity measures used to calculate genomic distance among tomato accessions from SNPs[a]

Distance	Equation	R-packages	References
Euclidean	$\mathbf{d_{i1i2}} = \sqrt{\sum_{k=1}^{K} (\mathbf{x_{i1k}} - \mathbf{x_{i2k}})^2}$	gstudio	[65]
Gower	$\mathbf{d_{i1i2}} = \frac{\sum_{k=1}^{K} \delta_{i1i2k} * d_{i1i2k}}{\sum_{k=1}^{K} \delta_{i1i2k}}$ For nominal or factor variables $\mathbf{d_{i1i2k}} = \mathbf{0}$, $(\mathbf{if}\ \mathbf{x_{i1k}} = \mathbf{x_{i2k}})$ $\mathbf{d_{i1i2k}} = \mathbf{1}$, $(\mathbf{if}\ \mathbf{x_{i1k}} \neq \mathbf{x_{i2k}})$	daisy	[66]
Allele share	$\mathbf{D}_{i1i2} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{d_{i1i2}}(\mathbf{k})$ Where $\mathbf{d_{i1i2}(k)}$ = {0, If individual $\mathbf{i_1}$ and $\mathbf{i_2}$ have two alleles in common at the $\mathbf{k^{th}}$ locus, 1, If individual $\mathbf{i_1}$ and $\mathbf{i_2}$ have only single alleles in common at the $\mathbf{k^{th}}$ locus, 2, If individual $\mathbf{i_1}$ and $\mathbf{i_2}$ have no alleles in common at the $\mathbf{k^{th}}$ locus}	Custom-R-script	[67]
Nei	$\mathbf{d_{nei}} = -\mathbf{ln}\left[\frac{(2N-1)\sum_{i=1}^{L}\sum_{j=1}^{l} p_{ij,x}p_{ij,y}}{\sqrt{\sum_{i=1}^{L}\left(2N\sum_{j=1}^{l} p_{ij,x}-1\right)\left(2N\sum_{j=1}^{l} p_{ij,y}-1\right)}}\right]$ Where, the summation L is across loci and l is across alleles at each locus in population x and y (here individual)	gstudio	[68]
Bray	$\mathbf{d_{i1i2}} = \frac{\sum_{k=1}^{K} \lvert x_{i1k} - x_{i2k} \rvert}{\sum_{k=1}^{K} x_{i1k} + x_{i2k}}$	vegan	[69]
Jaccard	$\mathbf{d_{i1i2}} = \frac{2B}{(1+B)}$	vegan	[70]
Kulczynski	$\mathbf{d_{i1i2}} = \mathbf{1} - \mathbf{0.5} * \left[\frac{\sum_{k=1}^{K} \min(x_{i1k}, x_{i2k})}{\sum_{k=1}^{K} x_{i1k}} + \frac{\sum_{k=1}^{K} \min(x_{i1k}, x_{i2k})}{\sum_{k=1}^{K} x_{i2k}}\right]$	vegan	[70]
GRM	$\mathbf{G} = \frac{ZZ'}{2\sum \mathbf{p_k}(\mathbf{1} - \mathbf{p_k})}$	Custom R-script	[22]

x_{i1k} and x_{i2k} = SNPs at locus k for accession x_{i1} and x_{i2} respectively
d_{i1i2k} = distance between i_1 and i_2 samples for SNPs at locus k
B Bray- Curtis dissimilarity
G Genomic relationship matrix
Z genotype information for all tomato accessions
p_k frequency of allele at locus k
[a] d_{i1i2} = distance between tomato accession i_1 and i_2

of DPLS with a prediction based on SNPs found as significant in a Univariate association analysis (where SNP was used as fixed effect) and GBLUP (where SNP was used as random effect). We demonstrate GP with DPLS on a single trait. The method can however be applied to simultaneously predict multiple, possibly correlated traits.

Results and discussion

The differences in information captured by various genomic distance measures for GP have not been explored. Therefore, we first explored the properties of the eight genomic distance measures of interest to this study by the Mantel test, heatmap visualization, and their application in Multi-Dimensional Scaling (MDS) before we studied their application in DPLS.

Comparison of genomic information captured by different distance measures

Mantel correlation

The Mantel test was used to compare the relation between two distance matrices in terms of correlation (*r*) statistics. The pair-wise correlation results obtained from the comparison of genomic distance measures by the Mantel test are presented in Table 2. On the basis of the Mantel correlation statistics, the eight genomic distance measures can be grouped into two categories for the data investigated in this study. Any two genomic distance matrices, which show a Mantel's test correlation > 0.70, are placed together in one group. The first group (hereafter Group-I) includes Euclidean, Gower, Nei and allele-share distance and the other four genomic distances i.e.,

Table 2 Summarized Mantel correlation statistics for analyzed genomic dissimilarity matrices*

Distances		Euclidean	Gower	Allele Share	Nei	Bray	Jaccard	Kulczynski	GRM
Euclidean	GROUP I	1							
Gower		0.95	1						
Allele share		0.95	0.95	1					
Nei		0.94	0.98	0.98	1				
Bray	GROUP II	0.55	0.46	0.46	0.42	1			
Jaccard		0.54	0.44	0.44	0.39	0.98	1		
Kulczynski		0.36	0.24	0.24	0.19	0.96	0.95	1	
GRM		0.53	0.43	0.48	0.44	0.85	0.92	0.78	1

**The cells are gray scaled according to correlation (r) values. The summary shows that the distance measures can be classified to two groups: Group I (consisting of Euclidean, Gower, Allele-share and Nei distances) and Group II (consisting of Bray, GRM, Jaccard and Kulczynski distances)*

Bray, Kulczynski, Jaccard and genomic relationship matrix (GRM) were placed in second group (hereafter Group-II).

