Noname manuscript No.
(will be inserted by the editor)

Developing a parsimonius predictor for binary traits in sugar beet (*Beta vulgaris*)

Filson Nazzarini · Simone Marini · Piergiorgio Stevanato · Nelppo Biscicari

Received: 05 August 2014 / Accepted:

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords binary traits · genomic predictions · parsimonious predictor · sugar beet

1 Introduction

The primary goal of breeding schemes in farm animals and crops is generally to increase the agricultural output. Production traits are typically quantitative continuous variables (e.g. milk yield in dairy cattle, or per hectare yield in maize and rice). Many traits of importance in plant and animal breeding follow nonetheless a discrete categorical distribution, both ordered (e.g. calving ease in cattle, grain texture in rice) and unordered (e.g. grain pigmentation in rice, coat colour in cattle). A special case is that of binomial traits, which can take up only two different values, like disease resistance/susceptibility or presence/absence of a morphological characteristic. Annual bolting (flowering) behaviour and root vigor are examples of binomial traits of agronomic importance in sugar beet (*Beta vulgaris*). [move this?]

Filippo Biscarini and Nelson Nazzicari contributed equally to the work.

F. Biscarini, N. Nazzicari
Fondazione Parco Tecnologico Padano
E-mail: filippo.biscarini@tecnoparco.org

S. Marini
Dipartimento di Ingegneria Industriale e dell'Informazione
Università di Pavia

P. Stevanato
DAFNE, Università di Padova
24105 Padova, Italy

Advances in biotechnology and genomics, and the advent of high-density molecular markers (especially sinlge-nucleotide polymorphisms, SNP) genotyping have led to the emergence of molecular breeding [16]. One exciting application of molecular breeding is genomic selection: the possibility of predicting the genetic value and future performance of selection candidates solely from their genotypes ([15]). The predictive equations are trained on reference individuals with both genotypic and phenotypic data and then applied to selection candidates with genotypes only. Genomic selection may bring about multiple benefits in breeding programmes such as shorter breeding cycles or more efficient (e.g. traits difficult or expensive to phenotype) and more accurate (e.g. traits with low heritability) estimation of breeding values/selection ([7, 10]). Key to the application of genomic selection to breeding programmes are reliable genomic predictions. The recent publication of the reference genome for *Beta vulgaris* genome [4] is facilitating the application of molecular breeding also in this crop species. Pioneering studies on genomic predictions for both continuous ([11,21]) and binary ([2]) traits in sugar beet have already been published.

Genomic predictions are being based on increasing number of molecular markers (e.g. 777K SNP-chip in cattle, 56K SNP-chip in maize, whole-genome sequence data). When a huge number of potential predictors is available, it may be useful to select a subset to limit laboratory and bioinformatics costs, and the time of analysis, while at the same time improving interpretability of results. There is therefore interest in finding the minimum necessary set of information for a specific problem. The principle of parsimony states that a model needs to be simpler than the data it explains (think for instance of K-nearest neighbors -KNN- classifier with k=1), and according to Occam's razor, given two models that explain the data equally well, the simplest has to be chosen ([3]).

The objective of this paper is to present the development of a parsimonious predictor for the binary trait root vigor in a population of sugar beet accessions. SNPs in the panel were ranked based on their relevance and used to classify observations: one SNP at a time was removed, progressively reducing the number of SNPs in the predictive model. We found that it was possible to strongly reduce the dimension of the predictor and still achieve high accuracy.

2 Material and methods

2.1 Plant material and SNP genotypes

The available population comprised 123 individual sugar beet (*B. vulgaris*) plants, 100 with high- and 24 with low-root vigor. Root vigor is related to nutrient uptake from the soil and plant productivity ([20]) and is recorded as a binary trait (either high or low) based on the root elongation rate of eleven-days-old seedlings. No predetermined root elongation rate threshold was used to classify sugar beets into high or low root vigour. The classification

was subjectively made upon phenotypic inspection and has nevertheless been shown to be robust over time ([20]). The plant material was provided by Lion Seeds Ltd. (UK).

