Multivariate statistical modelling for QTL detection and marker selection in a bi-parental grapevine population

Charlotte Brault, Marie Perrot-Dockès, Agnes Doligez, Julien Chiquet, Loic Le Cunff, Timothée Flutre

Multivariate statistical modelling for QTL detection and marker selection in a bi-parental grapevine population

Brault.C[1], Perrot-Dockès.M[2], Doligez.A[1], Chiquet.J[2], Le Cunff.L[3], Flutre.T[1]
[1]AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France
[2]AgroParisTech, UMR MIA, Paris, France
[3]Institut Français de la Vigne et du Vin, Montpellier, France

Abstract

In the present study, we perform variable selection with various flavours of the LASSO method (group Lasso, fused Lasso) adapted for multiple responses, extending the model and algorithm from Chiquet *et al.* (2017). We apply these methods on simulated data and on real data from Coupel-Ledru *et al.* (2014).

The aim of multivariate variable selection is to study multiple responses together in order to take into account the genetic correlation among responses (figure 1).

Genotyping

We worked on a 191-progeny of Syrah x Grenache. They have been genotyped with 153 SSR markers.

Phenotypic data

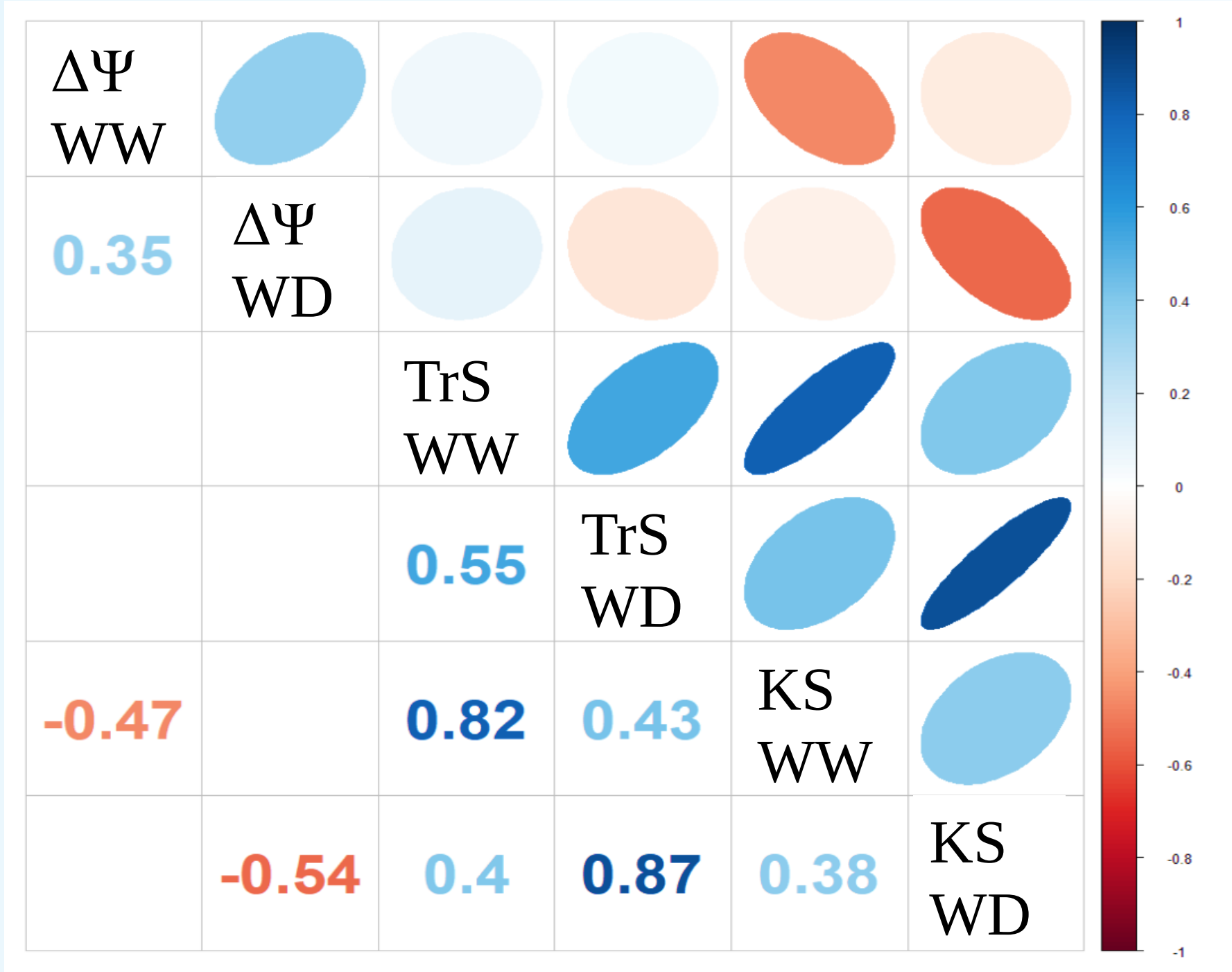


Figure 1
Genetic correlation table of 3 traits under 2 hydric conditions (WW and WD)

ΔΨ : difference of water potential between soil and leaves
TrS : specific transpiration rate
KS : hydraulic conductance

Three types of statistical modelling

$$\boldsymbol{Y} \sim \boldsymbol{XB} + \boldsymbol{E}$$

- Univariate with 1 trait in 1 hydric condition

We started by studying each response separately.

- Univariate with 1 trait in 2 hydric conditions

We then analysed 1 trait by adding in the statistical model a co-variable for each hydric condition (not shown).

- Multivariate with 6 responses

We finally considered the 3 traits and 2 conditions jointly (figure 2).

We used classical composite interval mapping with R/qtl only in the first modelling. For the last two, we used regularized regression with L1 and /or L2 penalties (classic lasso, group lasso and fuse lasso, structured elastic net).
In the multivariate case, the group lasso selects an allele of a marker if there is a non-null estimated effect in all responses. For this, we used the R/glmnet package (Friedman *et al.*, 2010).
The multivariate structured Elastic Net estimates correlations between responses and distinguishes between direct and indirect effects. Moreover, compare to the classical lasso, it can make use of the genetic map. For this, we used the R/spring package (Chiquet *et al.*, 2017).

