Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. *ssp sativa*) berry: a QTL to QTN integrated study

Fournier-Level Alexandre¹, Le Cunff Loïc¹, Gomez Camila², Doligez Agnès¹, Ageorges Agnès², Roux Catherine¹, Bertrand Yves¹, Souquet Jean Marc², Cheynier Véronique², This Patrice¹*

¹INRA UMR 1097 DIAPC Diversité et Adaptation des Plantes Cultivées, 2, place Viala F-34060 Montpellier, France
²INRA UMR Sciences pour l'œnologie, 2, place Viala F-34060 Montpellier, France

*corresponding author: Patrice This, tel: +33 4 99 61 22 92, fax: +33 4 99 61 20 64 e-mail: this@supagro.inra.fr

Other authors’ e-mails

fourniel@supagro.inra.fr; lecunff@supagro.inra.fr; gomezca@supagro.inra.fr; doligez@supagro.inra.fr; ageorges@supagro.inra.fr; rouxc@supagro.inra.fr; bertrany@supagro.inra.fr; souquet@supagro.inra.fr; cheynier@supagro.inra.fr.
Running title: Grape color quantitative genetics

Keywords: anthocyanin, transcription factor, QTL, association genetics, QTN, wine

Corresponding author: Patrice This,

INRA UMR 1097 Diversité et Adaptation des Plantes Cultivées,
2, place Viala F-34060 Montpellier, France
Tel: +33 4 99 61 22 92,
Fax: +33 4 99 61 20 64
E-mail: this@supagro.inra.fr
The combination of QTL mapping studies on synthetic lines and association mapping studies on natural diversity represents an opportunity to throw light on the genetically-based variation of quantitative traits. With the positional information provided through QTL mapping, which often leads to wide intervals encompassing numerous genes, it is now easily feasible to directly target candidate genes that are likely to be responsible for the observed variation in completely sequenced genomes and test their effects through association genetics. This approach was performed in grape, a newly sequenced genome, in order to decipher the genetic architecture of anthocyanin content. Grapes may either be white or colored, ranging from the lightest pink to the darkest purple tones according to the amount of anthocyanin accumulated in the berry skin, which is a crucial trait for both wine quality and human nutrition. Although the determinism of the white phenotype has been fully identified, the genetic bases of the quantitative variation of anthocyanin content in berry skin remain unclear. A single QTL responsible for up to 62% of the variation in the anthocyanin content was mapped on a Syrah x Grenache F1 pseudo-testcross. Among the 68 unigenes identified in the grape genome within the QTL interval, a cluster of four Myb-type genes was selected based on physiological evidences ($VvMybA1$, $VvMybA2$, $VvMybA3$, and $VvMybA4$). On a core collection of natural resources (141 individuals), 32 polymorphisms revealed significant association, and extended LD was observed. Using a multivariate regression method, we demonstrated that five polymorphisms in $VvMybA$ genes except $VvMybA4$ (one retrotransposon, 3 SNPs and one 2bp Indel) accounted for 84% of the observed variation. All these polymorphisms either led to structural changes in the MYB proteins
or differences in the $VvMybA$s promoters. We concluded that the continuous variation in anthocyanin content in grape was mainly explained by a single gene cluster of three $VvMybA$ genes. The use of natural diversity helped to reduce one QTL to a set of five QTNs and gave a clear picture on how isogenes combined their effects to shape grape color. Such analysis also illustrates how isogenes combine their effect to shape a complex quantitative trait and enables the definition of markers directly targeted for upcoming breeding programs.
Introduction

In the past decade, research on plant quantitative trait loci (QTL) has successfully identified numerous loci that control the genetic variation of complex traits in plants. Price (2006) reported no fewer than 30 studies of successful QTL cloning concerning 9 plant species, but these researches have only been conducted in biparental populations, revealing only a slice of the genetic architecture for the trait (Holland 2007). Now that we are entering the post-genomic era, the challenge for geneticists is to fully decipher the molecular bases of quantitative genetic variation in highly diversified resources and integrate the existing phenotypic variation extensively (Nordborg and Weigel 2008). The widespread availability of plant genomic and genetic resources has triggered the need for more integrated research (Flint-Garcia et al. 2003) that is likely to combine the findings of different approaches from various experimental designs, as has already been done for humans (Hirschhorn and Daly 2005), animals (Ron and Weller 2007) and plants (Osterberg et al. 2002; Aranzana et al. 2005). The combination of linkage and association genetics for this purpose constitutes a powerful tool (Yu and Buckler 2006) and newly considered model crops such as grape will greatly benefit from these advances.

The number of grape genomic resources has increased considerably over the past few years and the sequencing of the grape genome has recently been completed (Jaillon et al. 2007; Velasco et al. 2007). Taking advantage of these new resources will enable considerable progress in complex trait dissection, given that the access to candidate genes is straightforward, and will allow cloning of QTLs. Grape shows extended genetic variation with a high level of linkage disequilibrium (Barnaud et al. 2007) that
makes an association genetics strategy feasible, as has already been performed in model
(Flint-Garcia et al. 2005) or other perennial plants (Gonzalez-Martinez et al. 2007).

The variation in anthocyanin content is responsible for the continuous reddish color
tones of most plant species (grape, apple, petunia, sweet potato, snapdragon, and
Arabidopsis). In plant secondary metabolism, the anthocyanin pathway is one of the
best described: the genes coding structural enzymes have been cloned in different plant
systems (Sparvoli et al. 1994; Holton and Cornish 1995; Boss et al. 1996). Various
mutated genes affecting both enzymes and transcription factors have been reported in
maize, snapdragon, strawberry and petunia (Mol et al. 1999; Aharoni et al. 2001;
Winkley-Shirley 2001 and Quattrocchio et al. 2006). Grape, Vitis vinifera L., is an
important edible source of concentrated anthocyanins; their importance for final wine
quality (Flanzy 1998) and their antioxidant benefits for human health (Joseph et al.
2005) call for the need to increase the accumulation of anthocyanin in edible plants.