All plants were genotyped for 192 SNP markers with the high-throughput marker array QuantStudio 12K Flex system coupled with Taqman OpenArray technology. Additional details on the genotyping procedure are described in Stevanato et al., 2013 ([19]). After imputation and editing (call-rate $\geq 95\%$, MAF $\leq 2.5\%$) 175 SNPs were left for the analysis. A more detailed description of SNP genotypes and editing procedure can be found in Biscarini et al. ([2]).

2.2 Predictor development procedure

A two-step approach was adopted for the construction of a parsimonious predictor for root vigor in sugar beet. First, the 175 SNP available for the analysis after data editing were ranked based on their relevance for predicting the trait under study. In the second step the set of predictors was progressively reduced by removing the least useful predictors one at a time. At each iteration logistic regression was used to classify observations with the given set of SNPs. As many classification results as the number of SNPs (i.e. 175) were therefore obtained.

2.2.1 Rank of predictors

When many predictors are available -especially if $p > n$- it may be of interest to reduce the dimensionality of the problem by choosing the optimal subset of predictors that best describe the relationship between dependent and independent variables [or that best model the response variable, or are most informative with respect to the outcome to be predicted, etc ...]. A Bayesian model selection method, the binary outcome stochastic search (BOSS) algorithm [18], was applied to identify the best set of predictive SNPs by repeated sampling in a Markov Chain Monte Carlo (MCMC) approach. SNPs were ranked based on their probability of inclusion in the best predictive model.

In BOSS the relationship between binary observations and predictors is described by a Gaussian latent variable model with a probit link function:

$$P(Y = [0/1]|X) = \phi(X\beta) \tag{1}$$

where $P(Y = [0/1]|X)$ is the probability of having low or high root vigor given the SNP genotyps X, β is a vector of regression coefficients, and ϕ is the normal cumulative distribution function.

The n independent latent variables were normally distributed as $Z_i \sim N(X_i^T\beta, 1)$ (with $i = 1 \dots n$), and used to model the relationships between SNP genotypes and root vigor:

$$Y_i = \begin{cases} 0 & Z_i \leq 0 \\ 1 & Z_i > 0 \end{cases} \quad \mathbf{Z} = \mathbf{X}\beta + \epsilon \tag{2}$$

BOSS extensively explores the model space to identify relevant predictors, similar to what is done in Best Subset Selection ([9]).

The selection of predictors was performed by introducing a vector of indicator variables $\gamma = (\gamma_1, \ldots, \gamma_p)$ such that $\gamma_i = 0$ if $\beta_i = 0$ and $\gamma_i = 1$ if $\beta_i \neq 0$. A predictor was included in the model if its regression coefficient was not null and the associated indicator variable $\gamma = 1$. The model space to be searched was therefore given by the 2^p possible combinations of SNPs (either included or excluded). This yielded a vector β_γ of size p_γ with only the coefficients for which $\gamma_i = 1$.

The prior for vector γ was chosen from the following beta-binomial (B) distribution:

$$f(\gamma) = \frac{B(p_\gamma + a, p - p_\gamma + b)}{B(a, b)} \tag{3}$$

with parameters a and b related to model size p_γ.

Knowing that the posterior conditional density for γ is proportional to its prior and the marginal likelihood of $\mathbf{Z}$, the following holds:

$$f(\gamma|\mathbf{Z}) \propto f(\gamma) \cdot f(\mathbf{Z}|\gamma) \tag{4}$$

from which vector γ can be sampled.

