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
Tel.: +123-45-678910
E-mail: filippo.biscarini@tecnoparco.org

S. Marini
second address

N. Nazzicari
Fondazione Parco Tecnologico Padano
Tel.: +123-45-678910
E-mail: nelson.nazzicari@tecnoparco.org

P. Stevanato
address

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*).

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 [10]. One exciting application of molecular breeding is genomic selection ([9]): the possibility of predicting the genetic value and future performance of selection candidates solely from their genotypes. 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 ([6,7]). Key to the application of genomic selection to breeding programmes are therefore reliable genomic predictions.

The recent publication of the reference genome for *Beta vulgaris* genome [4] facilitated molecular breeding also in this crop species. Genomic predictions in sugar beet already done both for continuous ([8,14]) and binary ([2]) traits.

The concept of parsimony: when many possible predictors are available, it is useful to select a subset to limit analysis cost and time. Moreover: use the minimun necessary information set, occam razor ([3]), and so forth. A model need to be simpler than the data the it fits/explains (e.g. knn with k=1)

Given two models that fit the data, the simplest has to be chosen (Occam's razor)

As the technology advances, and available predictors grow, not only the prediction precision becomes important, but also the actual cost must be considered.

Sugar beets in particular: we work on root vigor [2].

In this paper we propose statistical methods to highlight and select the most useful predictors given a set. We started on real world data and validated our approach on a XXX dataset. We found that it is possible to strongly reduce the dimension of the predictors set and still achieve high performance.

2 Material and methods

2.1 Plant material and SNP genotypes

Root vigor. Available data. SNP technology used, imputation.
Copypaste from other articles. Dataset description. Text with citations [13] and [12].

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 [11]

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

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, weka, 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?

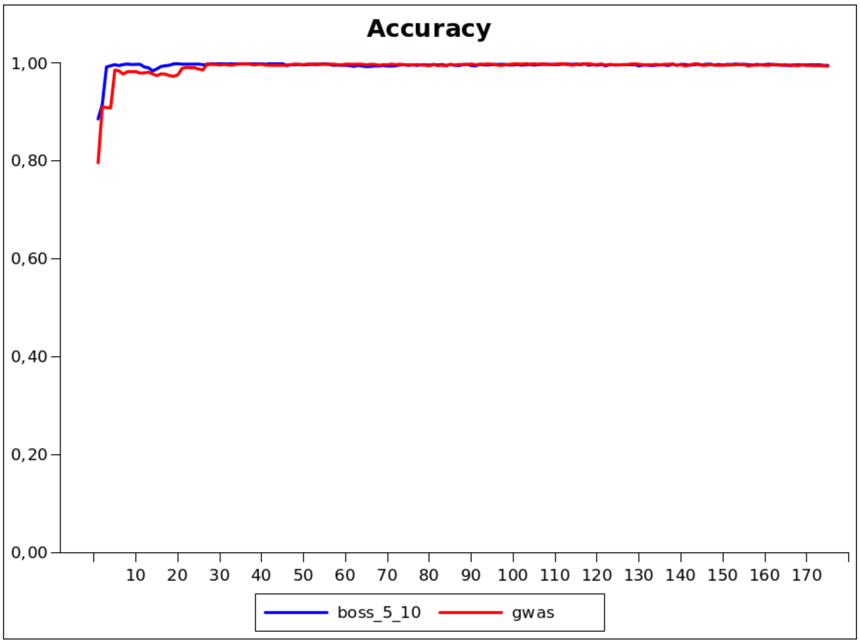

Fig. 1 Accuracy (1 - error rate) of prediction as a function of the number of SNPs included in the classifier: BOSS (blue line) vs logistic regression (red line)

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.

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 SNP effects

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

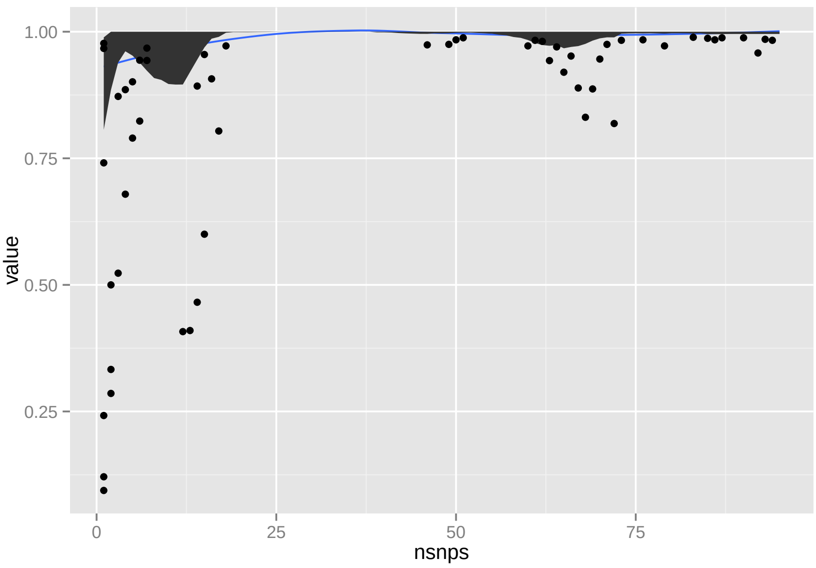

Fig. 2 Distribution of $P(Y = 1|x)$ as a function of the number of SNPs in the classifier

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.

4.2 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

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 (an-

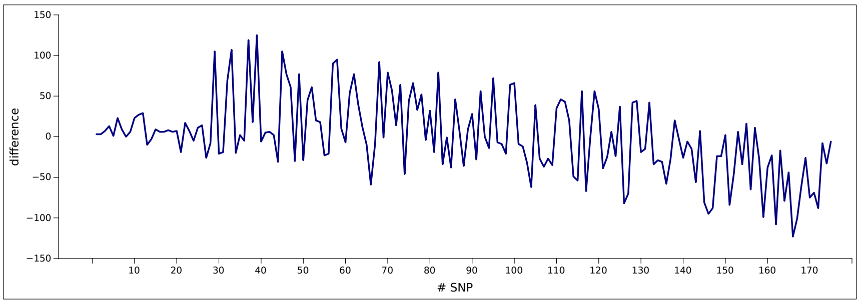

Fig. 3 Comparison of BOSS and logistic regression in terms of relative rank position of relevant SNPs

other 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?

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)

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		

7. 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)
8. 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)
9. Meuwissen, T., Hayes, B., Goddard: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**(4), 1819–1829 (2001)
10. Moose, S.P., Mumm, R.H.: Molecular plant breeding as the foundation for 21st century crop improvement. Plant physiology **147**(3), 969–977 (2008)
11. 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)
12. 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)
13. 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)
14. Würschum, T., Reif, J.C., Kraft, T., Janssen, G., Zhao, Y.: Genomic selection in sugar beet breeding populations. BMC genetics **14**(1), 85 (2013)