Notation

Environment Variables:

- Stage [Pea Size, Bunch Closure, 19Brix, Harvest]
- Cultivar [Cabernet Sauvignon, Sangiovese]
- Year [2011, 2012]
- Area [Riccione, Montalcino, Bolgheri]

We will use the term *experimental condition* to denote the 48 possible combinations of the 4 variables, e.g. {Stage:Pea Size x Cultivar:Sangiovese x Year:2011 x Area:Montalcino}.

Genetic Variables:

- Gene Expression [normalized fluorescence intensity]

We denote with x_{ijr} the expression of gene i in the jth experimental condition for the rth replicate, with m_{ij} its average expression (over the three replicates) in the jth experimental condition, with m_i its overall average expression:

$$m_{ij} = \frac{1}{3}\sum_{r=1}^{3} x_{ijr} \qquad m_i = \frac{1}{144}\sum_{r=1}^{3}\sum_{j=1}^{48} x_{ijr}$$

In the following the patterns of the gene expressions will be displayed using a 6-panels graphical representation such as, for examples, those visualized in Figure 2. The 6 panels are obtained by crossing variables Area and Year, so that each panel shows the expression patterns in a given Area, at a given Year. In each panel a two-dimensional graph is plotted, with the *x* and the *y* axis representing the levels of Stage and the expression values, respectively. A line shows the pattern of the average expressions m_{ij}. When a single gene is displayed, three points are also present for each Stage level, representing the three replicates. The variable Cultivar is accounted for by means of the color of points and lines (red=Cabernet sauvignon, cyan=Sangiovese).

General description of the pipeline

The gene expressions have been analyzed with data mining procedures, in order to summarize the most important relationships in data, with specific attention to discover the extent to which the variables Stage, Cultivar, Year and Area - separately or in interaction - affect gene expressions.

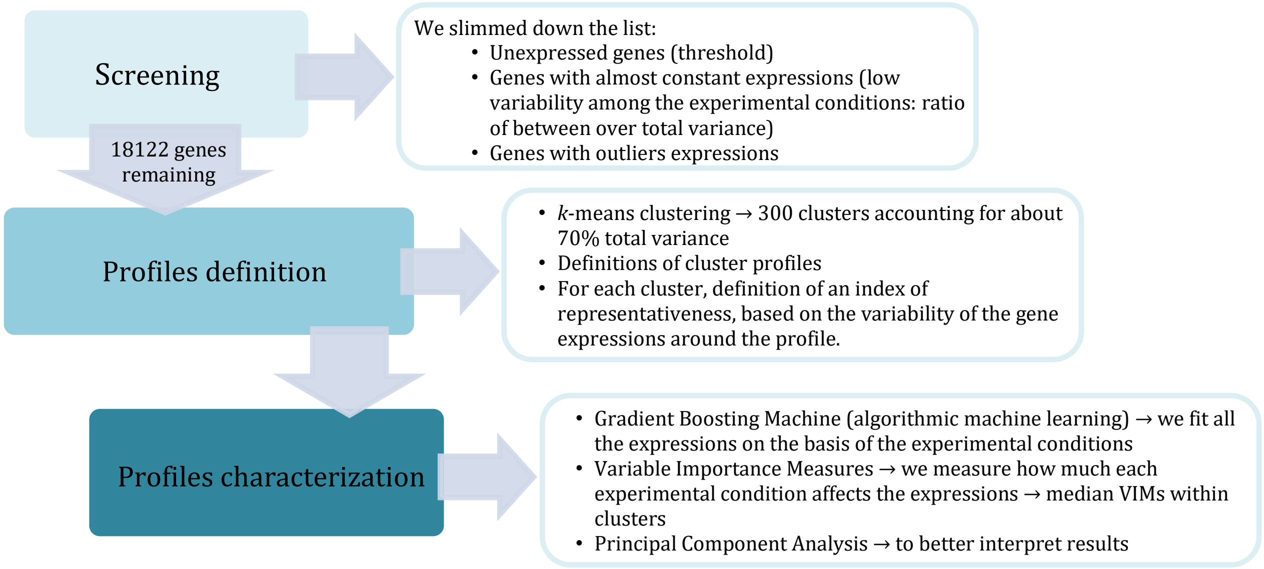

Screening
We slimmed down the list:
• Unexpressed genes (threshold)
• Genes with almost constant expressions (low variability among the experimental conditions: ratio of between over total variance)
• Genes with outliers expressions
18122 genes remaining
Profiles definition
• k-means clustering → 300 clusters accounting for about 70% total variance
• Definitions of cluster profiles
• For each cluster, definition of an index of representativeness, based on the variability of the gene expressions around the profile.
Profiles characterization
• Gradient Boosting Machine (algorithmic machine learning) → we fit all the expressions on the basis of the experimental conditions
• Variable Importance Measures → we measure how much each experimental condition affects the expressions → median VIMs within clusters
• Principal Component Analysis → to better interpret results

Figure 1 – Flow chart of the pipeline

The data mining process was composed of three steps (Figure 1):