In the case of grape, extensive molecular physiology studies provide evidence that two
adjacent transcription factors, VvMybA1 and VvMybA2, are able to induce the VvUFGT
transcription needed for berry pigmentation (Ageorges et al. 2006; Walker et al. 2007).
Recently, a last transcription factor, VvMyb5b, was also shown to marginally induce the
VvUFGT (Deluc et al. 2008). Furthermore, white/colored variation in grape
cosegregates as a monogenic locus with the VvMybA1 locus (Doligez et al. 2006;
Lijavetzky et al. 2006). The white grape phenotype has been linked to the homozygous
presence of a transposable element, Gret1, in the promoter of the VvMybA1 locus
(Kobayashi et al. 2002, 2005; Lijavetsky et al. 2006; This et al. 2007). The quantitative
variation in anthocyanin content in the skin of the berry was also shown to display
considerable fluctuation (Mazza 1995) but despite its economic importance, the
The determinism of anthocyanin accumulation in berry skin has until now not been elucidated in grape.

For the analysis of genetic variation of small effect such as traditionally observed for QTLs, traditional molecular physiology resources have reached their limit. Firstly, because transient assays in grape only allow to validate the effect of a single mutation in a bimodal way (Torregrosa et al. 2003; Vidal et al. 2006) and stable transformation is both labor- and time-consuming in the case of a long cycle crop such as grape (Bouquet et al. 2008). Secondly, if the overall genetic effect is due to many combined mutations, it may affect either the level of expression of a gene when located in the promoter or the functionality of the protein when located in the gene coding sequence. Both modifications can hardly be analyzed in single transformation experiments. The overall resulting genotypic effect can only be determined through linkage studies (Neale and Savolainen 2004). A QTL mapping approach in which no assumptions are made about the architecture of the trait and that provides accurate positional information is a powerful tool for identifying loci that control phenotypic variability (Salvi and Tuberosa 2005). Previous studies have illustrated the feasibility of cloning genes that underlie QTLs, but starting with the initial QTL detection on RIL population, these researches often took five to seven years, as in the case of the *Tb1* locus in Maize (Doebly et al. 1993; Doebly et al. 1997), the *COL1* gene in Brassica (Lagercrantz et al. 1996; Osterberg et al. 2002), the *fw2.2* gene in Tomato (Alpert et al. 1995; Frary et al. 2000) or the *ERECTA* gene in Arabidopsis (Alonso-Blanco et al. 1998; Masle et al. 2005). In upcoming years, the process from QTL detection to the QTN identification and soon, the functional validation of the effect of the identified polymorphisms, will become immediate even in the case of perennials such as grape, thanks to a more simplified
molecular data analysis. Finally, the quality of quantitative genetics studies will depend more on the careful definition of the experimental design and the acquisition of accurate phenotypes for the sample in question.

Relevant work concerning the genetic bases of flavonoid and phenylpropanoid biosynthesis have involved maize, for which abundant genetic resources are available. Great efforts have been done to fully dissect the loci underlying the genetic variation for maysin accumulation (McMullen et al. 1998; Lee et al. 1998). The QTL identified gave rise to a very extensive survey of loci controlling mayzin accumulation. This involved the use of both traditional molecular genetics tools as for the identification of the orange1 locus with epigenetic effect (Chopra et al. 2003) and association genetics technique for the p and c2 loci (Szalma et al. 2005). This revealed a complex regulatory mechanism involving both transcription factors and enzymes of the flavonoid pathway.

Anthocyanin, as one of the main flavonoid in grape, displays a continuous variation among cultivars, its genetic bases being thus necessarily affected by many genes and/or many mutations of small effect. The purpose of this study was to describe the genetic bases of quantitative variation of anthocyanin in grape based on the fewest possible assumptions about the genetic architecture of the trait. The most straightforward method to deal with the genetic architecture of a trait is to carry out QTL mapping as successfully done in grape (Doligez et al. 2002; Mejia et al. 2007). Furthermore, thanks to the extensive information obtained with the sequencing of the grape genome, we aimed to refine the findings of our QTL mapping strategy with a candidate gene approach through the use of an association study on a collection of genetic resources of great diversity, as has already been done for plant color with other models (Szalma et al. 2005, Chagné et al. 2007). This will allow the definition of genic markers believed to be
the cause of the variation instead of targeting linked neutral markers in breeding perspectives.

Nonetheless, these methods often lead to a huge number of positive associations, especially in the presence of linkage disequilibrium (LD), which makes the genetic effect of each polymorphism confusing. We established an original procedure for selecting the most stringent non-redundant associations. Moreover, as the associated polymorphisms are randomly combined in planta to shape a trait, we aimed to establish a single multivariate model that would provide an optimal fit of the anthocyanin berry content variation at the genotype level. Subsequently, a set of five putative quantitative trait nucleotides (QTNs), corresponding to one retrotransposon insertion, one 2bp indel and three SNPs, was selected and integrated into a multivariate model that accounts for 84% of the anthocyanin content of the highly diversified collection of cultivars. This study enabled us to explain most of the variation in grape color based on only the information of five polymorphisms on three distinct genes within a single gene cluster.

Materials and Methods

Plant materials

The plant material consisted of two populations: one cross-derived mapping population for QTL mapping and one natural population for the association study. The mapping population (SxG) was a F1 progeny of 191 individuals from a reciprocal cross between clone 73 of Syrah (S) and clone 516 of Grenache (G). Each offspring genotype was randomly displayed on two blocks (A & B). Both Syrah and Grenache are typical cultivars in the south of France. The natural population sample is a core collection (CC)
of 141 individuals, which maximizes the agromorphological diversity for 50 qualitative and quantitative traits (Barnaud et al. 2006).

