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

Filippo Biscarini and Nelson Nazzicari contributed equally to the work.

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

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

N. Nazzicari
Fondazione Parco Tecnologico Padano
E-mail: nelson.nazzicari@tecnoparco.org

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

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?]

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 [12]. 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 ([11]). 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 ([6,8]). 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 ([9,18]) 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 the 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

A population of 124 individual sugar beet (*B. vulgaris*) plants from 18 high- and low-root-vigor lines were available. These lines were characterised by different productivity and were provided by Lion Seeds Ltd. (UK). Root vigor is related to nutrient uptake from the soil and plant productivity ([17]), and is recorded as a binary trait (either high or low). The lines were phenotyped by measuring the root elongation rate of eleven-days-old seedlings grown under hydroponic conditions. There was no predetermined root elongation rate threshold to classify a sugar beet as having high or low root vigour, and the decision was subjectively made upon phenotypic inspection. The classification has nevertheless been shown to be robust: seedlings classified as "low" or "high" maintain the same class also at the adult plant stage ([17]). There were three low-root-vigor (24 individuals) and 15 high-root-vigor (100 individuals) lines. Root elongation rate was < 3 mm/day in the low-root-vigor lines and > 6 mm/day in the high-root-vigor lines.

All individual 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 ([16]).

The initial genotype screening led to the detection of one duplicated individual (100% matching genotypes) from a high-root-vigor line, which was removed. The average per-sample and per-marker call-rate was 0.984 and 0.969. Only one SNP had a per-marker call-rate $\leq 85\%$ and was removed from the analysis. After imputation data were edited for minor allele frequency (MAF): 16 SNPs with MAF $\leq 2.5\%$ were discarded. This left a total of 123 individuals and 175 SNP markers for the analysis. An overview of the data used in the paper is given in

Root vigor. Available data. SNP technology used, imputation. Copypaste from other articles. Dataset description. Text with citations [16] and [15].

2.2 Predictor development procedure

A two-step approach was adopted for the construction of a parsimonious predictor for root vigor.

- a ranker to rank the various available predictors (SNPs in our case). We used the BOSS algorithm - this is an iterative step. we progressively reduced the predictors set, taking away the laest useful predictor and applying to the resulting subset a ridge logistic regression apprach. Thus, we obtained as many performances estimation as the number of original predictors.

2.2.1 Rank of predictors

This explain the BOSS algorithm [14]

2.2.2 Selection of predictors and classification method

We take one predictor out at each iteration You put the model formula for ridge logistic regression

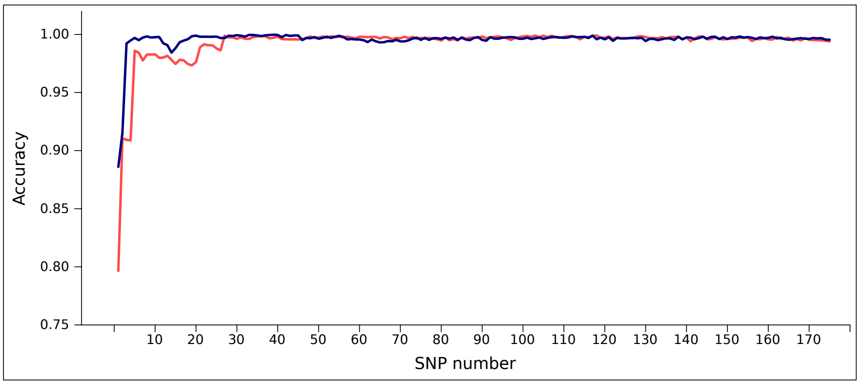

Fig. 1 Accuracy (1 - error rate) of prediction as a function of the number of SNPs included in the classifier: BOSS (blue line) vs logistic regression (red line)

2.2.3 Predictive ability

Cross validation: how many times, what fractions. Explanation of error rate and other parameters (ROC?)

2.3 Comparison with another method to rank predictors

Another ranker: why use one, and its description. P value and SNP effect (as it is done in GWAS)

SNP variance [5]

2.4 Software

R [13], weka [7], perl.

3 Results

Figure 1: Accuracy as a function of the number of predictors, BOSS vs logistic [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?