1. **Screening:** we identified, out of the whole set of 29549 genes, a subset of 11427 genes with uninteresting patterns with respect to the aim of the present study (unexpressed genes, genes with almost constant expressions, genes with outlier expressions). After deleting this "noisy" subset, the resulting dataset was composed of 18122 genes deserving statistical analysis.
2. **Profiles definition:** in the second step we performed a cluster analysis using the *k*-mean algorithm, in order to reduce the 18122 genes into a number of clusters containing genes with similar patterns among the experimental conditions. We obtained 300 clusters, globally accounting for about 70% of the total variance of the genes expressions. For each cluster we defined an average profile, to be used as a representative of the cluster, and an index of its representativeness, based on the variability of the expressions around the average profile.
3. **Profiles characterization:** in the third step we exploited an advanced machine learning algorithm, the Gradient Boosting Machine (GBM – Friedman, 2001), with the aim of evaluating to what extent each experimental condition affects the gene expression patterns. To this aim Variable Importance Measures (*VIM*s) have been computed, a tool able to describe the impact of given variables on a selected outcome, taking into account the presence of possible, even complex, interactions among the variables themselves. In other words, for each gene, we defined how the variability of its expressions can be accounted for by stage, cultivar, year, site, or an interaction of them, with a multivariate approach. The median *VIM*s within the clusters defined in step 2, allow to characterize the clusters under this point of view, thus determining the relationship between the clusters and the experimental conditions. In the end, the *VIM*s have been crossed with the results coming from a Principal Component Analysis of the cluster profiles, in order to better interpret results.

Step 1: Screening

We identified the genes to remove according to the criteria summarized in Table 1.

Table 1 – Screening guidelines

Phase	Feature	Rule	Number of genes
I	Unexpressed genes	A gene was considered *unexpressed* if the value of its expression is below the threshold 400 in at least 2 of 3 replicates, for all the 48 experimental conditions. Examples of removed genes in Figure 2.	5507
II	Genes with low variability among the experimental conditions	For gene i, the variability of the expressions among the 48 experimental conditions was measured by means of the Pearson ratio $$\eta_i^2 = \frac{\sum_{j=1}^{48}\left(m_{ij} - m_i\right)^2}{\sum_{r=1}^{3}\sum_{j=1}^{48}(x_{ijr} - m_i)^2}$$ Gene i was considered to have a low variability between the experimental conditions if $\eta_i^2 < 0.8$. Examples of removed genes in Figure 3.	4209
III	Genes with outliers	Some genes showed expressions characterized by a high peak in one of the 48 experimental conditions, which in general has to be considered an outlier, i.e. it can be due to errors in recording data or specific conditions. It is reasonable to remove these genes from the analysis. They were identified as follows: *i.* for gene i, let $m_{i(1)}, \cdots, m_{i(48)}$ be the averages $m_{i1}, \cdots, m_{i48}$, sorted in increasing order; *ii.* let $\Delta m_i = m_{i(48)} - m_{i(47)}$ and let SD$_i$ be the standard error of $m_{i(1)}, \cdots, m_{i(47)}$; *iii.* identify a peak in $m_{i(48)}$ if $\Delta m_i > k \cdot SD_i$, where k is a fixed threshold. In our analysis we found that a reasonable value for k was 5. Examples of genes identified with this procedure are given in Figure 4.	238
IV	Genes with low *VIM*s	For gene *i* we computed Variable Importance Measures (*VIM*s) for the 4 variables Stage, Cultivar, Year, Area, using the procedure that will be described later in Step 3. For each variable, the *VIM* measures its contribution to the overall variability of the gene expressions. In other words, it gives an idea of the extent to which the pattern of the gene expressions is affected by the considered variable. In this phase we remove gene *i* if its *VIM*s are all below the following thresholds: *VIM*$_{ih}$(*h* = Stage) < 0.0145612 (30th percentile) *VIM*$_{ih}$(*h* = Cultivar) < 0.0016767 (50th percentile) *VIM*$_{ih}$(*h* = Year) < 0.0012431 (50th percentile) *VIM*$_{ih}$(*h* = Area) < 4.6207353 × 10−4 (50th percentile)	1473

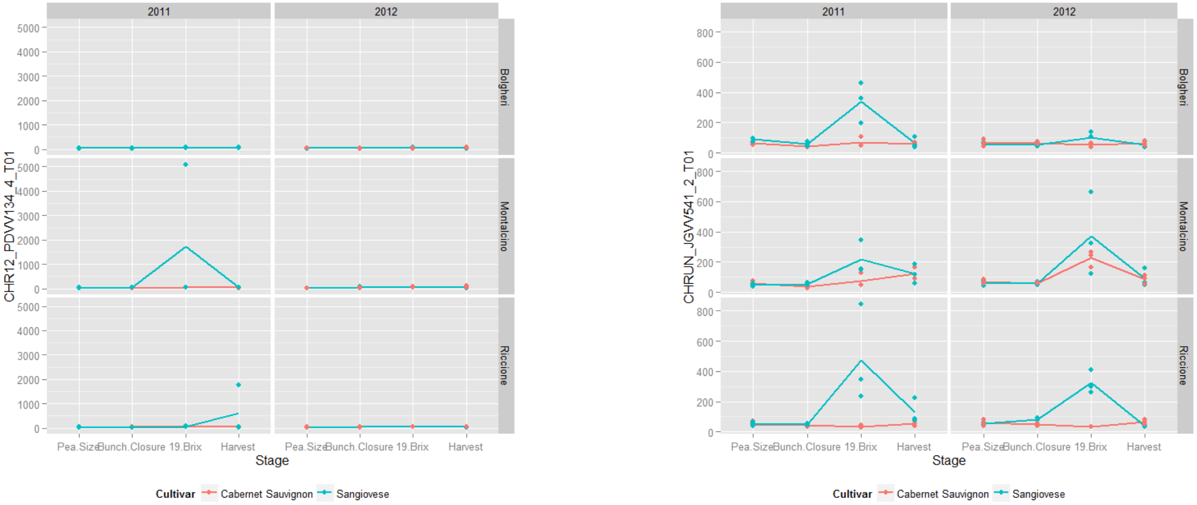


Figure 2 – Step 1, phase I; examples of removed genes.