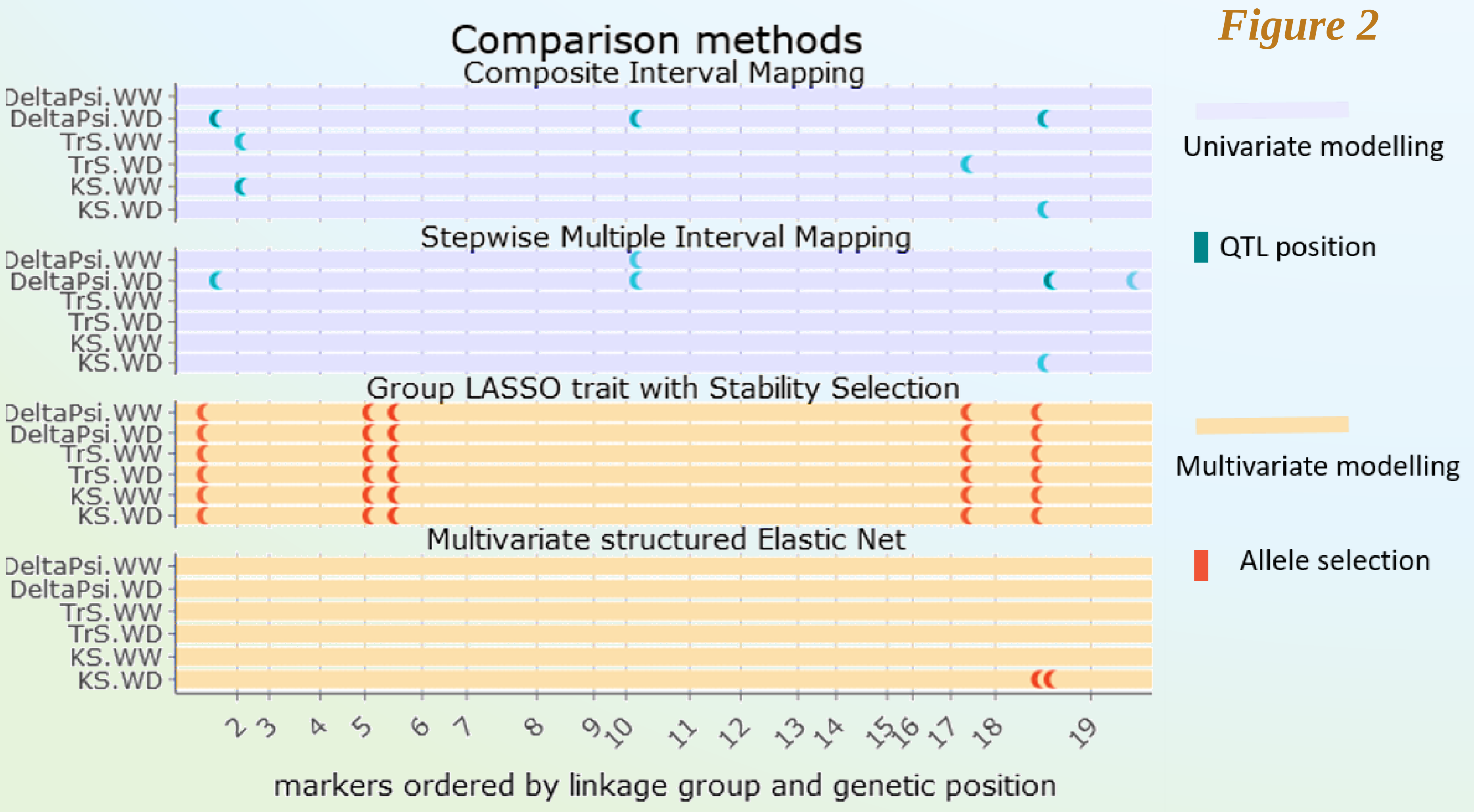


Figure 2

Simulation

We used the design matrix of SSR genotypes from the Syrah x Grenache progeny, and simulated twenty times two responses with the aim of comparing various variable selection methods.
In each simulation, a single marker has a non-null effect. Its magnitude is similar to what was estimated in the real data. But in the simulations, we vary heritability (h2), genetic correlation (rhoB) among responses and add a dominance effect.
As we know the true position of the QTL, we calculate the True Positive Proportion (TPP) and the False Positive and Negative Proportions (FPP, FNP) for each method for these parameters (figure 3).
For all variable selection method (except structured elastic net), we used stability selection.