**Phenotyping**

Grapes were harvested at maturity (20° Brix): (1) in 2005 in the SxG A block and in the core collection, and (2) in 2006 in both SxG A and B blocks and in the core collection. Eight representative clusters were harvested among the clones. Twenty-five berries with a density of between 130 and 160 g/L were randomly selected for further analysis. Their skins were powdered under liquid nitrogen and anthocyanins were analyzed by high performance liquid chromatography (HPLC) according to Fournand et al. (2006).

Total anthocyanin content was then log-transformed with a Log(1+x) function in order to unskew their distribution whilst conserving the power of the 0 class. The normality of the distribution was checked using the Shapiro-Wilks test for individuals carrying anthocyanins (non-white cultivars). To improve extraction of the genetic variance components, the data from 2005 and 2006 were treated together using the mixed procedure of SAS software (SAS INSTITUTE, NC, USA, www.sas.com) in order to extract the best linear unbiased predictor (BLUP) for each genotype.

**DNA extraction, genotyping and sequencing**

One square inch (80-100 mg) of fresh young leaf was harvested for each genotype. DNA was extracted using a Qiagen DNA Plant Mini Kit (QIAGEN S.A., Courtaboeuf, France, www.qiagen.com) with slight modifications as described in Adam-Blondon et al. (2004).
The 191 SxG offspring were genotyped for 97 SSR markers. Marker selection was carried out using both (1) their position on reference maps (Adam-Blondon et al. 2004; Doligez et al. 2006) to cover the genome with minimal inter-marker space (maximum of 10cM) and (2) the most informative polymorphisms (priority given to 1:1:1:1 segregating markers).

SSR markers were genotyped as described by Doligez et al. (2006). Amplified fragments were analyzed with an ABI PRISM 3100 Genetic Analyzer (APPLERA, Norwalk, CT, USA, http://www3.appliedbiosystems.com/index.htm). An additional indel corresponding to a 10 Kbp Gret1 retro-element in the MybA1 locus promoter was genotyped as previously described (Kobayashi et al. 2002; This et al. 2007). PCR conditions were identical to those described by Kobayashi et al. (2002). Amplified fragments of the two pairs described above were then bulked and run together on a 1% agarose gel, stained with Ethidium-Bromide and photographed under UV light.

Amplification primers were designed using the Primer 3 software and are listed in supplementary Table S1. PCR fragments were amplified, sequenced, and analyzed as described by Le Cunff et al. (2008).

Framework genetic maps

Segregation distortions for parental and consensus data were assessed with chi-square tests. The parental maps were constructed according to the pseudo-testcross strategy (Grattapaglia and Sederof 1994) using Carthagene 0.999R software (de Givry et al. 2005). The best marker order was determined starting with a first order and then optimizing this order. Finally, a sliding window likelihood calculation was used to detect local changes, and markers with uncertain order at LOD 2 were discarded. A
consensus map was built using the same procedure. Kosambi’s mapping function was
applied to all maps for the computation of genetic distance.

**QTL detection**

QTL detection was performed on both parental and consensus maps. Composite interval
mapping (CIM) on total anthocyanin content was performed separately on the 2005 A,
2006 A, and 2006 B block data with MapQTL 4.0 (Van Ooijen et al. 2002) for the
consensus map, and QTL Cartographer (Basten et al. 2001) for parental maps. For CIM
on the parental maps, we first used the Forward & Backward regression method for
cofactor selection with 0.1 as the in-and-out threshold for the P-value of the partial F-
test. Then, a genome scan was performed with a maximum of five cofactors within a 10
cM window. LOD thresholds corresponding to an experimentwise error rate of 5% were
then determined through 1000 permutations. For CIM on the consensus map, we
determined the simple interval mapping LOD thresholds through 1000 permutations
with a genome-wide error rate of 5%. Confidence intervals of the first and second rank
of the QTL’s position were determined as one- and two-LOD support intervals.

**Local blast**

The intervals defined by QTL mapping were blasted on the grape’s genome browser
(http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis) and supercontigs were
isolated. A local blast on the NCBI UniGene set was performed on the same
supercontigs to determine the total number of genes present in the QTL intervals
Haplotype reconstruction

Haplotypes of the \textit{VvMybA} genes have been reconstructed using a Partition-Ligation Expectation Maximization (PLEM) algorithm described in Qin \textit{et al.} (2002) and implemented in PHASE v2.1 (Stephens and Donelly 2003), using a 200 burn-in with 200 iterations in total and a thinning interval of 1; this was repeated 10 times until convergence was validated. The algorithm was run again on the most highly associated polymorphisms to reconstruct an entire macrohaplotype combining the linkage signal of all three genes.

Association tests

Association tests were carried out using SAS and TASSEL software packages (Buckler, http://www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=). To test for associations, we used the structured association method (Thornsberry \textit{et al.} 2001). Firstly, population structure was calculated via a Bayesian approach implemented in the STRUCTURE software (Pritchard \textit{et al.} 2000) that used 20 SSR markers well scattered throughout the 19 grape linkage groups (Lacombe \textit{et al.} 2007). 500,000 iterations were performed for each population number between one and ten with a burn-in period of 500,000. The optimal subpopulation model was selected using Evanno’s correction (Evanno \textit{et al.} 2003). The best population subdivision was obtained for K=2 subpopulations and the corresponding Q matrix was used for association analyses. Secondly, kinship was calculated in two ways, as implemented in TASSEL according to Ritland’s calculation (1996): first by using the 20 SSRs described above, then by using 129 SNPs from 10 unigenes of the anthocyanin pathway (\textit{VvAM1}, \textit{VvAM2}, \textit{VvAM3}).
A naïve General Linear Model test, a structured association test and two Mixed Linear
Model tests using the two different kinship matrices were performed with TASSEL on
each gene to identify the highly associated polymorphisms; this was based on P-values
for all models and on the adjusted \( R^2 \) of the GLMs (TASSEL). The linkage
disequilibrium calculations were performed using the LD option implemented in
TASSEL, and the effect of the gene on the LD level was tested using the ANOVA
procedure in SAS. Only a reduced set of closely associated and non-redundant markers
was selected based on a stepwise cofactor selection of the GLM procedure with partial
risk set to \( \alpha=0.01 \).