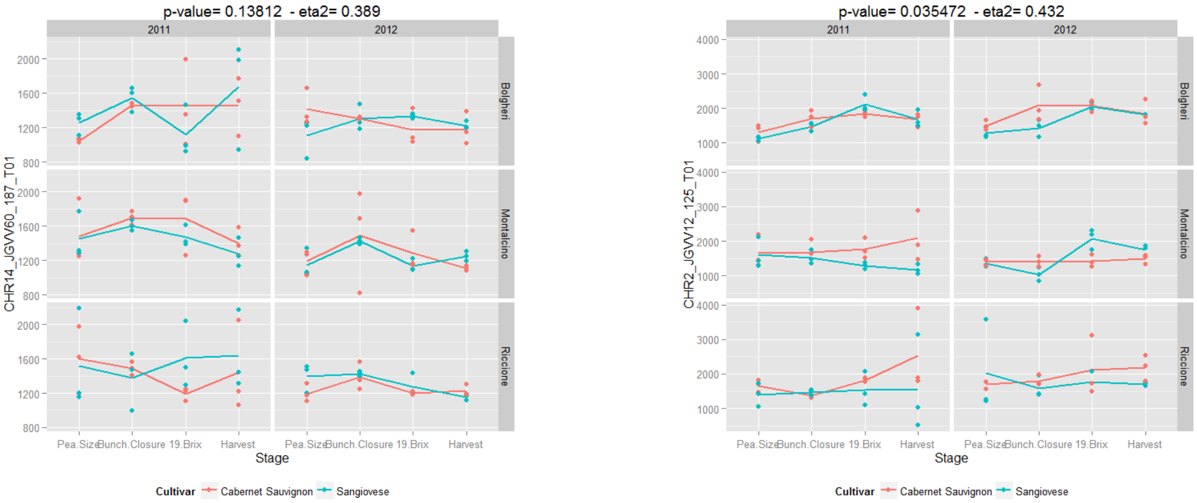


Figure 3 – Step 1, phase II; examples of removed genes.

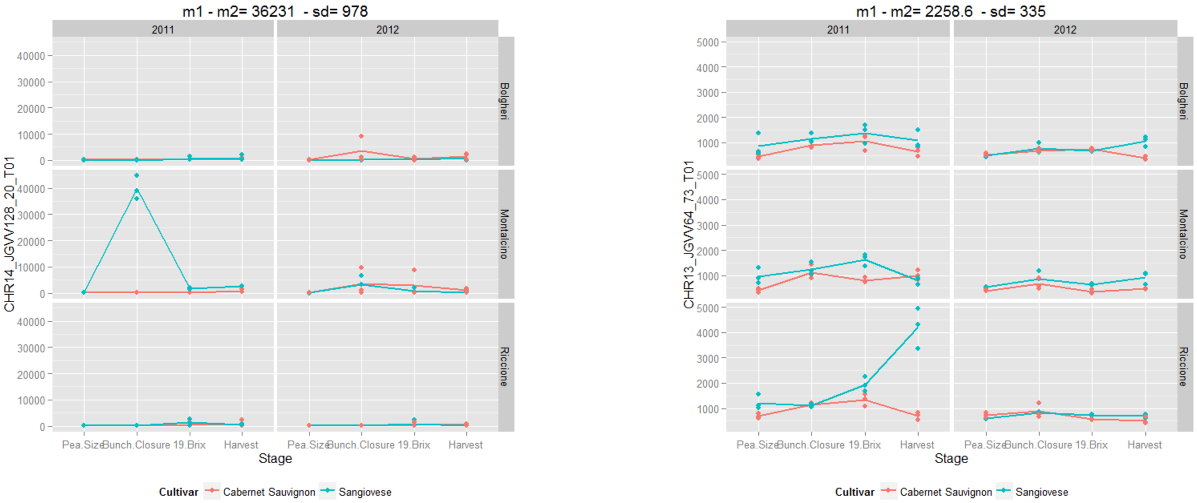


Figure 4 – Step 1, phase III; examples of removed genes.

In the end, 11427 genes were identified for removal and the remaining 18122 genes were selected for the analysis and were passed to the following two Steps.

Step 2: Profiles definition

The data matrix for this step was given by the (18122 x 48) table containing the average expressions m_{ij} of the genes, standardized by rows, that will be denoted $\widetilde{m}_{ij}$ in the following. A cluster analysis was performed with the *k*-means algorithm, in order to define groups of genes with similar patterns along the experimental conditions. After examining different alternatives, the number of cluster was determined so as to explain about 70% of the total variability of the genes expressions. This choice was judged an acceptable tradeoff between the need of a number of clusters as low as possible (easier interpretation) and a good separation between the clusters (high internal cohesion); a set of k = 300 clusters was identified (Figure 5).