The BOSS algorithm is summarised below:

1. initialize the latent variables $\mathbf{Z}$;
2. sample γ through a Metropolis-Hastings sampler (check the Theory that would not die) from its conditional posterior distribution in equation 4;
3. given $\mathbf{Z}$ and γ, sample β_γ through standard Bayesian linear regression ($\mathbf{Z} = \mathbf{X}_\gamma \beta_\gamma + \epsilon$);
4. given β_γ and $\mathbf{Y}$, sample the vector of latent variables $\mathbf{Z}$ from equation 2;
5. restart from step 2 until $m = 1\,000\,000$ iterations are completed

From the $m = 1\,000\,000$ MCMC iterations the inclusion probabilities for the 175 SNPs were computed and used to rank the predictors. A more detailed description of the BOSS algorithm can be found in Russu et al. ([18]).

2.2.2 Selection of predictors and classification method

SNPs were selected based on the BOSS rank (see section 2.2.1) by progressively removing one SNP at a time. The full set of 175 SNPs was used at first in the prediction model. Subsequently, the m^{th} least relevant SNP (from the BOSS rank) was removed each time, and the resulting $175-m$ SNPs model was fitted. As a result of this, 175 different predictive models (from 175 to 1 predictor) were fitted.

With each set of SNP, a logistic regression model for binary outcomes was used to classify observations into low and high root vigor based on the SNP genotypes. The probability having high root vigor ($P(Y = 1|X) = p(x)$) was

modeled as a linear combination of the predictors (SNPs) through a *logit* link-function in a generalised linear model:

$$logit(p(x_i)) = log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \mu + \sum_{j=1}^{m} z_{ij} SNP_j \quad (5)$$

where $p(x_i)$ is the $P(Y = 1|X)$ for individual i with vector of predictors x_i; SNP_j is the effect of the j_{th} marker; z_{ij} is the genotype of individual i at locus j (0, 1 or 2 for AA, AB and BB genotypes). Equation 5 returns the odds of $p(x)$ which are backtransformed to $P(Y = 1|X)$ through the cumulative distribution function of the logistic distribution (i.e. the logistic function). Individuals with $p(x) > / < 0.5$ were classified as high/low root vigor individual plants.

Since there were more predictors than observations in model 5 ($p > n$), a ridge logistic regression fitting method [13] was adopted, which consisted in maximizing the following penalized log-likelihood:

$$L(\mu, SNP) = \sum_{i=1}^{n} [y_i log(p_i) + (1 - y_i) log(1 - p_i)] - \frac{1}{2}\lambda \sum_{j=1}^{p} SNP_j^2 \quad (6)$$

where μ is the intercept, SNP is the vector of SNP effects (i.e. the regression coefficients of SNP genotypes) and λ is a tuning parameter that was specified through cross-validation and controls the amount of penalization.

2.2.3 Predictive ability

For each set of SNPs (175 to 1), the predictive ability of the ridge logistic regression model (equation 5) was assessed through a two-fold cross-validation. The 123 samples were randomly split into a training and testing set of approximately equal size. The training set was used to fit the model and estimate SNP effects which were then used to predict root vigor in the testing set. This process was repeated 100 times, each time sampling different training and testing sets. The test error rate in each replicate was computed as:

$$ER_{(n)} = \frac{1}{n} \sum_{i=1}^{n} Err_i \quad (7)$$

where n is the number of observations in the test set and $Err_i = I(y_i \neq \hat{y}_i)$, with $I(\cdot)$ an indicator function which returns a value of 1 if the predicted and observed phenotypes are different, 0 otherwise. The cross-validation error rate was then estimated averaging the test error rate over all $k = 100$ replicates:

$$CV_k = \frac{1}{k} \sum_{i=1}^{k} ER_i \quad (8)$$

The prediction accuracy was then defined as $1 - CV_k$. Besides the total error rate, also the false positive (FPR) and false negative (FNR) rates were computed. False positives (FP) were defined as true low-root vigor plants predicted as high-root vigor, whereas false negatives (FN) were defined as true low-root vigor plants predicted as high-root vigor. Then, $FPR = {}^{FP}/_{N}$ and $FNR = {}^{FN}/_{P}$, where N and P are the total number of true negative (low-root vigor) and true positive (high-root vigor) samples.