**Regression model selection**

The genotypes of the most closely associated polymorphisms selected above were
broken down into additive (allele doses) and dominance (heterozygote versus
homozygote contrast) effects. The set of polymorphisms was then entered in a
multivariate regression model using the REG procedure of SAS software. First, the
population structure effect was included as an initial effect. Second, for model ranks
ranging from 1 to \( n \) covariates, \( n \) being the number of covariates selected in the previous
step, we calculated all alternative models and selected the best model of each rank based
on the adjusted \( R^2 \). Finally, the model with the optimal number of parameters was
selected according to both Akaike and Bayesian (Schwarz) information criteria.

**Results**
QTL mapping

Syrah x Grenache (SxG) sample showed extensive variation ranging from 0 mg of anthocyanin for white cultivars to a maximal content of 31.6 mg of anthocyanin per g of fresh skin with a variance of $\sigma_{SxG}=0.40$ (data for 2005 shown in Figure 1). In the SxG progeny, 30% of the individuals were white/green and 70% were dark berried cultivars showing all 15 anthocyanin compounds, suggesting the 1:3 segregation of a major locus. QTL mapping was performed on the consensus map and on both parental maps, separately on each block for each year and collectively using Best Linear Unbiased Predictors (BLUPs). In all detections, the BLUPs tended to improve both the score intensity and the position accuracy of the QTL (Table 1).

A single QTL located on LG2 between the *VMC6B11* and *VVIU20* markers was identified in all analyses with a 5.9 cM confidence interval at LOD-1 defined on the consensus map (Table 1). This locus accounted for 48 to 62% of the total variation in anthocyanin content in the berry and was repeated across all blocks and years. It was restricted to a 1.7 cM interval on the Grenache map, centered on the *VvMybA1 Gret1* and accounting for 47% of total variation, and to a wider interval of 9.2 cM on the Syrah map located between the *VMC6B11* marker and the *VvMybA1* locus, accounting for 54% of total variation.

Candidate genes and polymorphisms within the QTL interval

The 5.9 cM confidence interval at LOD-1 corresponded to a physical region of 2.85 Mbp. This segment was anchored on scaffold 97 of LG2 from the grape reference genome sequence (http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis). The
local blast on the grape non-redundant EST database enabled us to identify 68 unigenes with a perfect match. Among these was a cluster of transcription factors belonging to the VvMybA family including four isogenes, of which VvMybA1 and VvMybA2 had already been described as being involved in the anthocyanin biosynthesis pathway (Kobayashi et al 2004; Walker et al. 2007). The four VvMybA genes were retained as candidate genes (Table S1 in supplementary material). The sequencing of a 6.2 Kbp region in 141 genotypes (Figure 2 and supplementary Table S2) enabled the identification of a total of 75 SNPs, two InDels, and one retrotransposon, Gret1: for VvMybA1, 20 polymorphisms on 1.5 kbp including Gret1 in the promoter; for VvMybA2, 28 polymorphisms on 2.2 kbp including an InDel in the ORF; for VvMybA3, 21 polymorphisms on 1.75 kbp including an InDel in the ORF; for VvMybA4, nine on 0.75 kbp.

Association genetics

Core collection sample (CC) showed greater variation than the SxG sample ($\sigma_{CC}=1.15$), as was expected, with anthocyanin concentration ranging from 0 mg of anthocyanin to a maximal content of 33.2 mg of anthocyanin per g of fresh skin. The association tests were performed following the Structured Association (SA, Thornsberry et al. 2001) and Mixed Linear Model (MLM, Yu and Pressoir et al. 2006) procedures to control for false positives, using either the kinship calculated on SSRs or on SNPs in independent genes. The GLM test of the population structure effect was significant (P-val=0.0019) according to the hypothesis that the sample was divided into two admixed subpopulations. Thus, the population structure effect was included in all the association tests. Considering that we performed 78 tests, we applied a Bonferroni correction for
multiple hypotheses testing, whereby the threshold of the tests was set to 6.25E-4. At this threshold, SA and MLM with kinship calculated on SSRs provided the same number of significant associations, while MLM with SNP kinship appeared slightly less conservative (Supplementary Figure 1), detecting one additional significant association. We thus relied on the SA method.

Out of the 78 tests performed on a BLUP predicting anthocyanin berry content across two years, 32 gave a significant result, with P-values ranging from 3.83E-04 to 5.72E-25 (Table 2), which explains 10 to 59% of the (anthocyanin) variation. Out of these 32 associated polymorphisms, 10 were identified on \(VvMybA1\), 12 on \(VvMybA2\), 10 on \(VvMybA3\), but none on \(VvMybA4\). The location of the candidate polymorphisms on each gene, the effect on protein sequences, and the test statistics for association are presented in Table 2.

In order to determine whether the positive association was due to LD throughout the zone or a particular effect of each polymorphism, we genotyped the core collection with polymorphic SSR loci from scaffold 97. Two SSRs at both ends of scaffold 97 were already available (\(VMC6B11\) and \(VVIU20\)), and six more were developed within scaffold 97 (\(VVTM1\) to 6), scattered between the \(VvMybA\) genes (Figure 2 and supplementary Table S1). We performed the same GLM tests as in the association genetics section, including eight SSR markers surrounding the three loci of interest. Each of the neutral markers located between each \(VvMybA\) isogene showed a high level of association. Only one test performed on SSR \(VVTM4\) was non-significant, revealing the presence of LD along the supercontig (Table 2 and Figure 2c). Nonetheless, 21 of the 32 polymorphisms linked to the phenotype had a smaller P-value than the most closely associated flanking SSR marker (\(VVTM2\), Table 2). Ten of the thirty-two
associated polymorphisms and five out of six polymorphisms selected using the stepwise cofactor selection method showed an adjusted $R^2$ higher than that of $VVNTM2$, the most closely associated SSR. With $R^2$ and P-values of the SSRs bordering each side of the supercontig that decrease with increasing distance from the $VvMybA$ genes, we successfully restricted the associated interval to the $VvMybA$ gene cluster only.