Heatmap visualization

In Fig. 1, the quantitative distance patterns between the 242 tomato accessions are visualized as a heatmap for the Euclidean and Bray distance. Both heatmaps show many sub-clusters of closely related tomato accessions. However, in contrast to the heatmap of the Euclidean distance (Group-I distance) the heatmap of the Bray distance (Group-II distance) shows many small clusters. The heatmap plot clearly shows that tomato accessions cluster differently in Euclidean and Bray distance space. In Fig. 1, the Euclidean and Bray distance were selected as representative distances of Groups-I and II identified by the Mantel test. Similar results were observed for the other distance measures within each group.

Multi-Dimensional Scaling (MDS) analysis

MDS models based on the two selected representative distance measures (i.e., Bray and Euclidean) were used to visualize the relations between the tomato accessions in a scatter plot (Fig. 2). The MDS plot of the Euclidean distance matrix suggests that most tomato accessions are genetically similar and form a big cluster with few smaller clusters of genetically less similar accessions. In the analogous representation of the Bray distances, tomato accessions were distributed throughout the entire plot with few accessions forming small clusters in MDS space.

In both plots the observed clusters do not clearly relate to the phenotype. We conclude that the two groups of distance metrics represent different structures within the genotype data, neither of which can be strongly associated to the phenotype measures by MDS. However, this insight could only be obtained from the MDS score plot that represents the relative differences and similarities between all accessions used in the study.

Phenotype prediction with DPLS

In this study, DPLS was used to relate genomic information captured by the distance measures indicated in Table 1 to the BW. For each distance measure repeated double cross validation (rDCV) (see Methods) was used to choose the optimal number of latent variables (LVs) to fit the

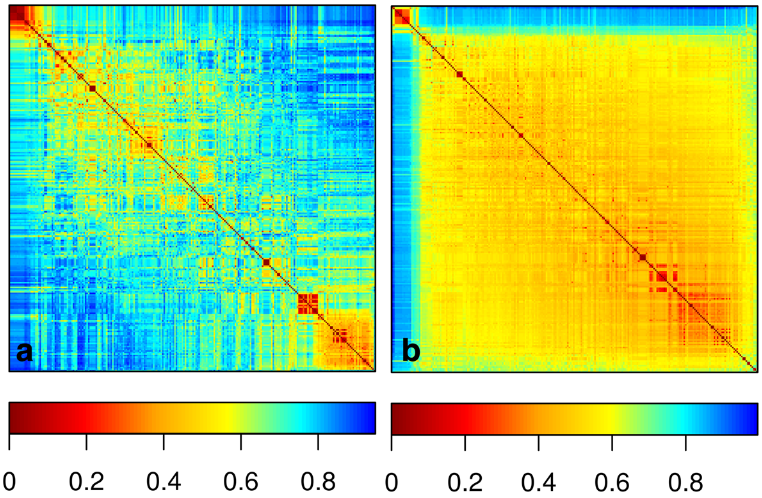

Fig. 1 Heatmap representation of dissimilarity scores between 242 tomato accessions for Bray (**a**) and Euclidean (**b**). The pixels are colored in proportion to the genotypic dissimilarity between tomato accessions. Euclidean and Bray heatmap represents distance group-I and II respectively

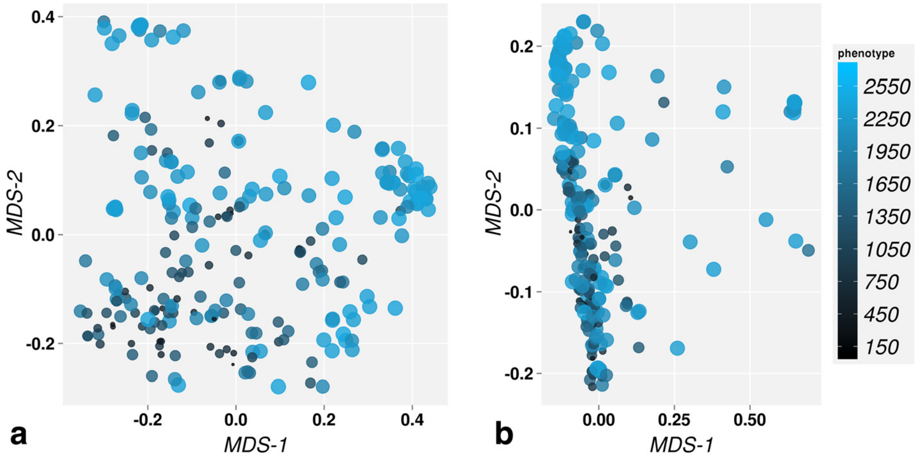

Fig. 2 Multi-dimensional scaling (MDS) scores representation of Bray (**a**) and Euclidean (**b**) distances. MDS scores are visualized in first two dimension of MDS space, where MDS1 and MDS2 represents scores in first and second dimension respectively. The size and colors of bubbles are in proportion to the actual trait values (the measure of resistance against Bacterial wilt) of tomato accessions. The bigger bubble size represents higher resistance accessions

DPLS model (see Table 3). As explained in the Methods section (equation 3) a big advantage of the DPLS method is that it also returns so-called score values for each accession. These scores represent the relative position of each accession in terms of their genomic distances associate to the trait values. As shown in Fig. 3, these scores can be visualized in a plot similar to the MDS analysis (Fig. 3), where large distances between accessions in the plot (the different dots) indicate large genomic differences. The score plots show a better arrangement of tomato accessions in the space of DPLS LVs when compared to the original data structure observed in exploratory analysis by MDS (see Fig. 2). The accessions are arranged according to their trait values; tomato accessions with similar trait values are close together in DPLS LVs space. The DPLS model is therefore better able to predict the trait values from the genotype dissimilarity scores data. Although the score plots of DPLS models based on different distance measures differ considerably, a direction within the space along which the trait value increases can be identified in both score plots.