2.3 Comparison with another method to rank predictors

The BOSS algorithm yielded a rank of the SNPs based on their probability of inclusion in the best predictive model. SNPs could be ranked also based on different metrics. For instance, the relative effect on the trait under analysis could be chosen as an indirect measure of the predictive importance of the SNPs.

A logistic model of the form $logit(p_i) = \mu + SNP_m$ ($i = 1, \ldots, 123$ plants; $m = 1, \ldots, 175$ SNPs) was fit one SNP at a time in order to estimate marker effects. From estimated marker effects and allele frequencies, the additive genetic variance at each SNP locus was estimated as:

$$V_A = 2pqa^2 \tag{9}$$

where p and q are the frequencies of the two alleles in the population and a is the marker additive effect ([5]). SNPs were then ranked based on decreasing genetic variance and used for predictor selection and classification as in section 2.2.2. Prediction accuracy was estimated as in section 2.2.3.

2.4 Software

Matlab scripts [14] were used to implement the BOSS algorithm for model selection. Classification of observations was done with the Java data mining software *Weka 3* [8]. The open source statistical environment *R* [17] and the open source spreadsheet *Gnumeric 1.12.12* ([6]) were used for data manipulation, the creation of figures and all other statistical analyses.

3 Results

The ranking of SNPs in descending order of predictive relevance from the BOSS algorithm is presented in supplementary table S1. SNPs were ranked based on their probability of inclusion in the best predictive model over 1 000 000 MCMC iterations. The most relevant SNPs were those most often included in the predictive model. The first SNP of the ranking (predictor with the highest probability of inclusion) was *SNP109* on scaffold00456 on chromosome 9: this SNP was therefore included in all 175 SNP subsets (from all 175 SNPs

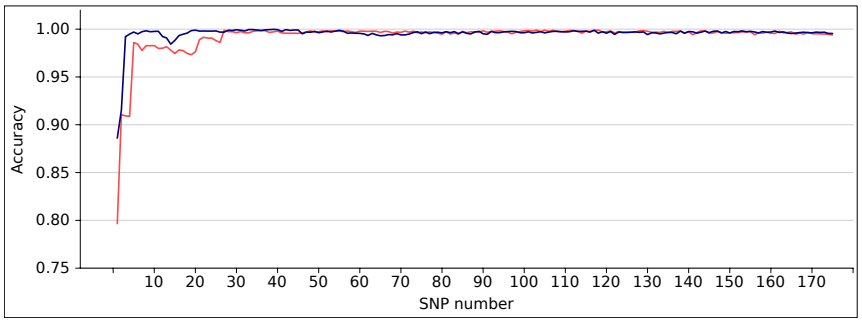

Fig. 1 Accuracy (1 - error rate) of prediction as a function of the number of SNPs included in the classifier: BOSS (blue line) vs logistic regression (red line)

to just one single SNP) used for classification. Of the top 20 SNPs, 6 were on chromosome 9, 2 each on chromosomes 3, 7 and 8, and 1 on chromosome 2 of the sugar beet genome. Seven of the top 20 SNPs were not mapped to any chromosome. The last (least relevant) 20 SNPs were distributed as follows: 6 on chromosome 4, 4 on chromosome 1, 3 on chromosome 7, 1 on chromosomes 3, 5, 6 and 9; three were unmapped. Chromosome and scaffold information for each ranked SNP can be found in the supplementary table S1.

Figure 1 shows the accuracy of prediction as a function of the number of SNPs used for classification based on both the BOSS (blue line) and SNP-variance (red line) rankings. The prediction accuracy of classifying individuals as high- or low-root vigor was defined as $1 - \overline{test\_error\_rate}$ as in equation 8.

improve plot: no need to go down to 0.0 in the y-axis; legend names and position; color of the lines? The "bump" at around 20-30 SNPs is not visible]

Table 1: TER, FPR, FNR for the first 30/35 SNPs + average for the rest of the SNP (error close to 0). BOSS + GWAS (6 columns)

Probability of assignment as a function of predictors: Figure 2. Better a table? Maybe in discussion?