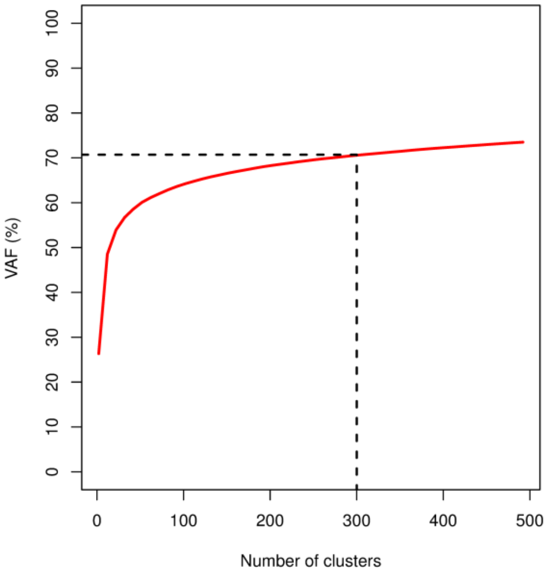


Figure 5 – Variance accounted for (VAF) by clusterization vs number of clusters

For each cluster c, an average profile P_c = (P_{1c}, P_{2c}, …, P_{48c}) was defined, where P_{jc} denotes the average standardized expression of the genes belonging to cluster c in the experimental condition j, to be used as a representative of the cluster:

$$P_{jc} = \frac{1}{n_c} \sum_{i \in c} \widetilde{m}_{ij}$$

where n_c is the number of genes belonging to cluster c. The internal cohesion of each cluster was measured by the following homogeneity index

$$R_c = 1 - \frac{\sum_{j=1}^{48} \sum_{i \in c} \left(\widetilde{m}_{ij} - P_{jc}\right)^2}{\sum_{j=1}^{48} \sum_{i \in c} \left(\widetilde{m}_{ij} - \bar{P}_c\right)^2}$$

where $\bar{P}_c = \frac{1}{48} \sum_{j=1}^{48} P_{jc}$.

Figure 6 shows two examples of clusters with high and low R_c, respectively.

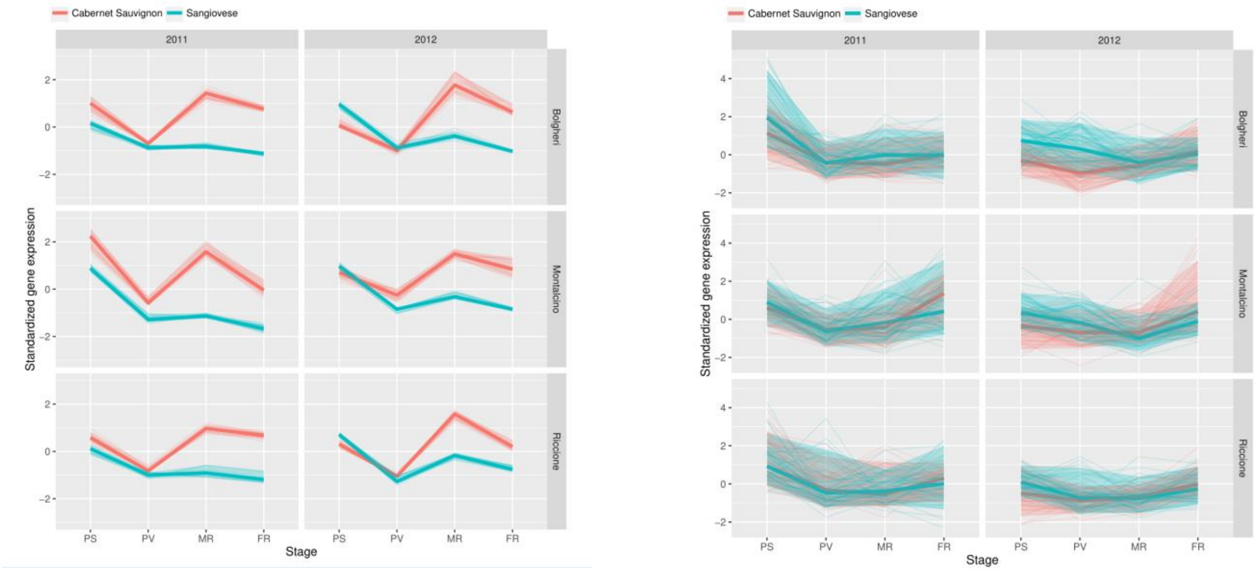


Figure 6 – Step 2; examples of clusters with high (left, cluster no. 299) and low (right, cluster no. 111) homogeneity index R_c (R_c=0.976 and R_c=0.406, respectively). Thin lines represent the expressions of genes belonging to each cluster; bold lines depict the median profiles P_c of clusters.