Table 3 Dissimilarity based partial least squares (DPLS) prediction results over all dataset in a 10-fold CV setup

Distance	PQ[c] (R^{2d})	RMSE[a]	Optimal LVs[b]
Euclidean	0.62 ± 0.005	370 ± 2.7	4
Gower	0.60 ± 0.0052	380 ± 2.8	6
Allele share	0.61 ± 0.005	380 ± 2.8	6
Nei	0.59 ± 0.005	390 ± 2.9	6
Bray	0.63 ± 0.004	370 ± 2.6	4
Jaccard	0.64 ± 0.0043	360 ± 2.8	4
Kulczynski	0.61 ± 0.0053	380 ± 2.8	4
GRM	0.62 ± 0.005	370 ± 2.9	5
GBLUP	0.61 ± 0.001	369.9 ± 0.66	NA

All the results presented in table are significant (with respect to *p*-value computed from permutation analysis). The results are averaged over 10-fold CV scheme. The 10-fold CV procedure was repeated 50 times. The standard error (se) calculated over 10-fold CV repetition. The last row present prediction results obtained from GBLUP. The PQ (R^{2d}), RMSE and LVs represents prediction quality, root mean square error and latent variables respectively

[a] *RMSE* stands for root mean square error

[b] *LVs* stands for latent variables used for model building

[c] *PQ* represent prediction quality

[d] R^2 presented in the table are estimated for testset and not from training model. The value is calculated in a cross validation setup (some time indicated as Q^2). This value is refer as prediction quality in this study

The DPLS prediction performance with respect to each distance measure is presented in Table 3. The performance statistics for each distance measure consist of PQ (i.e. R^2 estimated on testset), model error measured in terms of Root Mean Square Error (RMSE) and the optimal number of LVs used for building prediction models. PQ range from 0.59 to 0.64 for analyzed genomic distance measures. The resulting RMSE was found to be similar for each genomic distance measure. These results (Table 3) indicate that the DPLS models may predict the trait similarly well between all distance measures for the studied BW-tomato data. The correlation (r) between measured and predicted BW values is visualized in Fig. 4 (hereafter prediction plot). The prediction plot shows a linear trend for the prediction based on both the distances (i.e., Bray and Euclidean). However, it seems from the prediction plot (Fig. 4) that accessions with higher BW values were predicted better than accessions with lower values. This follows from least-squares criterion used within DPLS, which it shares with conventional PLS, MLR and most other 'conventional' data analysis methods. This criterion gives more importance to the prediction of more distant accession. The heritability of the trait has also direct

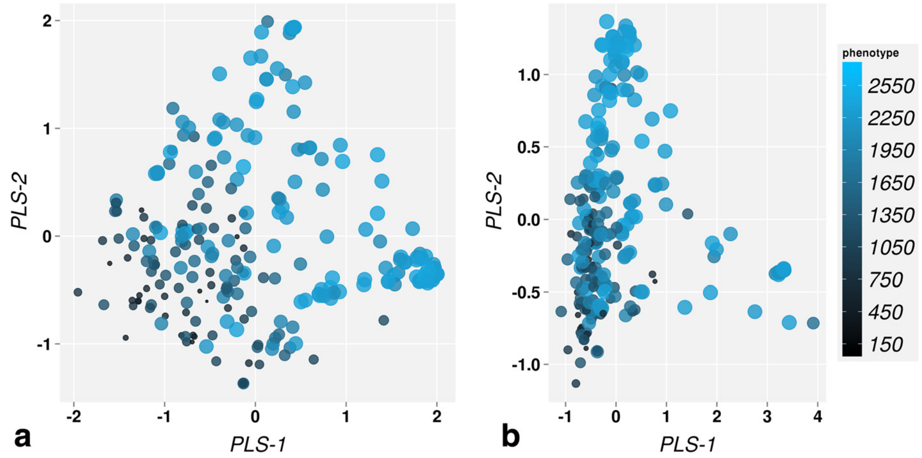

Fig. 3 Dissimilarity based partial least squares (DPLS) scores representation for Bray (**a**) and Euclidean (**b**) distances. The DPLS scores are visualized in first two latent variables (LVs) where PLS1 and PLS2 present scores in first and second LVs space respectively. Each bubble represents a tomato accession. The size and color of the bubbles are corresponding to the actual trait values of tomato accessions where bigger size of the bubble corresponds to higher resistance accession. The PVE represents phenotypic variance explained by the DPLS prediction model

influence on GP. This is considered as theoretical upper limit for prediction accuracy and maximum variance explained due to genetic effects [8]. The estimated heritability for the BW is 0.76. The prediction results PQ (which is also a measure for variance explained in test set from prediction model) in Table 3, indicates that the variance explained by the DPLS model for the BW is close to the upper limit set by the estimated BW heritability.

Phenotype prediction with GBLUP

The performance of DPLS was compared to that of GBLUP in a 10-fold double cross-validation setup. The BW prediction results from the GBLUP model are indicated in Table 3. The trait can be predicted similarly well by GBLUP and DPLS. However, GBLUP does not provide any visual representation of the relations between individual accessions. A disadvantage of DPLS compared to GBLUP is that it does not provide information about SNPs effect directly. The genotype-based distance matrix from DPLS implies that information on individual SNP is lost during the modeling, which is retained by GBLUP. However, approaches such as pseudo-sampling [36, 40, 41] are available to interpret the effect of each individual SNP in GP. From this, the SNPs most relevant to the DPLS model can be obtained. However, Pseudo-sampling has not yet been applied to high-dimensional genomic data. Therefore this will be subject of a future study.