From ROC curves only the AUC. No plot, use AUC as result in the text (e.g. comparison between ranker: overall average AUC, average AUC per # SNPs + std). Table?

4 Discussion

General overview why error rates are not evenly distributed? Reminder: it works very well because of LD and H2

Unstable below 30/40 SNPs; little "bump" around 20 SNPs: more marked with BOSS, but also visible with GWAS. Why there? SNPs with large effect on the trait, but low significance? SNPs with large effect but low LD (with the QTL)? In the latter case, the marker might sometimes be in the opposite phase. Look also at marker frequency.

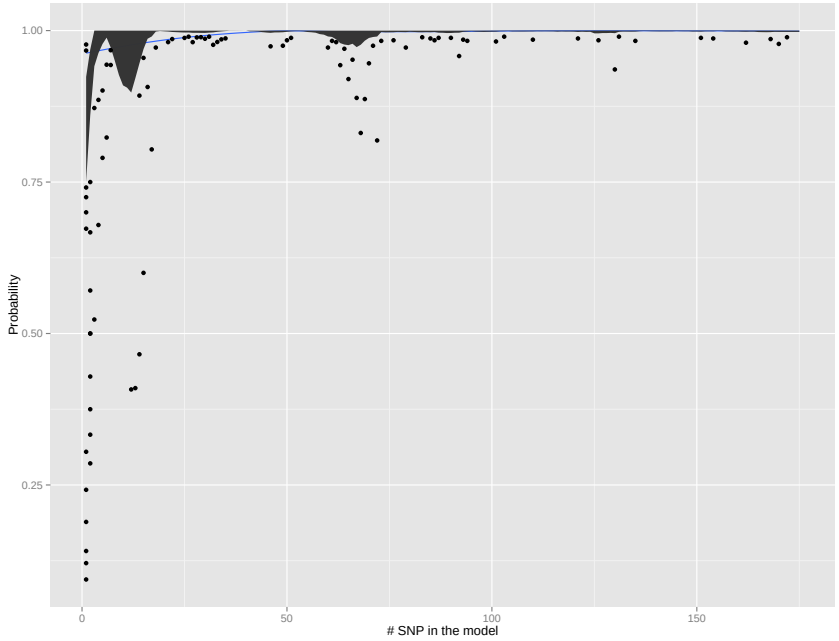

Fig. 2 Distribution of $P(Y = 1|x)$ as a function of the number of SNPs in the classifier

Based on results, a panel of 30-35-40 SNPs is recommended for accurate prediction of root vigor (move to breeding applications? Together with development of a custom-chip?)

4.1 Relative performance of rankers

why using Pvalues and not other standard rankers (e.g. backward stepwise selection)? Because of the specific nature of the problem. ROC and AUC?

Comparison of rankers: spearman correlation + plot (ranker1 vs ranker2).

Figure 4

4.2 SNP effects [alternative: LD and probabilities]

Manhattan plot with BOSS weights and weights from the other articles, somehow compared (same chart? two charts? only ten best?).

Let's not use effects from GWAS (the paper is on BOSS!!): something on the top ranked SNPs (position ...)?

Do the peaks make sense from the biological perspective? (something on sugar beet genes?)

Variance of SNPs vs genetic variance: → missing heritability? (cite Brachi 2011, Manolio 2009?).

Fig. 3 Average linkage disequilibrium (LD) for increasing number of SNPs in thepredictive model

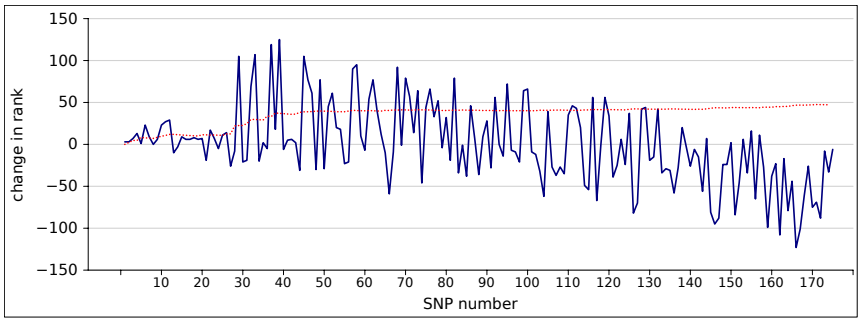

Fig. 4 Comparison of BOSS and logistic regression in terms of relative rank position of relevant SNPs