Step 3: Profiles characterization

Step 3.1: Modelling the relationship between gene expressions and environmental variables

In this step the genes have been analysed one-to-one in order to measure the impact of the experimental conditions on gene expressions. Then, the analysed data were given by the 18122 matrices of dimension (48 x 5), where each row describes one experimental condition according to the variables Stage, Cultivar, Year, Area, Gene Expression. The idea was to fit a predictive model to the outcome variable Gene Expression, with the environmental variables as predictors, in order to inspect, for each gene, the extent to which environmental variables affect the gene expression. From a statistical point of view, two main issues have to be taken into account in setting the predictive model: the possible nonlinearity of relationships between the outcome and the predictors and the possible presence of interactions among the predictors in affecting the outcome. At the same time, we are not only interested to prediction itself but also to the extrapolation of the role played by covariates, for interpretative purposes. This is the main reason why, in spite of the need of complex models, it is often inopportune to take advantage of powerful nonlinear regression techniques such as, for example, neural networks, which, in change of a significant predictive accuracy, are impenetrable black-boxes. In the last decades the predictive and interpretative power of statistical tools has been greatly improved by the introduction of data mining algorithms able to deal with complex relationships among variables from a multivariate perspective. These new algorithms generate, together with predictions, variable importance measures (*VIM*s) identifying the most important predictors of the outcome. Within this innovative approach, called algorithmic modelling (Breiman, 2001a), a great deal of attention has been devoted to the ensemble learning philosophy (Friedman, 2003, 2006; Friedman and Popescu, 2003, 2005). Learning ensembles are sequences of ensemble members, i.e. statistical models predicting the outcome as a function of the set of predictors. Final predictions are obtained by a linear combination of the prediction of each ensemble member. Learning ensembles can be built using different prediction methods, i.e. different base learners as ensemble members. The most interesting proposals use decision trees (more specifically CART, Classification And Regression Trees - Breiman et al., 1984) as base learners and are called tree-based learning ensembles. Popular examples are the Random Forest technique (RF - Breiman, 2001b) or the tree-based Gradient Boosting Machine (GBM - Friedman, 2001). Both these algorithmic techniques identify the most important predictors within the set of covariates, by means of the computation of *VIM*s. In this case study we have used GBM to fit each gene expression as a function of the environmental variables. Then, for each gene we have been able to compute 4 *VIM*s, measuring the extent to which each environmental variable affects the gene expression, on its own or jointly/in interaction with the others.

Let VIM_{ih} (h=1,2,3,4) be variable importance measure of the hth environmental variable for gene *i*. Firstly, we used the *VIM*s to identify genes whose expression is almost unaffected by environmental variables, in order to remove them from the analysis (see Step 1, phase IV). Then, we computed the average *VIM*s for the 300 clusters obtained in Step 2

$$VIM_h^c = \frac{1}{n_c} \sum_{i \in c} VIM_{ih}$$

and used them to characterize the cluster profiles according to the relationship between gene expression and environmental variables (see the example of cluster no. 202 in Figure 7).

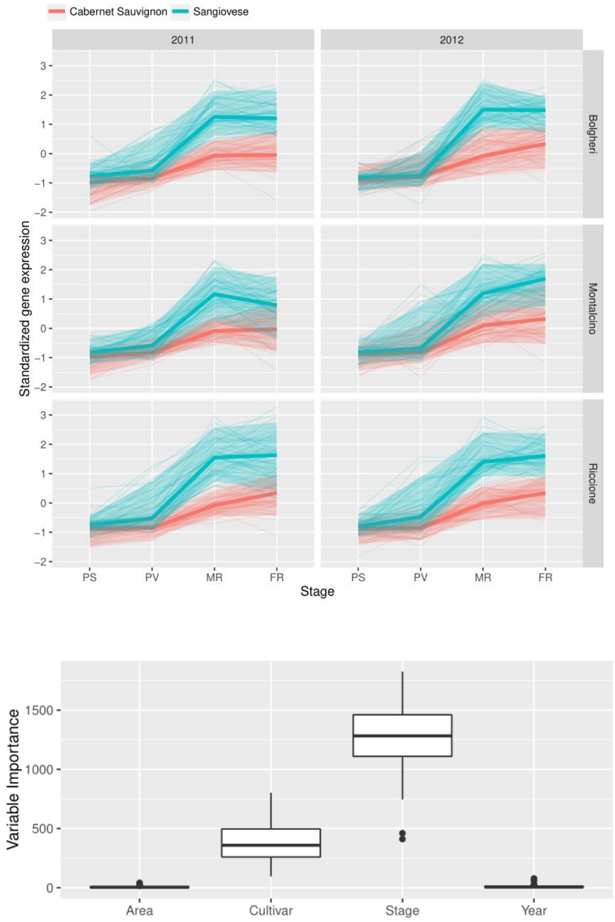


Figure 7 – Expression profiles (top) and boxplots of VIMs (bottom) for cluster no. 202.

Step 3.2: Principal Component Analysis of the cluster average profiles

Subsequently, we performed a dimensionality reduction via Principal Component Analysis (PCA) of the (48 x 300) matrix containing, by columns, the average profiles $P_c = (P_{1c}, P_{2c}, \ldots, P_{48c})$ of the 300 clusters.

Figure 8 displays some selected two-dimensional objects scores plots, which highlight that the Principal Components, computed as linear combinations of the cluster profiles, are able to discriminate the environmental variables characterizing the 48 experimental conditions with a remarkable accuracy.