Linkage disequilibrium and haplotype structure among associated polymorphisms

On the overall associated polymorphisms, intragenic LD was observed to be higher than intergenic LD in this region (Figure 3), resulting in significantly higher intragenic $r^2$ values than on the overall SNP markers. The overall level of LD differed among genes (ANOVA P-value<0.0001): tight LD on $VvMybA2$, moderate on $VvMybA3$, and moderate to low on $VvMybA1$ ($r^2_{A2}$=0.38, $r^2_{A3}$=0.28, and $r^2_{A1}$=0.22 respectively). $VvMybA2$ and $VvMybA3$ appeared to be in strong intergenic LD ($r^2_{A2A3}$=0.28) and significantly different from other intergenic patterns of LD ($VvMybA1$ vs. $VvMybA2$ and $VvMybA1$ vs. $VvMybA3$; P-value<0.0001).

Phased haplotypes of each gene were then reconstructed using associated polymorphisms. On all three genes, the reconstruction enabled us to identify 96 haplotypes with a single haplotype representing a frequency of 45%. We identified 26 haplotypes on $VvMybA1$, 24 on $VvMybA2$, and 42 on $VvMybA3$ separately on each gene.

In the haplotype group carrying the A1 Gret1 insertion, only 2 polymorphisms segregated with a frequency higher than 5% (Y847 on $VvMybA1$ and R1020 on $VvMybA2$), while their frequencies ranged from 11 to 50% in the haplotype group carrying no A1 Gret1. This reveals that most of the associated polymorphisms were only variable in the absence of A1 Gret1; hence, their effect may be distinguished from
that of A1 Gret1. Although unbalanced haplotype classes were involved, the tests performed on single-gene haplotypes led to reinforced association and LD levels within and among genes (Table 2). Due to weak statistical power we did not carry out the tests on the haplotype set combining all 32 polymorphisms.

Stepwise Cofactor selection for QTN identification

In order to define a set of highly associated and independent polymorphisms within a strong LD context, we applied a stepwise cofactor selection method for the associated polymorphisms separately on each gene. A set of six putative QTNs was retained (Figure 2c). On VvMybA1, three polymorphisms were retained: the retrotransposon A1 Gret1 (promoter), SNP A1 K244 (change from R to S in position 188 of the amino acid sequence) and A1 M318 (change from Q to P in position 213); on VvMybA2, 2 polymorphisms were integrated, SNP A2 K980 (change from R to L in position 44) and A2 InDel1748 (frame shift leading to truncated 265 amino acid protein instead of 344); on VvMybA3, only one, SNP A3 K403 (promoter); none on VvMybA4.

The PHASE reconstruction with the 6 retained QTNs led to the identification of 17 distinct macrohaplotypes combining the linkage information of the 3 genes for the 141 individuals. A single macrohaplotype represented an overall frequency of 0.58 and nine haplotypes were present in less than 4 individuals (frequency of 0.01). The macrohaplotype was highly associated with anthocyanin berry content in a structured association model with a P-value of 1.22E-30 and a marker R² of 0.82. Among the six putative QTNs, there were only two pairs of loci showing substantial LD: A1 Gret1 and A2 InDel1748 that were moderately linked with r²=0.25 and D’=0.73, and A2 K980 and A3 K403 that were more closely linked with r²=0.51 and D’=0.90. All other pairs
showed low LD, considering their physical linkage with an $r^2$ ranging from 0.17 to 0.25 and a $D^\prime$ ranging from 0.52 and 0.61.

Multivariate regression model and haplotype association

To further dissect the combinatorial effect of each of the six selected polymorphisms in terms of additive and dominance effects at the genotypic level (Table 4), we applied a regressive procedure for multivariate model testing. Both Akaike and Bayesian information criteria led to the selection of a 7-cofactor model explaining 84% of variation in anthocyanin content (including population structure, additivity and dominance of $A1 \text{Gret}1$, additivity of $A1 \text{K244}$, dominance of $A2 \text{K980}$, additivity of $A2 \text{InDel1748}$, and additivity of $K403$). We showed that $A1 \text{Gret}1$ was the polymorphism that explained most of the variation (the partial additive effect of Gret1 in an SA model accounted for 59% of the variation in anthocyanin content of the berry) and was the only one included in the final model with its two genetic components. Including four other polymorphisms enabled us to account for an additional 23% of the variance. SNP $A1 \text{M318}$, which was initially selected as a putative QTN, appeared to be useless in the full rank model, limiting the final set of putative QTNs to five.

Discussion

In the previous section, we presented the results of a detailed study from QTL to QTNs for a quantitative trait in grape. These results elucidate the primordial role played by the cluster of the $\text{MybA}$ genes in the quantitative determinism of grape berry color, since only five polymorphisms in the three genes are enough to explain 84% of the total
phenotypic variation of the core collection with reasonable statistical power (df_{model}=10 and df_{error}=67). Whilst great emphasis was previously placed on the role played by allelic variants of the \textit{VvMybA1} promoter in explaining the qualitative distinction between white and black grapes (Kobayashi \textit{et al.} 2002; Lijavetsky \textit{et al.} 2006; This \textit{et al.} 2007), here, we have instead developed an accurate model for a quantitative understanding of variation in grape color.

\textit{Increasing the accuracy of the analysis of grape color: molecular vs. empirical classification}

One way of increasing the accuracy of the analysis is to improve the estimation of the genotypic value. In this work, the data repetition across years and blocks was valorized through the definition of BLUP calculated from mixed models already used for the extraction of genetic variance components (Borevitz \textit{et al.} 2002). Consequently, the overall year effect for two years was approximately taken into account, and was shown to be helpful in explaining the phenotype. Other environmental effects such as saccharose concentration or light exposure proved to have an influence on \textit{Myb} genes related to grape anthocyanin biosynthesis (Matus \textit{et al.} 2009). Full control of environmental effects can be achieved by repeated phenotyping across environments, years, and their covariance, as successfully done in apple by Segura \textit{et al.} (2008).