Phenotype prediction based on single SNP analysis

In previous studies using Univariate models for analysis of the BW-tomato data, 29 SNPs were found to be significantly associated with the studied BW (Unpublished). The phenotypic variance predicted from each SNP ranges from ~0.04 to ~50 %, which is lower than the variance explained by DPLS (59–64 %). This analysis clearly shows DPLS prediction has edge over single-SNP regression

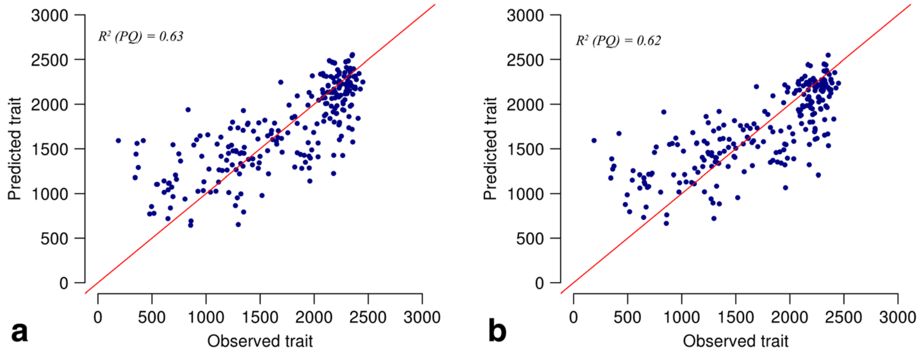

Fig. 4 Dissimilarity based partial least squares (DPLS) prediction plot for Bray (**a**) and Euclidean (**b**). The prediction for each accession obtained in repeated 10-fold-CV scheme. Each point indicates mean value of accession prediction. Original and predicted value of BW traits are plotted on X and Y axis respectively. The R^2 represent prediction quality and the red line indicates trend line for regression model

approach. The analysis clearly indicates that BW is a complex trait, which should be analyzed by multivariate methods that observe all SNPs within all accessions.

Advantages of DPLS as genomic prediction model

The results obtained from the proposed DPLS method indicate that its prediction performance is on par with that of GBLUP. Together with that DPLS provides some other beneficial characteristics. It can be applied to dataset of any dimension. DPLS reduces the dimensionality drastically and can handle missing values while computing distances or dissimilarities. It can handle the multicolinearity that is omnipresent in genomic data and can be easily implemented using widely available software and methodology for conventional PLS. A major advantage of DPLS over other methods is the DPLS score plot, which represents arrangement of tomato accessions in DPLS space. This visualization provides a tool for the breeders to select the optimal candidates for their future breeding program. For instance, breeders can select the specific tomato accessions from the right panel of the score plot presented in Fig. 3, as candidates to specifically breed for a BW resistant tomato variety. This score plot based on the Bray distance shows that the first two LVs of DPLS explain about 56 % of variation in the phenotype. Additionally, the arrangement of tomato accessions with respect to the trait values in the plot shows that there seem to be at least two discrete groups of disease resistant accessions in the dataset. By inspecting this score plot, breeders can select candidates from both groups to breed for resistant varieties, to grasp more trait variability than with selection based on high resistant accessions. The score plot therefore enables, selection of multiple germplasm sources, which would be impossible if a single phenotype summary such as Estimated Breeding Values (EBVs) or other transformations of the trait. No existing method for GP provide such scores to compare individual accessions.

The other advantage of DPLS is that, it is flexible to various genotype representations. For Example, SNPs are usually encoded as discrete variables (i.e. 0, 1, 2 or 1, 0, −1) and many models uses this encoding as standard input for GP. The DPLS prediction model does not rely on such standard genotype representation since a distance matrix between accessions is used as input for the model. This makes it more flexible to data representation and may possibly be better applicable for GP in diploid or polyploid crops. The approach can also be very useful for analyzing complex phenotypes which are often collected in form of multiple traits to gather more information [42]. These traits are generally correlated and share a common genetic mechanism. The analysis of multiple traits together in a multivariate model may bring more power and increase chances of detecting SNPs, which have effect on individual or multiple traits [42, 43]. However, there is limited number of methods available, which can be applicable to multivariate trait analysis [44]. Successful prediction of multivariate responses with PLS has been reported in numerous references [45, 46]. We therefore expect that the DPLS may efficiently exploit the information from high-dimension SNPs to predict multiple potentially correlated traits. Assessing the DPLS performance on simultaneous prediction of multiple traits is a topic for future study.

Conclusions

In this study DPLS, a novel approach for genomic prediction, is proposed for dealing with genomic data. This method employs the strengths of multivariate partial least squares (PLS) based prediction with the expression of genomic distances (calculated from SNPs) between individual accessions. This way, problems in the data such as the categorical nature of the variables, the large number of variables and their multi-collinearity are avoided. It was found that DPLS performs on par with GBLUP and better than Univariate prediction approach for GP. The prediction performance of the proposed method was close to the biologically imposed upper limit boundary set by the heritability of the trait. DPLS allows for visualization of the accessions with respect to the trait of interest, which may be invaluable for selection of specific candidates in agricultural breeding programmes.