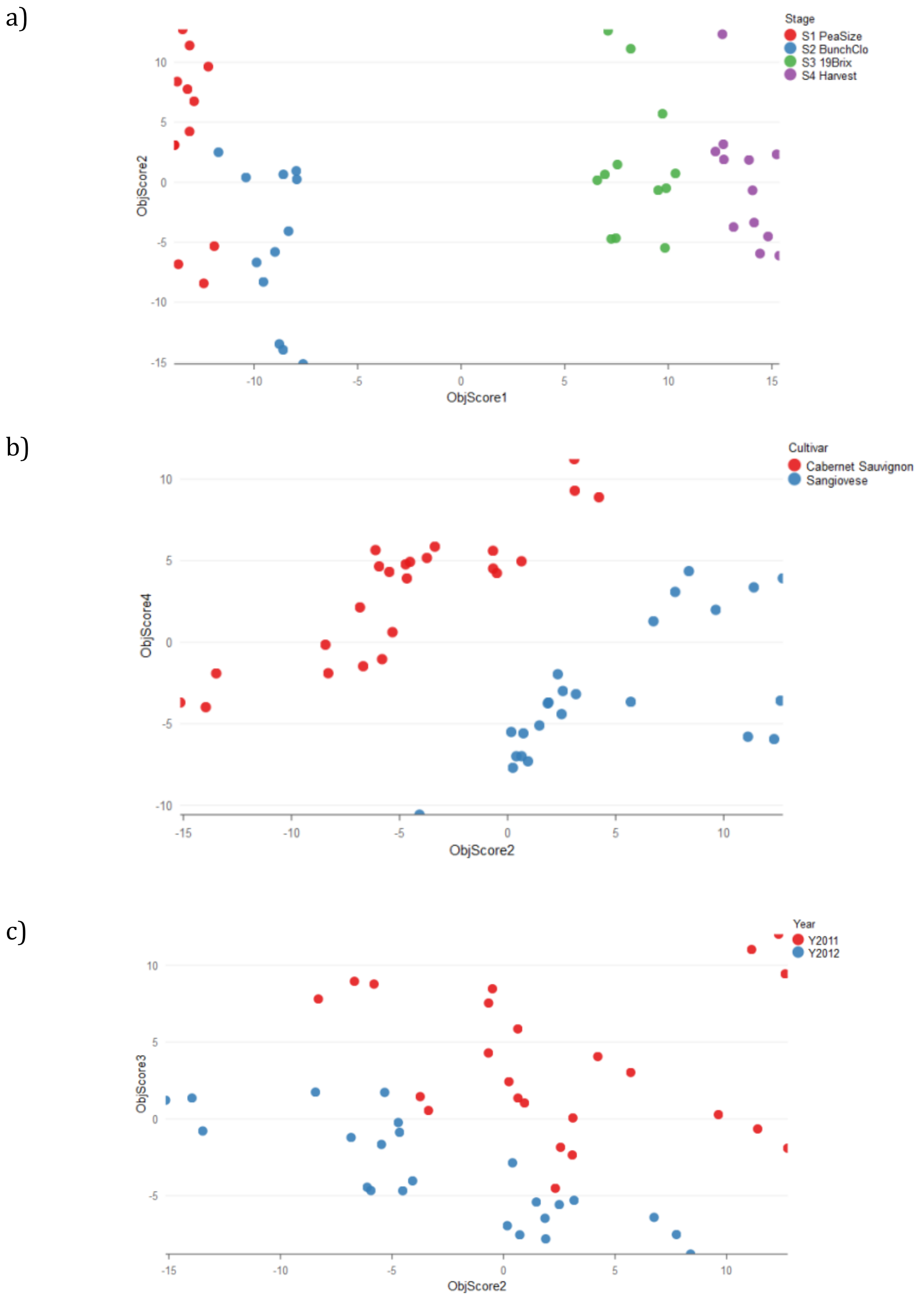


Figure 8 – Selected two-dimensional object scores plots of the 48 experimental conditions (points coloured according to the environmental variables)

Step 3.3: Combining PCA, VIMs and cluster homogeneity index

In order to gain an immediate and comprehensive insight on the cluster profiles, we provide in the Supplementary Material a set of interactive graphical tools for the investigations of the characteristics of the 300 clusters according to several variables. Figures 9 to 12 show some noteworthy examples.

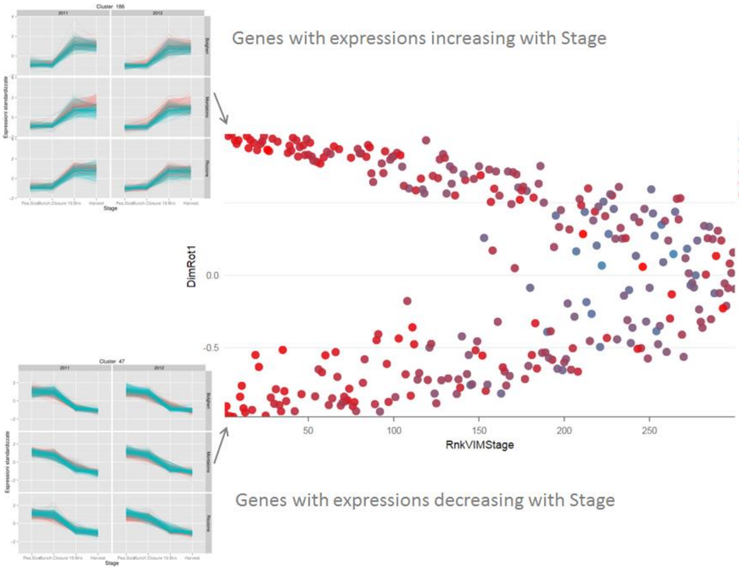
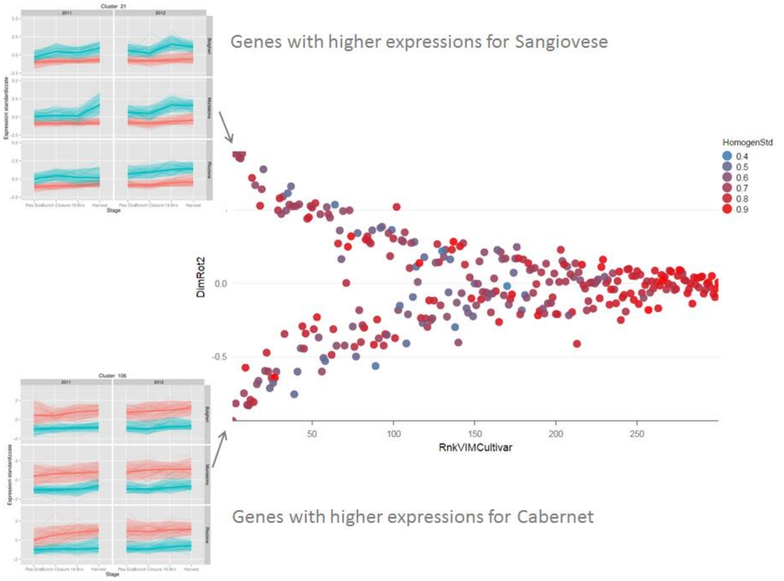