\textit{A complex trait with a simple genetic architecture}

Previous work in quantitative genetics emphasizes that the \textit{Gret1} homozygous presence is the determinant for white color (This \textit{et al.}, 2007); our study confirms this finding.
because a single QTL was found and no other polymorphism was able to explain the white-to-colored grape bimodal variation.

Genetic variation for anthocyanin berry content was convincingly (LOD score up to 29.78, and $R^2$ up to 0.62) and consistently (detected across years and blocks) explained by a single QTL. Nonetheless, the anthocyanin content of the berry show high level of heritability (Barritt and Einset 1969) and a continuous variation; thus, the anthocyanin content of the berry is necessarily polygenically determined. Knowledge of the grape genome sequence enabled the reduction of an interval of 5.9 cM into a set of 68 genes.

Given previous evidence (Kobayashi et al. 2005; Lijavetsky et al. 2006; This et al. 2007; Walker et al. 2007), we reduced a priori this number to a single gene cluster. One of the main limitations of this study is the fact that we did not sequence all remaining 65 unigenes present in the QTL confidence interval and thus, there may be other loci that are also involved in anthocyanin biosynthesis in the same region; clearly, we have made a candidate gene assumption. Furthermore, the testing of new putative variation factors as candidate genes has recently been suggested (Deluc et al. 2008). Nonetheless, both the high scores of the QTL and association tests and the overall goodness-of-fit of the final model are sufficient to hereby confirm a posteriori that most of the genetic effect has been included and unambiguously analyzed in different genetic backgrounds.

The LD pattern led us to consider a gene haplotype structure strongly associated with anthocyanin content. Nonetheless, with unbalanced frequency in the haplotypes set, the test results have to be interpreted with caution because exaggerated numbers of classes lead to over-parametered non-conservative tests. To avoid redundancy in the presence of LD and to identify the most informative sites, we used a regression technique with stepwise cofactor selection, which is more convenient in the case of a huge number of
linked markers. Starting with 32 associated polymorphisms, we reduced the dataset to five putative QTNs with non-redundant genetic effect. In future, an alternative method would be to adopt a composite association mapping strategy similar to the composite interval mapping in QTL detection but applied to natural populations (Boitard et al. 2006).

The very high percentage of explained variance leads us to consider that nearly all the variation is determined by the *VvMybA* gene cluster. Even if a huge number of genes underlying QTL are described as transcription factors (Salvi and Tuberosa 2005; Price 2006), only a few studies identify a transcription factor gene cluster as shaping a quantitative trait (Francia et al. 2007). An interesting feature is that each gene of the cluster appears to have a dose effect. Nearly all the SNPs except one (A2 K980) essentially display additive behaviour (Table 3), which is coherent with the fact that we were investigating a transcription factor family where such a dose effect is to be expected: more functional MYBA protein synthesis leads to more *VvUFGT* transcript abundance and thus to more anthocyanin synthesis. The retroelement of *VvMybA1*, *Gret1*, is the main component in generating color variation; nevertheless, the four other polymorphisms explaining an additional 23% of variance represent a very significant contribution to the phenotype (Table 3). For instance, in the case of the SxG population, Syrah and Grenache had the same genotype at the *Gret1* locus but differed in their A2 K980 genotype, Syrah being heterozygote while Grenache was homozygous for the weak allele. This explained both the lack of color observed in Grenache and the smaller size of the QTL compared to Syrah on the Grenache parental map.

These findings are also relevant in terms of breeding due to the close physical linkage between the *VvMybA* genes. Any new breeding strategy for anthocyanin potential in
grape should first ensure the presence of the appropriate combination of \textit{VvMybA} alleles
in the parents at the haplotypic level. Otherwise, the low recombination rate will be very
restrictive.

Validity of the detected associations

The starting point in dissecting a quantitative phenotype is to obtain accurate positional
information about the loci responsible for the variation. Both QTL studies and genome
scans are very straightforward approaches as they do not rely on \textit{a priori} selection of the
polymorphisms to be tested. The use of collections of natural diversity enables very
accurate resolution in quantitative genetics studies, leading to putatively causal
polymorphisms (Yu and Buckler 2006). Furthermore, the results are easily transposed
as the tests are performed in different genetic backgrounds, contrary to a cross-derived
population, where the results are cross-dependent (Flint-Garcia \textit{et al.} 2007).

A general concern in association genetics on natural samples is the control of false
positives due to genetic neutral covariance, generally referred to as “population
structure” (Pritchard \textit{et al.} 2000) and distinct levels of kinship. Methods of dealing with
such problems are widely available but depend on the technical possibilities of the crop
being studied (Thornsberry \textit{et al.} 2001; Yu \textit{et al.} 2006; Zhao \textit{et al.} 2007). Convenient
structure matrices to be used as cofactors in association studies rely on inferences about
the past history of the crop concerned. The hypothesis relating to two major populations
corresponding to Western and Eastern genetic pools was emphasized by Arroyo-Gracia
\textit{et al.} (2005) and Le Cunff \textit{et al.} (2008), and was integrated as such into our model.
Recent studies have shown that the relative relatedness between individuals (kinship)
performed better than overall population covariance (structure) in controlling false
positives (Zhao et al. 2007). In the case of grape, we tested both structured association and mixed models using either the 20 SSRs or a set of 129 SNPs located in 10 genes of the anthocyanin biosynthesis pathway to calculate kinship. All alternative models appeared in our case to perform equally well (Supplementary Figure 1), probably owing to the fact that either structure or kinship effects are weak compared to the very strong detected associations. Therefore, we simply relied on the structured association procedure, a purely deterministic model. Multiple testing is also challenging in association studies where the number of markers is often huge. We chose to apply a Bonferroni correction systematically and set an extreme threshold level for each test to ensure a very high level of stringency.