Methods

Genotype

329 tomato breeding accessions were genotyped for 7321 SNPs using “SolCAP” array. The data was kindly provided by Bezo Zaden and East–west Seed. The SNPs are distributed across 12 chromosomes of tomato genome. A quality check (QC) was performed on the genotype data to exclude low quality SNPs and accessions from the analyses. SNPs were excluded if minor allele frequency (MAF) <0.01, proportion of missing value (PMV) > 0.10 or both. Tomato accessions with genotypes, but not phenotyped were excluded from the analyses. 242 tomato accessions and 6517 SNPs remains after the quality filtering and were used for GP.

Phenotype

BW disease is caused by the *Ralstonia solanacearum* bacteria. It is one of the most destructive crop diseases in tropical, subtropical and some warm temperate regions of the world [39]. BW is a complex trait. In the current analysis, the trait was measured as percentage wilted plants at several time points (14, 21 and 28 days after disease inoculation, four replicates per accession), under greenhouse conditions. Based on the percentage-wilted plants, an area under disease progression curve (AUDPC) was calculated. Thus AUDPC is quantitative summarization of BW disease intensity over time. In this analysis we have focused on the interval between 0 and 28 days because measurements in this interval were available for most accessions. The AUDPC of the BW was calculated using equation below

$$\mathbf{AUDPC} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \frac{\left(\mathbf{y_i} + \mathbf{y_{i+1}}\right)}{\mathbf{2}} \left(\mathbf{t_{i+1}} - \mathbf{t_i}\right)$$

Where $\mathbf{n}$ = total number of observation, $\mathbf{y_i}$ = percentage wilted plat at the $\mathbf{i^{th}}$ observation and $\mathbf{t}$ = time at the $\mathbf{i^{th}}$ observation.

Genomic distance measurement

Eight dissimilarity measures (see Table 1) were explored and analyzed to find the most appropriate distance measure to use for genomic prediction. These measures were used to calculate distances between the tomato accessions based on SNPs. Furthermore, these calculated distances between tomato accessions were used in a DPLS model to predict BW. The evaluation of distances measures was based on the trait prediction accuracy for BW with the DPLS model. Additionally, initial exploratory analysis with the Mantel test [47, 48], heatmap visualization and MDS analysis were performed first on the distance measures to understand the correlation between the selected measures and their respective behaviors in the dataset.

Exploratory analysis of distance measures

Mantel test

Relationships between distance matrices based on different genomic measures (see Table 1) were quantified using the Mantel test [48]. This statistical test constructs a linear comparison of two genomic distance matrices. The Mantel test first calculates the correlation between two distance matrices followed by a randomization procedure (permutation) to evaluate whether the observed correlation between two distance matrices is random or not [47, 49]. The Mantel test was performed with 1000 permutations, using the R package ade4 [50].

Heatmap visualization

Visual inspection of the distance matrices based on the measures in Table 1 was performed by plotting heatmaps of the genomic distances between the accessions within the BW-tomato data using the R-package ‘seriation’ [51].

Multi-Dimensional Scaling (MDS)

MDS is generally used to examine multivariate structure within a dataset by representing dissimilarity measures between specific accessions as distances in a much lower-dimensional space [52, 53]; we chose two dimensions for all MDS models. MDS analysis on genomic distance measures (see Table 1) was performed using the R-package ‘stats’ [54].

Prediction model building and validation

Dissimilarity-based Partial Least Squares (DPLS)

The aim of this study is to determine if DPLS is a viable method for genomic prediction from whole genome marker data. The DPLS method employs the strengths of PLS based prediction with the expression of genomic distances (calculated from SNPs) between individual accessions. The general goal of PLS is to predict a set of response (dependent) variables**Y** from very large set of independent variables **X** (predictors), where for the BW- tomato data example **X** is the genotype matrix and **Y** is disease trait or response vector [55]. This prediction is achieved by first extracting a set of orthogonal factors called latent variables (LVs) from the predictors set. These latent variables (LVs) are considered to have best predictive power [31, 55]. The PLS model can be represented as

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E} \tag{1}$$

$$\mathbf{Y} = \mathbf{UC^T} + \mathbf{F} \tag{2}$$

Where,

T = **X**-factors scores (analogous to principal components in PCA although they are not the same)

$$\mathbf{T} = \mathbf{XW} \tag{3}$$

U = **Y**-factors scores
P = **X**-factors loadings and
C = **Y**-factors loadings
E and **F** = matrices of residuals

The regression coefficient that relates **X** to **Y** is obtained by:

$$\mathbf{B} = \mathbf{W}\left(\mathbf{P^T W}\right)^{-1}\mathbf{C^T} \tag{4}$$

Where,

W = **X**-factors weights i.e., projections of the objects of **X**-space onto **Y**-factor scores. The decomposition of **X** and **Y** are made so as to maximize the covariance between **T** and **U**.

In DPLS [56], the original data matrix **X** (genotype matrix) is replaced by dissimilarity matrix **D** (distance between tomato accessions). A summarized overview of genomic distance measures used to fit a DPLS prediction model in this study are given in Table 1. The DPLS model can be presented in following form.

$$\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{5}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{C}^{\mathbf{T}} + \mathbf{F} \tag{6}$$

The **D** matrix is a square matrix. A double centering (i.e., subtracting row and column means of a distance matrix from its elements) has to be applied to matrix **D**. This means for DPLS the score matrix is same as the loading matrix (linear combination of predictors in **D** matrix), up to a scaling constanαt **α** to identify the model.

$$\mathbf{T} = \boldsymbol{\alpha}\mathbf{P} \tag{7}$$

The regression coefficient from the DPLS model can be calculated in similar equation as in classical PLS (see equation (4)). However, the regression coefficient **B** obtained from DPLS model is based on **D** matrix and not on **X** matrix as in PLS. The DPLS method also calculates scores for each accession. These scores are linear combinations of the predictor variables and are calculated in similar fashion as presented in equation (3). In DPLS, the predictors are the columns of the distance matrix, which contains the distance information on the samples in the dataset. The scores are calculated by using an appropriate weight matrix (**W**), which reflects the covariance structure between predictors and response variables. We have used the R-package 'pls' [57] and a custom script to perform DPLS for GP. Extensive validation methodology is available for PLS, that we adapt here for DPLS [58].