BOSS probability: 1 big peak + smaller peaks. Compare against SNP density? Maybe the big peak corresponds to a physically isolated SNP, whereas smaller peaks correspond to a cluster of SNPs in LD which individually account for a smaller part of the variatino, but together play an important predictive role.

4.3 Genotyping strategies and applications to breeding

Several techology choices are commonly available when genotyping strategies must be decided. Assuming knowledge of SNP flanking sequences, we examine four options: SNP chips, genotyping by sequencing (GBS), targeted sequencing (TS), and a commercial solution, Illumina BeadXpress.

Genotyping by sequencing is

genotyping strategies: Costs, possible technologies (gbs, snp chip, macroarrays), implications

applications to breeding: why is it important root vigor early detection. Other binomial traits (e.g. disease resistance) May be applied to bolting (another trait which exhibits binomial distribution), which has been shown to be controlled by multiple genes and influenced by environmental factors ([1]).

sugar beet: 30% of world's sugar production (cite Dohm? FAO?). Root vigor linked to yield.

Sugar beet: sugar + energy (citation?)

Other binomial traits: resistance to viral and fungal diseases, bolting (cite Dohm? Someone else?)

Breeding has shaped the genome of sugar beet (comparison with *Beta maritima*, [4]).

Extensions to multinomial traits? Examples?

potential and challenges of genomic selection in plant breeding ([12])

5 Conclusions

Concluding remarks.

Acknowledgements This research was financially supported by the Marie Curie European Reintegration Grant "NEUTRADAPT".

References

1. Abou-Elwafa, S., Büttner, B., Kopisch-Obuch, F., Jung, C., Müller, A.: Genetic identification of a novel bolting locus in *Beta vulgaris* which promotes annuality independently of the bolting gene *B.* Molecular Breeding **29**, 989–998 (2012)
2. Biscarini, F., Stevanato, P., Broccanello, C., Stella, A., Saccomani, M.: Genome-enabled predictions for binomial traits in sugar beet populations. BMC Genetics **18**(5), 1–9 (2014)
3. Chaitin, G.: The limits of reason. Scientific American **294**(3), 74–81 (2006)
4. Dohm, J.C., Minoche, A.E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sörensen, T.R., Stracke, R., Reinhardt, R., et al.: The genome of the recently domesticated crop plant sugar beet (beta vulgaris). Nature (2013)
5. Gianola, D., de Los Campos, G., Hill, W.G., Manfredi, E., Fernando, R.: Additive genetic variability and the bayesian alphabet. Genetics **183**(1), 347–363 (2009)
6. GNOME-Project: The gnumeric spreadsheet (2014). URL http://www.gnumeric.org/
7. Goddard, M., Hayes, B.: Genomic selection. Journal of Animal Breeding and Genetics **124**(6), 323–330 (2007)

Table 1 Total error rate (TER), false positive (FPR) and false negative (FNR) rates as a function of the number of SNPs incluided in the model, for both the BOSS and SNP variance rankings