The use of a stepwise cofactor selection method and the definition of a multivariate model is the most precise way to interpret the genetic determinism of the \textit{in vivo} accumulation of anthocyanin. The stepwise selection method ensures that all polymorphisms selected within a gene are independent; the linkage information is therefore synthesized in the best possible way. The haplotypic treatment of the QTNs led to a higher level of association than by taking each polymorphism independently, but the R\textsuperscript{2} is 2\% lower than the selected multivariate model. These results must be interpreted with caution due to the non-homoscedasticity of the haplotype classes with the overrepresentation of a single haplotype and numerous low frequency ones. In terms of the information provided, the multiple regression method allowed the ranking of the various QTNs in order of association magnitude and guaranteed the independence of the various effects tested, so as to ultimately establish the best non-redundant model. This approach represents a step forward in the field of complex trait dissection, where the combined effects of different genes are believed to exist but have yet to be
demonstrated. In a system where all three genes may interact and have many different alleles, a statistical approach is very helpful in defining new hypotheses and in creating a hierarchy of the many polymorphisms that have been shown to shape the complex phenotype. Nonetheless, due to the extent of LD between \( VvMybA2 \) and \( VvMybA3 \) in the sample, the additive effect of A2 K980 cannot be differentiated from the effect of A3 K403. Replacing the additive effect of A3 K403 by A2 K980 decreased the \( R^2 \) by only 0.2%.

**Functional effect of the polymorphisms**

Polymorphisms on \( VvMybA2 \) and \( VvMybA3 \) were essentially located in the promoter and the first exon, while the \( VvMybA1 \) promoter and the first exon appeared to be less variable, showing more variability in the 3’UTR instead. On these three genes, the third exon was rather monomorphic, showing only three SNPs on MybA1 that were all associated, one associated indel on \( VvMybA2 \), and one unassociated indel on \( VvMybA3 \). This exon corresponds to a CR domain putatively involved in DNA binding activity. Previous work led by Walker *et al.* (2007) showed the capacity of both \( VvMybA1 \) and \( VvMybA2 \) to activate anthocyanin synthesis, and Kobayashi *et al.* (2002) showed that the CR domain in \( VvMybA3 \) was truncated, leading to a non-functional protein. One possible explanation for the involvement of non-functional \( VvMybA3 \) in berry pigmentation is the steric competition between \( VvMybA3 \) and the other two functional isogenes. This hypothesis is supported by the fact that on \( VvMybA3 \), only one polymorphism within the promoter is linked, while on \( VvMybA1 \) and \( VvMybA2 \), exonic polymorphisms also appear to be linked with anthocyanin content. Nonetheless, given the strong LD level between A2 K980 and A3 K403, we may also consider that the
association of VvMybA3 is purely due to linkage. For locus VvMybA4, reported as being unexpressed in berries (Walker et al. 2007), we were unable to detect any effect on grape color due to the limited variability of the gene.

Out of the six polymorphisms, three had already been identified in previous works as being associated with qualitative color variation (Kobayashi et al. 2005; This et al. 2007) because of gene silencing or because of a change in the structure of the white allele (Kobayashi et al. 2005; Walker et al. 2007). Nevertheless, for the first time, we have detected an association between A1 K244 and A1 M318 and the anthocyanin content of the berry, with both mutations leading to a change in the amino acid in the VvMybA1 protein structure: from R to S in position 188 and from Q to P in position 213 respectively. For the association characterized on VvMybA3 with A3 K403, only a statistics-based quantitative approach could enable the identification of putative effects on transcript regulation.

A final validation of all the effects described here could be performed through transient assays. In the case of a quantitative trait, due to very high technical variance in the phenotype of the transformed plants, this would require a huge number of replicates of independent transformations, which is not technically feasible at the present time. The present study directly investigated the polymorphism that may cause the variation at the phenotypical level. As the revealed associations are very strong and are supported by additional physiological information (data not shown here), we believe that association genetics per se is sufficient validation to draw final conclusions about the genetic determinants of naturally occurring variation. The color quality aspect, a highly relevant trait for viticulturists, may now be easily targeted in any grape cross offspring, thereby
providing the opportunity to identify very early genotypes of interest, from light red for rose wines to the most intense purple.

Acknowledgments

In the memory of Dr A. Bouquet and Dr D. Fournand. Thanks to Dr V. Segura for many fruitful and inspiring discussions and to two anonymous reviewers for helpful comments.

This work was funded in part by the European project, FLAVO (number 513960), a French Genoplante project, COREGRAPGEN (Trilat 017), and a PhD grant from the Languedoc-Rousillon region for AFL. The authors wish to thank the technical staff for the phenotypic analysis, in particular M. Farnos, P. Ortigosa, C. Morel, and the staff of Domaine du Chapitre and Domaine de Vassal for grape cultivation.
ADAM-BLONDON, A. F., C. ROUX, D. CLAUX, G. BUTTERLIN, D. MERDINOGLU et al., 2004
 Mapping 245 SSR markers on the Vitis vinifera genome: a tool for grape genetics.
 Theoretical and Applied Genetics 109: 1017-1027.

AGEORGES, A., L. FERNANDEZ, S. VIALLET, D. MERDINOGLU, N. TERRIER et al., 2006 Four specific isogenes of the anthocyanin metabolic pathway are systematically co-expressed with the red colour of grape berries. Plant Science 170: 372-383.


ALPERT, K. B., S. GRANDILLO and S. D. TANKSLEY, 1995 Fw-2.2 - a Major Qtl Controlling Fruit Weight Is Common to Both Red-Fruited and Green-Fruited Tomato Species.
 Theoretical and Applied Genetics 91: 994-1000.


BARRITT, B. H., EINSET, J., 1969 The inheritance of three major fruit colors in grapes. Journal of
the American Society for Horticultural Science 94: 87-89.

BASTEN, C.J., B. S. WEIR and Z. B. ZENG, 2001 QTL Cartographer, Department of Statistics,
North Carolina State University, Raleigh, N.C.