Repeated double cross validation (rDCV)

Three steps are critical when building a PLS/DPLS prediction model of 1) selection of optimal number of latent variables (LVs), and model building 2) the assessment of the overall model quality (or model reproducibility) 3) assessing significance of prediction model (model transportability).

Several approaches including cross-validation (CV) setups are recommended for selection of LVs [58–60]. Here we used a so-called Double Cross-Validation (DCV) setup. In DCV [61], the data is first split into a test and calibration set. The calibration set is then further split into training and a validation set. In DPLS, the predictors are columns of the distance matrix, which contain the distance information on the samples in the dataset. The distance matrix is square, which requires specific splits for DCV. This matrix segmentation needs to be done in a specific fashion, to truly separate the information contained within them. The distance matrix is segmented such that the calibration training sets are square, as depicted in Fig. 5. The training and validation sets were used to determine the number of latent variables with optimal model error statistics. The error rate of the model

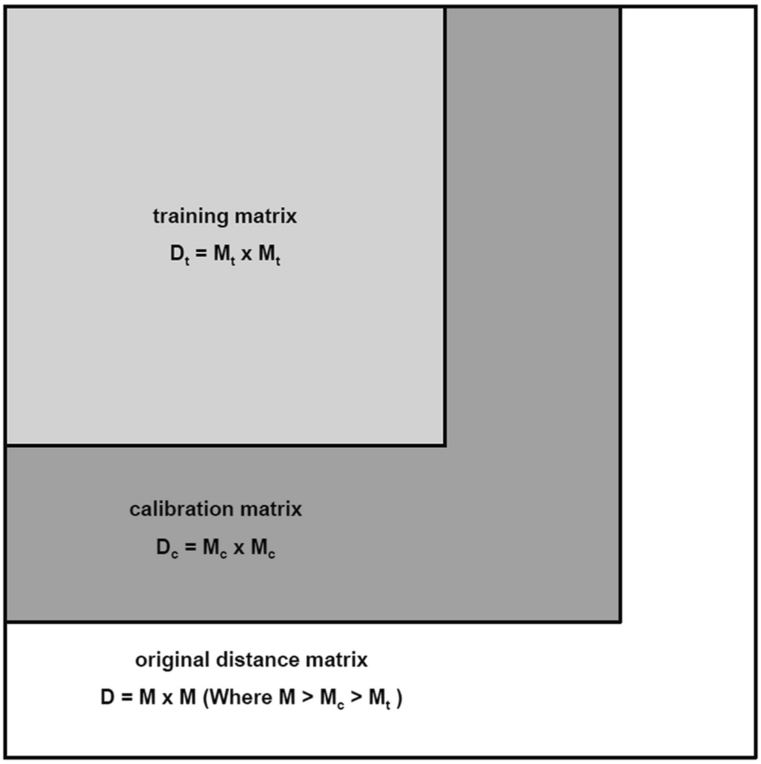

Fig. 5 Illustration of distance matrix segmentation in double cross validation. Where D, D_c and D_t are squared distance matrix and represents distance scores between total accessions (M), accessions in calibration set (M_c) and accessions in training set (M_t) respectively

with this optimal number of LVs to predict the phenotype was then determined by the test set. This procedure was repeated several times (in this case 50 times) hence in this study, we called this strategy as repeated-DCV (rDCV). The details concerning DCV/rDCV can be found elsewhere [58, 61].

Permutation analysis

Permutation analysis is often used for validation of PLS-based classification or regression [58, 61, 62]. Predictive multivariate models may be highly prone to overfit. Therefore it is common practice to assess their predictive performance against a benchmark where the effect of interest has been removed by randomization of the dependent variable, i.e. the trait value in this case. The model should not have any predictive power for such randomized data. A permutation test may be used to assess significance of reproducibility and transportability (prediction of new external validation set) of the model [63]. A permutation analysis was carried out in this study to validate the prediction accuracy observed with double-cross validation. The BW labels were permuted and randomly assigned to different accessions. A new DPLS model was then fitted to the permuted BW and the same model statistics were calculated. This procedure was repeated 5000 times and the model statistics were compared to the statistics obtained from the DPLS model on un-permuted labels. A *p*-value was obtained by combining all obtained statistics (mean and standard deviation RMSE) to assess the significance of difference between statistics (mean and standard deviation of RMSE) from original and permuted dataset.

Conventional analysis methods

Genomic best linear unbiased prediction model (GBLUP)

Genomic BLUP [22] is considered a standard statistical method for genomic prediction. Several variations of BLUP models have been proposed in literatures for genomic prediction. We have used ridge regression best linear unbiased prediction (RR-BLUP) to predict BW t from SNPs. RR-BLUP assumes that all SNPs effect are normally distributed and have equal variance [28]. The model considered is:

$$\mathbf{y} = \mathbf{1\mu} + \mathbf{Zg} + \mathbf{e}$$

Where, $\mathbf{y}$ is a vector of phenotype, μ is the overall mean, $\mathbf{Z}$ is design matrix corresponding to $\mathbf{g}$, $\mathbf{g}$ is the vector of SNP effects and $\mathbf{e}$ is the vector of residuals. It was assumed that $\mathbf{g} \sim \mathbf{N}(\mathbf{0}, \mathbf{G}\boldsymbol{\sigma}_{\mathbf{g}}^{2})$ where $\boldsymbol{\sigma}_{\mathbf{g}}^{2}$ is additive genetic variance and $\mathbf{G}$is genomic relationship matrix derived from SNP markers. These analyses were performed using the R-package rrBLUP [64].