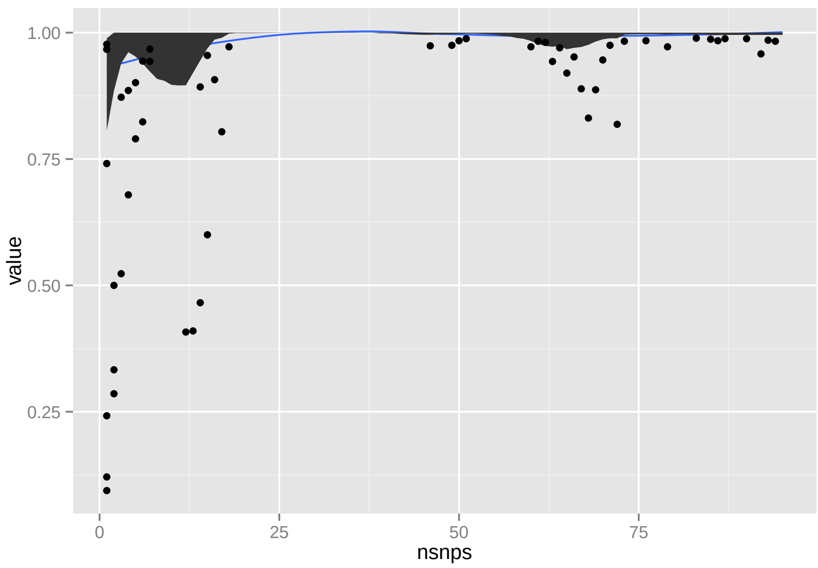

Fig. 2 Distribution of $P(Y = 1|x)$ as a function of the number of SNPs in the classifier

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.

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

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

Figure 3

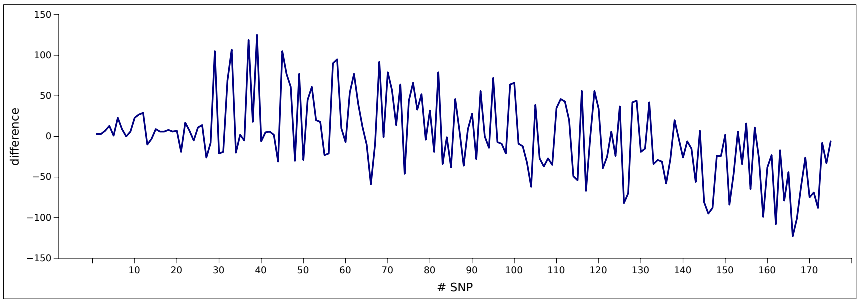

Fig. 3 Comparison of BOSS and logistic regression in terms of relative rank position of relevant SNPs

4.2 SNP effects

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

Do the peaks make sense from the biological perspective?

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

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

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 ([10])

Table 1 Total error rate (TER), false positive (FPR) and false negative (FNR) rates as a function of the number of SNPs ranked according to BOSS or logistic regression

# SNPs	TER	FPR	FNR
1	0.114	0.065	0.049
2	0.085	0.037	0.047
3	0.008		
4	0.005		
5	0.003		
6	0.005		
7	0.003		
8	0.002		
9	0.003		
10	0.002		
11	0.002		
12	0.008		
13	0.009		
14	0.016		
15	0.012		
16	0.007		
17	0.005		
18	0.004		
19	0.002		
20	0.001		
...	...	...	...
21–30	0.002		
31–40	0.001		
41–100	0.003		
101–175	0.001		

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. Goddard, M., Hayes, B.: Genomic selection. Journal of Animal Breeding and Genetics **124**(6), 323–330 (2007)
7. 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)
8. 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)
9. 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)
10. Jonas, E., de Koning, D.J.: Does genomic selection have a future in plant breeding? Trends in biotechnology **31**(9), 497–504 (2013)
11. Meuwissen, T., Hayes, B., Goddard: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**(4), 1819–1829 (2001)
12. Moose, S.P., Mumm, R.H.: Molecular plant breeding as the foundation for 21st century crop improvement. Plant physiology **147**(3), 969–977 (2008)
13. 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
14. 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)
15. Saccomani, M., Stevanato, P., Trebbi, D., McGrath, J.M., Biancardi, E.: Molecular and morpho-physiological characterization of sea, ruderal and cultivated beets. Euphytica **169**(1), 19–29 (2009)
16. 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)
17. 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)
18. Würschum, T., Reif, J.C., Kraft, T., Janssen, G., Zhao, Y.: Genomic selection in sugar beet breeding populations. BMC genetics **14**(1), 85 (2013)