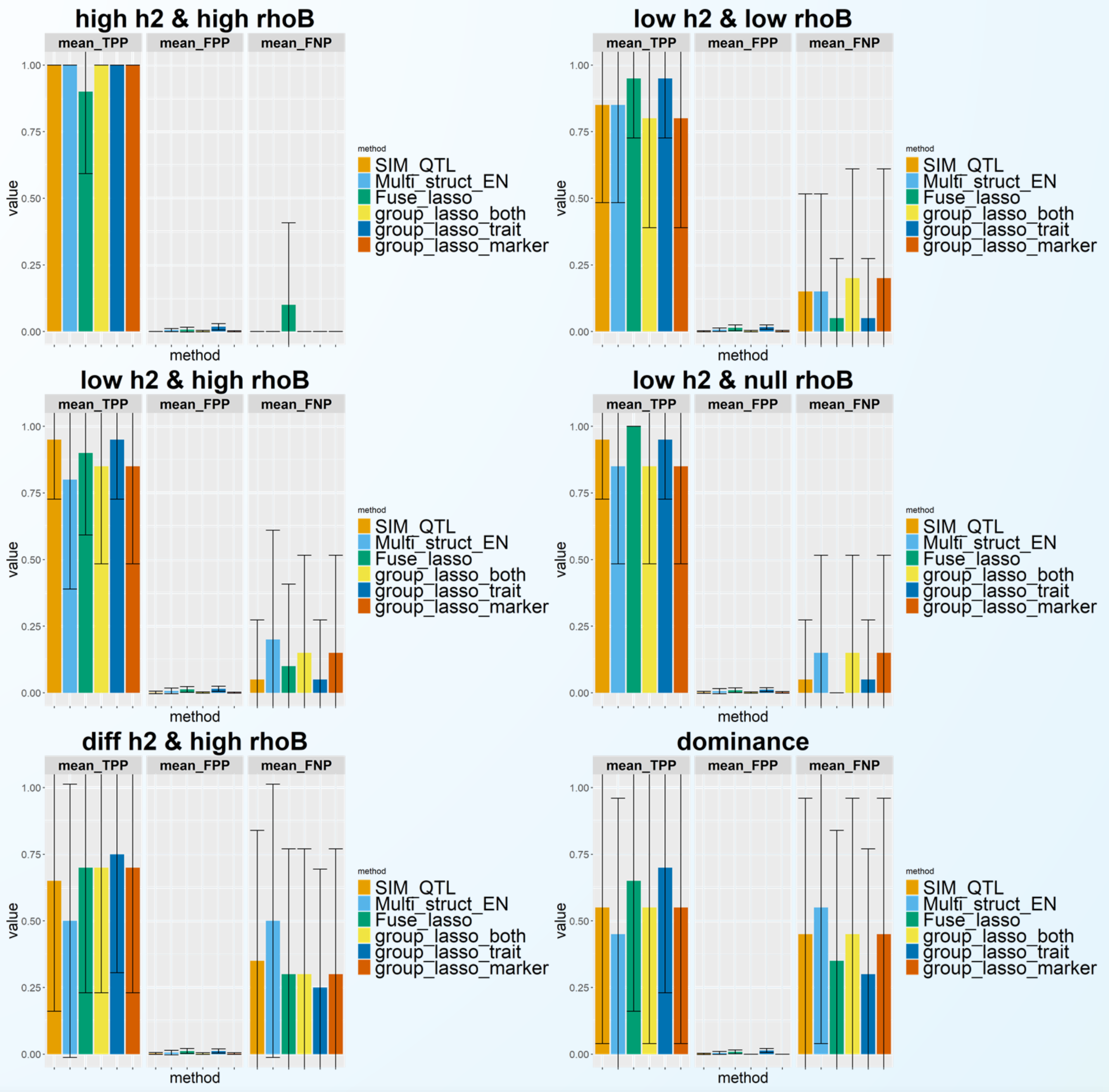
Figure 3
Results of Simulations

High h² = 0.8
Low h² = 0.18
Different h² = 0.18 / 0.10

High rhoB = 1
Low rhoB = 0.5
Null rhoB = 0

Dominance : with low h² & high rhoB

SIM : Simple Interval Mapping



Conclusion

Given that traits may share the same genetic basis, we expect the multivariate statistical modelling to be appropriate to analyse multiple responses. In this preliminary work, to experiment with various modelling assumptions, we tried several methods, which gave contrasted results on real data. Further work via simulations are required to clarify the impact of these assumptions.

Main references

Coupel-Ledru, A., Lebon, É., Christophe, A., Doligez, A., Cabrera-Bosquet, L., Péchier, P., Hamard, P., This, P., Simonneau, T. (2014). Genetic variation in a grapevine progeny (Vitis vinifera L. cvs Grenache×Syrah) reveals inconsistencies between maintenance of daytime leaf water potential and response of transpiration rate under drought. *Journal of experimental botany*, *65*(21), 6205-6218.
Chiquet, J., Mary-Huard, T., & Robin, S. (2017). Structured regularization for conditional Gaussian graphical models. *Statistics and Computing*, *27*(3), 789-804