Univariate analysis

In an unpublished study, 29 SNPs were found to be significantly associated with bacterial wilt s resistance. These SNPs were extracted from the BW-tomato data and fit in a univariate linear model. The SNPs were treated as fixed effect in this univariate model. The predictive ability (R^2) was calculated for each individual SNP and compared with the R^2 value obtained from multivariate approach DPLS.

Ethics

No human or animal samples were used in this study. As far as applicable, all experiments have been performed according to legal guidelines.

Consent to publish

The authors would like to thank Bejo Zaden and East-west Seed for access to the tomato dataset and to publish these results.

Availability of data and materials

The distance matrices to perform genomic predictions can be provided upon request. The raw genotyping data cannot be shared due to the fact that the authors are not the owners of this data. Therefore the main contribution of this paper is on the methodological part of genomic prediction.

Abbreviations

AUDPC: area under disease progression curve; DCV: double cross validation; DPLS: Dissimilarity based Partial Least Squares; GBLUP: genomic best linear unbiased prediction; GP: genomic prediction; GRM: genomic relationship matrix; GWAS: genome wide association studies; LVs: latent variables; MDS: Multi-Dimensional Scaling; PLS: Partial Least Squares; PQ: prediction quality; rDCV: repeated double cross validation; RMSE: root mean square error; SNP: single nucleotide polymorphism.

Competing interest

The authors declare that they have no competing interests.

Authors' contributions

PS and JDH collected, preprocessed and prepared the GP study. LB, JE, JJ designed the data analysis methodology. PS performed the data analysis. PS, JE, JDH, JJ and LB wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Bejo Zaden and East–west Seed for access to the tomato dataset, which was used to develop the proposed methodology. Furthermore, the authors would like to thank the Department of Analytical Chemistry, IMM, Radboud University for providing knowledge support required to carry out this research. Finally, the authors acknowledge the financial support from Genetwister Technologies BV to conduct the study.

Funding

The research work in this study was funded by Genetwister Technologies BV, Wageningen, The Netherlands. The data was provided by Bejo Zaden and East–west Seed.

Author details

[1]Department of Bioinformatics, Genetwister Technologies B.V., Wageningen, The Netherlands. [2]Radboud University Nijmegen, Institute for Molecules and Materials, Nijmegen, The Netherlands.

Received: 24 July 2015 Accepted: 22 April 2016
Published online: 04 May 2016

References

1. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–9.
2. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010;465(7298):627–31.
3. Olsen H, Hayes B, Kent M, Nome T, Svendsen M, Larsgard A, Lien S. Genome-wide association mapping in Norwegian Red cattle identifies quantitative trait loci for fertility and milk production on BTA12. Anim Genet. 2011;42(5):466–74.
4. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185(3):1021–31.
5. Hayes B, Bowman P, Chamberlain A, Goddard M. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009;92(2):433–43.
6. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
7. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet. 2008;9(4):255–66.
8. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14(7):507–15.
9. Lee SH, van der Werf JH, Hayes BJ, Goddard ME, Visscher PM. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet. 2008;4(10):e1000231.
10. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. Curr Opin Genet Dev. 2015;33:10–6.
11. Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
12. Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2014;112(1):48–60.
13. Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics. 2010;9(2):166–77.
14. Daetwyler HD, Calus MP, Pong-Wong R, Delos Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 2013;193(2):347–65.
15. Hayes B, Bowman P, Chamberlain A, Verbyla K, Goddard M. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol. 2009;41(1):51.
16. Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 2010;6(9):e1001139.
17. Daetwyler H, Hickey J, Henshall J, Dominik S, Gredler B, Van Der Werf J, Hayes B. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Anim Prod Sci. 2010;50(12):1004–10.
18. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17(10):1520–8.
19. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 2001;157(4):1819–29.
20. De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 2009;182(1):375–85.
21. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC bioinform. 2011;12(1):186.
22. VanRaden P. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
23. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013;194(3):597–607.
24. Gianola D, van Kaam JB. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008;178(4):2289–303.
25. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH. Reducing dimensionality for prediction of genome-wide breeding values. Genet Sel Evol. 2009;41(1):29.
26. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet. 2009;10(10):681–90.
27. Colombani C, Legarra A, Fritz S, Guillaume F, Croiseau P, Ducrocq V, Robert-Granié C. Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. J Dairy Sci. 2013;96(1):575–91.
28. Wang X, Yang Z, Xu C. A comparison of genomic selection methods for breeding value prediction. Science Bulletin. 2015;60(10):925–935.
29. Bączek T, Wiczling P, Marszall M, Heyden YV, Kaliszan R. Prediction of peptide retention at different HPLC conditions from multiple linear regression models. J Proteome Res. 2005;4(2):555–63.
30. Çamdevýren H, Demýr N, Kanik A, Keskýn S. Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. Ecol Model. 2005;181(4):581–9.
31. Tobias RD. An introduction to partial least squares regression. In: Proc Ann SAS Users Group Int Conf. 20th ed. Orlando, FL: Citeseer; 1995. p. 2–5.
32. Wold H. Soft modeling by latent variables: the nonlinear iterative partial least squares approach, Perspectives in probability and statistics, papers in honour of MS Bartlett. 1975. p. 520–40.
33. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta. 1986;185:1–17.
34. Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. Anal Chim Acta. 1997;348(1):71–86.
35. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002;18(1):39–50.
36. Engel J, Postma G, van Peufflik I, Blanchet L, Buydens L. Pseudo-sample trajectories for variable interaction detection in Dissimilarity Partial Least Squares. Chemometrics Intell Lab Syst. 2015;146:89–101.
37. Ickstadt K, Selinski S, Müller T. Cluster Analysis: A Comparison of Different Similarity Measures for SNP Data. In. Technical Report/ Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen; 2005.
38. Libiger O, Nievergelt CM, Schork NJ. Comparison of genetic distance measures using human SNP genotype data. Hum Biol. 2009;81(4):389–406.
39. Hayward A. Biology and epidemiology of bacterial wilt caused by Pseudomonas solanacearum. Annu Rev Phytopathol. 1991;29(1):65–87.
40. Krooshof PW, Üstün B, Postma GJ, Buydens LM. Visualization and recovery of the (bio) chemical interesting variables in data analysis with support vector machine classification. Anal Chem. 2010;82(16):7000–7.
41. Smolinska A, Blanchet L, Coulier L, Ampt KA, Luider T, Hintzen RQ, Wijmenga SS, Buydens LM. Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis. PLoS One. 2012;7(6):e38163.
42. Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. J Probab Stat. 2012;2012:13.
43. Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genet Epidemiol. 2010;34(5):444–54.
44. Maity A, Sullivan PF. Tzeng Ji: Multivariate Phenotype Association Analysis by Marker-Set Kernel Machine Regression. Genet Epidemiol. 2012;36(7):686–95.
45. Heussen PC, Janssen H-G, Samwel IB, Van Duynhoven JP. The use of multivariate modelling of near infra-red spectra to predict the butter fat content of spreads. Anal Chim Acta. 2007;595(1):176–81.
46. Galtier O, Abbas O, Le Dréau Y, Rebufa C, Kister J, Artaud J, Dupuy N. Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions. Vibrational Spectrosc. 2011;55(1):132–40.
47. Legendre P, FORTIN MJ. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Mol Ecol Resour. 2010;10(5):831–44.

48. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967;27(2 Part 1):209–20.
49. Peres-Neto PR, Jackson DA. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia. 2001;129(2):169–78.
50. Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. J Stat Softw. 2007;22(4):1–20.
51. Buchta C, Hornik K, Hahsler M. Getting things in order: an introduction to the R package seriation. J Stat Soft. 2008;25(3):1–34.
52. Schiffman SS, Reynolds ML, Young FW, Carroll JD. Introduction to multidimensional scaling: Theory, methods, and applications. New York: Academic press; 1981.
53. Young FW. Multidimensional scaling: History, theory, and applications. Lawrence Erlbaum Associates, Inc., Publishers, 356 Broadway, Hillsdale, New Jersey: Psychology Press; 2013.
54. Team RC. The R Stats Package. Vienna, Austria: R Foundation for Statistical Computing; 2002. Available from: https://www.r-project.org/.
55. Abdi H. Partial least square regression (PLS regression). In: Lewis-Beck M, et al. (eds). Encyclopedia of social sciences research methods. Sage. 2003;792–795.
56. Zerzucha P, Daszykowski M, Walczak B. Dissimilarity partial least squares applied to non-linear modeling problems. Chemometrics Intell Lab Syst. 2012;110(1):156–62.
57. Mevik B-H, Wehrens R. The pls package: principal component and partial least squares regression in R. J Stat Soft. 2007;18(2):1–24.
58. Westerhuis JA, Hoefsloot HC, Smit S, Vis DJ, Smilde AK, van Velzen EJ, van Duijnhoven JP, van Dorsten FA. Assessment of PLSDA cross validation. Metabolomics. 2008;4(1):81–9.
59. Fearn T: Double cross-validation. In: News 3 Interview: Katherine Bakeev 4 Meetings: NIR on the Go 6 Quasi-imaging spectrometer with programmable field of view 8 Laboratory Profile: Regional Breeders Association of Lombardy 11: 2010; 2010: 201014.
60. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. 2009. na.
61. Szymańska E, Saccenti E, Smilde AK, Westerhuis JA. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics. 2012;8(1):3–16.
62. Lindgren F, Hansen B, Karcher W, Sjöström M, Eriksson L. Model validation by permutation tests: Applications to variable selection. J Chemometr. 1996;10(5–6):521–32.
63. Nieboer D, van der Ploeg T, Steyerberg EW. Assessing Discriminative Performance at External Validation of Clinical Prediction Models. PLoS One. 2016;11(2):e0148820.
64. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2011;4(3):250–5.
65. Li Q, Wacholder S, Hunter DJ, Hoover RN, Chanock S, Thomas G, Yu K. Genetic background comparison using distance-based regression, with applications in population stratification evaluation and adjustment. Genet Epidemiol. 2009;33(5):432–41.
66. Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971;27:857–71.
67. Gao X, Starmer J. Human population structure detection via multilocus genotype clustering. BMC Genet. 2007;8(1):34.
68. Nei M, Roychoudhury AK. Genetic relationship and evolution of human races. Evol Biol. 1982;14(1–59):2.
69. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr. 1957;27(4):325–49.
70. Ickstadt K, Selinski S, Network G. Similarity Measures for Clustering SNP Data. SFB 475, Fachbereich Statistik, Universität Dortmund: The Genica network Interdisciplinary Study Group on Gene Environment Interaction and Breast Cancer in Germany. HT014602036 2005.