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

To cite this version:
Charlotte Brault, Marie Perrot-Dockès, Agnes Doligez, Julien Chiquet, Loic Le Cunff, et al.. Multivariate statistical modelling for QTL detection and marker selection in a bi-parental grapevine population. 17. Meeting of the EUCARPIA Section Biometrics in Plant Breeding, Sep 2018, Gand, Belgium. hal-01868802
Genotyping
We worked on a 191-progeny of Syrah x Grenache. They have been genotyped with 153 SSR markers.

Phenotypic data

Three types of statistical modelling

\( Y \sim XB + E \)

- 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 fused 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).

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 \( (h^2) \), genetic correlation \( (\rho_{B}) \) 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.

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