	BOSS			SNP Variance		
# SNPs	TER	FPR	FNR	TER	FPR	FNR
1	0.114	0.333	0.061	0.204	0.500	0.132
2	0.085	0.192	0.059	0.089	0.002	0.111
3	0.008	0.033	0.002	0.091	0.001	0.113
4	0.005	0.025	0.001	0.091	0.001	0.113
5	0.003	0.006	0.002	0.014	0.000	0.017
6	0.005	0.011	0.004	0.016	0.001	0.019
7	0.003	0.005	0.002	0.022	0.001	0.028
8	0.002	0.002	0.002	0.017	0.000	0.022
9	0.003	0.000	0.003	0.017	0.000	0.022
10	0.002	0.001	0.003	0.017	0.000	0.021
11	0.002	0.001	0.002	0.020	0.001	0.025
12	0.008	0.001	0.009	0.020	0.002	0.024
13	0.009	0.001	0.011	0.018	0.001	0.023
14	0.016	0.000	0.019	0.022	0.001	0.027
15	0.012	0.000	0.015	0.025	0.003	0.032
16	0.007	0.001	0.008	0.022	0.002	0.027
17	0.005	0.000	0.007	0.022	0.001	0.027
18	0.004	0.001	0.005	0.025	0.002	0.032
19	0.002	0.001	0.002	0.027	0.002	0.033
20	0.001	0.000	0.001	0.022	0.001	0.029
21–30	0.002	0.001	0.002	0.007	0.003	0.008
31–40	0.001	0.000	0.001	0.003	0.002	0.003
41–50	0.004	0.001	0.003	0.004	0.002	0.004
51–60	0.002	0.001	0.004	0.004	0.004	0.003
61–70	0.003	0.001	0.007	0.003	0.003	0.002
71–80	0.004	0.001	0.005	0.003	0.003	0.003
81–90	0.002	0.004	0.002	0.003	0.004	0.003
91–100	0.001	0.001	0.001	0.003	0.003	0.002
101–175	0.001	0.002	0.001	0.002	0.005	0.002

8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter **11**(1), 10–18 (2009)
9. Hastie, T., Tibshirani, R., Friedman, J.: Model assessment and selection. In: The elements of statistical learning, pp. 219–260. Springer, New York, NY, USA (2009)
10. Heffner, E.L., Lorenz, A.J., Jannink, J.L., Sorrells, M.E.: Plant breeding with genomic selection: gain per unit time and cost. Crop science **50**(5), 1681–1690 (2010)
11. Hofheinz, N., Borchardt, D., Weissleder, K., Frisch, M.: Genome-based prediction of test cross performance in two subsequent breeding cycles. Theoretical and Applied Genetics **125**(8), 1639–1645 (2012)
12. Jonas, E., de Koning, D.J.: Does genomic selection have a future in plant breeding? Trends in biotechnology **31**(9), 497–504 (2013)
13. Liu, Z., Shen, Y., Ott, J.: Multilocus association mapping using generalized ridge logistic regression. BMC bioinformatics **12**(1), 384 (2011)
14. MATLAB: version 7.10.0 (R2010a). The MathWorks Inc., Natick, Massachusetts (2010)
15. Meuwissen, T., Hayes, B., Goddard: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**(4), 1819–1829 (2001)
16. Moose, S.P., Mumm, R.H.: Molecular plant breeding as the foundation for 21st century crop improvement. Plant physiology **147**(3), 969–977 (2008)

17. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008). URL http://www.R-project.org. ISBN 3-900051-07-0
18. Russu, A., Malovini, A., Puca, A.A., Bellazzi, R.: Stochastic model search with binary outcomes for genome-wide association studies. Journal of the American Medical Informatics Association **19**(e1), e13–e20 (2012)
19. Stevanato, P., Broccanello, C., Biscarini, F., Del Corvo, M., Sablok, G., Panella, L., Stella, A., Concheri, G.: High-throughput rad-snp genotyping for characterization of sugar beet genotypes. Plant Molecular Biology Reporter pp. 1–6 (2013)
20. Stevanato, P., Trebbi, D., Saccomani, M.: Root traits and yield in sugar beet: identification of aflp markers associated with root elongation rate. Euphytica **173**(3), 289–298 (2010)
21. Würschum, T., Reif, J.C., Kraft, T., Janssen, G., Zhao, Y.: Genomic selection in sugar beet breeding populations. BMC genetics **14**(1), 85 (2013)