Figure 9 – Scatterplot of the 300 clusters according to Rank in VIM^{c}_{Stage} and loading in the first rotated Principal Component (points coloured by cluster homogeneity index R_c).

Figure 9 shows that the loadings of the clusters in the first rotated Principal Component (DimRot1) are associated with the importance of the variable Stage (RnkVIMStage = rank of clusters according to VIM^{c}_{Stage}; low values denote high importance of Stage). Specifically, the higher the importance of Stage, the higher the loadings (in absolute value). The expressions of genes belonging to clusters with positive loadings exhibit a positive trend with respect to Stage (i.e. a pattern with increasing values as stage moves forward), whereas the opposite happens for genes with negative loading. In addition, clusters with high absolute loading are characterized by excellent homogeneity (represented by the colour of points, with red denoting the highest values of the index R_c). This graph is a valid tool for the identification of clusters composed by genes with expressions highly associated to Stage, and with very similar patterns within the clusters..

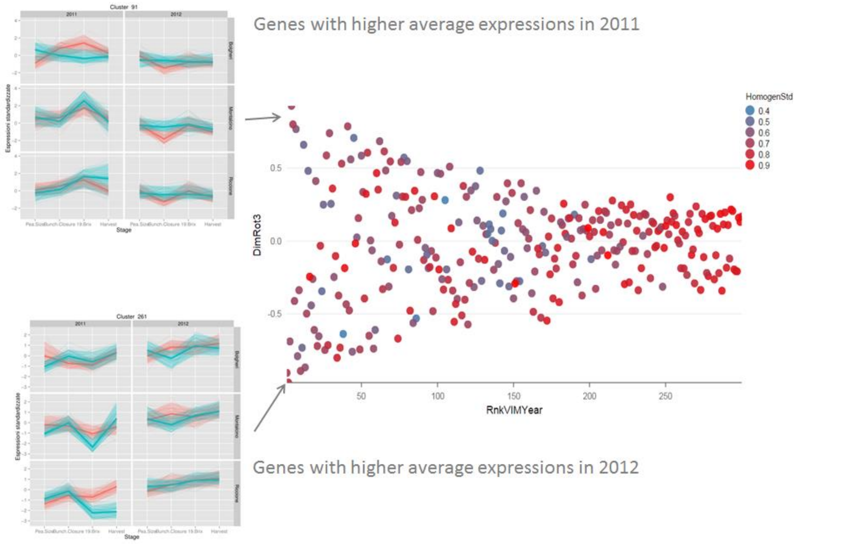
Similar to the preceding figure, Figure 10 shows that cluster loadings in the second rotated Principal Component (DimRot2) are associated with the importance of the variable Cultivar (RnkVIMCultivar = rank of clusters according to $VIM^{c}_{Cultivar}$; low values denote high importance of Cultivar for that cluster). Also in this case, the higher the importance of Cultivar, the higher the absolute values of the loadings. The expressions of genes belonging to clusters with positive loadings exhibit a median higher expression for Sangiovese, whereas the opposite happens for genes with negative loadings. Clusters with high absolute loading are characterized by good homogeneity, but lower than the homogeneity of clusters associated to the variable Stage. This graph helps to identify clusters composed of genes associated to the Cultivar, with quite similar patterns within the clusters.



Genes with higher expressions for Sangiovese

Genes with higher expressions for Cabernet

Figure 10 – Scatterplot of the 300 clusters according to Rank in $VIM^{c}_{Cultivar}$ and loading in the second rotated Principal Component (points coloured by cluster homogeneity index R_c).



Genes with higher average expressions in 2011

Genes with higher average expressions in 2012

Figure 11 – Scatterplot of the 300 clusters according to Rank in VIM^{c}_{Year} and loading in the third rotated Principal Component (points coloured by cluster homogeneity index R_c).

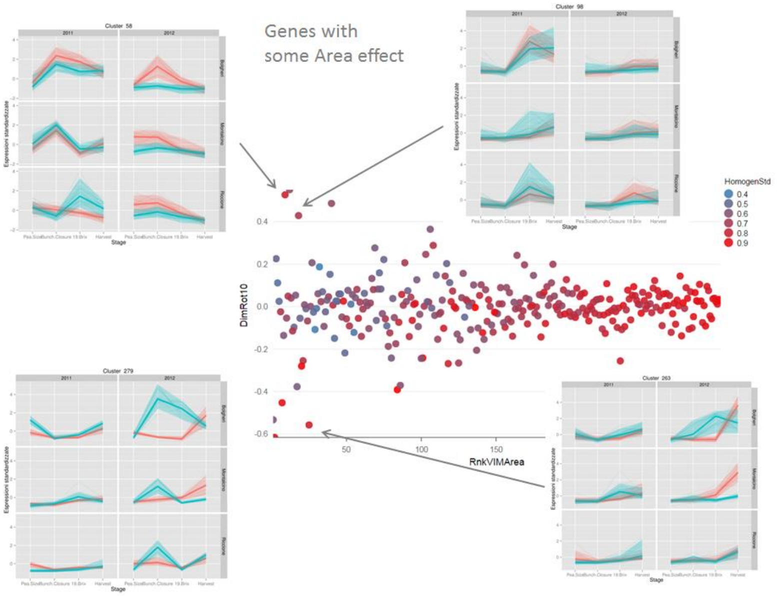


Figure 12 – Scatterplot of the 300 clusters according to Rank in VIM^{c}_{Area} and loading in the tenth rotated Principal Component (points coloured by cluster homogeneity index R_c).

Figures 11 and 12 can be interpreted analogously to Figures 9 and 10, with reference to Year and Area, respectively. It should be noted that the association of loadings (DimRot3 and DimRot10, respectively) loadings and variable importance is weaker for these two variables, that the homogeneity within clusters is markedly lower and that for variable Area we have to reach the tenth Principal Component to find some noteworthy association between loadings and variable importance. The patterns of gene expressions in these cases cannot easily be summarized with some standard behaviours, as they are more complex and their effects are often activated in interaction with other factors. This finding confirms the appropriateness of the choice of the GBM algorithm and its ability to take into account interactions among variables in determining the VIMs.

Each interactive graph also allows the researcher to easily obtain reference information about the clusters he/she is interested to (Figure 13).

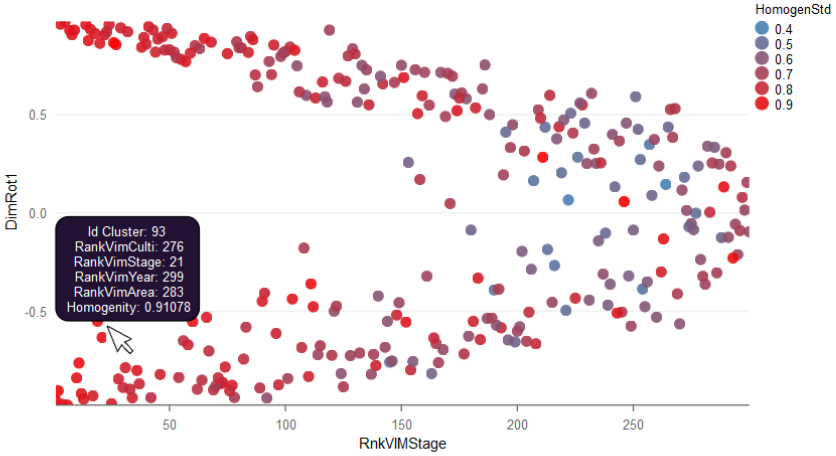


Figure 13 – Selection of one cluster in the scatterplot of the 300 clusters according to Rank in VIM^{c}_{Stage} and Loading in the first rotated Principal Component.

To sum up, thanks to these interactive graphs we are able to

- identify variable-dependent clusters, i.e. clusters composed of genes whose expression is strongly affected by one specific environmental variable;
- characterize their pattern of expression (e.g., increasing or decreasing with Stage, higher or lower for one Cultivar, …);
- evaluate the degree of cluster homogeneity;
- investigate the existence of clusters characterized by the interaction of two or more environmental variables.

References

Breiman L., (1996a), The heuristic of instability in model selection, Annals of Statistics, 24, 6, 2350-2383.

Breiman L., (1996b), Bagging Predictions, Machine Learning, 24, 2, 123-140.

Breiman L., (2001a), Statistical Modeling: The Two Cultures, Statistical Science, 16, 3, 199-231.

Breiman L., (2001b), Random Forests, Machine Learning, 45, 1, 5-32.

Breiman L., Friedman J.H, Olshen R.A. and Stone C.J. (1984), Classification and Regression Trees, Chapman & Hall, New York.

Friedman J.H. (2001), Greedy function approximation: a gradient boosting machine, Annals of Statistics, 29: 1189-1232.

Friedman, J.H. (2006) Recent advances in predictive (machine) learning, Journal of classification 23(2): 175-197.

Friedman, J. H. (2006), Separating signal from background using ensembles of rules, In Lyons L. and Ünel M.K. (eds.), Statistical Problems in Particle Physics, Astrophysics and Cosmology, Proceedings of PHYSTAT05, Imperial College Press, London

Friedman, J. H., and Popescu, B. E. (2003), Importance Sampled Learning Ensembles, Stanford University, Department of Statistics, Technical Report.

Friedman, J. H., and Popescu, B. E. (2008), Predictive learning via rule ensembles, The Annals of Applied Statistics, 2 (3): 916-954