Linkage disequilibrium interval mapping of quantitative trait loci. Bmc Genomics 7.

BOREVITZ, J. O., J. N. MALOOF, J. LUTES, T. DABI, J. L. REDFERN et al., 2002 Quantitative Trait
Loci Controlling Light and Hormone Response in Two Accessions of Arabidopsis

BOSS, P. K., C. DAVIES and S. P. ROBINSON, 1996 Analysis of the expression of anthocyanin
pathway genes in developing Vitis vinifera L cv Shiraz grape berries and the

of Transgenic Crop Plants: Transgenic Temperate Fruits and Nuts. Eds C. Kole and T.

CHAGNE, D., C. M. CARLISLE, C. BLOND, R. K. VOLZ, C. J. WHITWORTH et al., 2007 Mapping a
candidate gene (MdMYB10) for red flesh and foliage colour in apple. Bmc Genomics
8.

CHANDLER, S., and Y. TANAKA, 2007 Genetic modification in floriculture. Critical Reviews in
Plant Sciences 26: 169-197.

CHOPRA, S., S. M. COCCIOLONE, S. BUSHMAN, V. SANGAR, M. D. McMULLEN et al., 2003 The
maize Unstable factor for orange1 is a dominant epigenetic modifier of a tissue


DOEBLEY, J., and A. STEC, 1993 Inheritance of the Morphological Differences between Maize and Teosinte - Comparison of Results for 2 F2 Populations. Genetics 134: 559-570.


PRICE, A. H., 2006 Believe it or not, QTLs are accurate! Trends in Plant Science 11: 213-216.


QUATTROCCHIO, F., W. VERWEIJ, A. KROON, C. SPELT, J. MOL et al., 2006 PH4 of petunia is an R2R3 MYB protein that activates vacuolar acidification through interactions with basic-helix-loop-helix transcription factors of the anthocyanin pathway. Plant Cell 18: 1274-1291.


Table 1: Properties of the QTL detected on linkage group 2 between markers VVMC5G7 and VVMybA1 in composite interval mapping for anthocyanin content of berry skin detected on the SxG map. Four detections were carried out separately on three different maps. BLUPs of the genotypic effect extracted from mixed models were used to synthesize the information in the 2 blocks and 2 years, and performed systematically better in QTL detection. \( R^2 = \) adjusted percentage of variance explained by the model, LOD= likelihood ratio of the Odds.

<table>
<thead>
<tr>
<th>Map</th>
<th>Year/Block</th>
<th>LOD</th>
<th>QTL confidence interval size</th>
<th>R squared in MQM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LOD-1 LOD-2</td>
<td></td>
</tr>
<tr>
<td>Consensus</td>
<td>2005/A</td>
<td>18.37</td>
<td>9.3 12.9</td>
<td>0.479</td>
</tr>
<tr>
<td></td>
<td>2006/A</td>
<td>18.37</td>
<td>7.6 11.5</td>
<td>0.486</td>
</tr>
<tr>
<td></td>
<td>2006/B</td>
<td>17.22</td>
<td>7.2  8.2</td>
<td>0.506</td>
</tr>
<tr>
<td></td>
<td>Blup</td>
<td>29.78</td>
<td>5.9 10.6</td>
<td>0.623</td>
</tr>
<tr>
<td>Syrah</td>
<td>2005/A</td>
<td>11.07</td>
<td>10.8 17.3</td>
<td>0.445</td>
</tr>
<tr>
<td></td>
<td>2006/A</td>
<td>16.22</td>
<td>12 21.8</td>
<td>0.560</td>
</tr>
<tr>
<td></td>
<td>2006/B</td>
<td>18.36</td>
<td>14.5 22.8</td>
<td>0.533</td>
</tr>
<tr>
<td></td>
<td>Blup</td>
<td>30.72</td>
<td>9.2 10.6</td>
<td>0.543</td>
</tr>
<tr>
<td>Grenache</td>
<td>2005/A</td>
<td>7.91</td>
<td>2.2  4.2</td>
<td>0.320</td>
</tr>
<tr>
<td></td>
<td>2006/A</td>
<td>10.86</td>
<td>1.8  3.6</td>
<td>0.390</td>
</tr>
<tr>
<td></td>
<td>2006/B</td>
<td>11.75</td>
<td>3.5  5.8</td>
<td>0.422</td>
</tr>
<tr>
<td></td>
<td>Blup</td>
<td>15.67</td>
<td>1.7  3.6</td>
<td>0.473</td>
</tr>
</tbody>
</table>
Table 2: Markers and results of association tests for 32 associated polymorphisms, 3 haplotypes, and 8 SSRs framing the candidate genes. P-values below 6.25E-4 were considered to be significant. $R^2_{mk}$ is the percentage of anthocyanin variation explained by the polymorphism alone when integrating the structure. Promo= Promoter, 3'UTR= 3' Untranslated sequence, Ex= Exon, Sy= Synonymous change, and NS= Non-synonymous change. Frequencies are given for the SNP minority allele and the number of alleles are given for SSRs and haplotypes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Marker</th>
<th>type</th>
<th>Promoter / Exon / Intron / 3'UTR</th>
<th>Sy / NS</th>
<th>frequency or num of alleles</th>
<th>F-test</th>
<th>P-value</th>
<th>$R^2_{mk}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMC6B11</td>
<td>SSR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
</tr>
<tr>
<td>VVNTM1</td>
<td>SSR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
</tr>
<tr>
<td>VVNTM2</td>
<td>SSR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
</tr>
<tr>
<td>S84</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.07</td>
<td>13.36</td>
<td>3.83E-04</td>
<td>0.10</td>
</tr>
<tr>
<td>R153</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.23</td>
<td>33.69</td>
<td>2.45E-12</td>
<td>0.35</td>
</tr>
<tr>
<td>W393</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.23</td>
<td>65.25</td>
<td>6.63E-20</td>
<td>0.50</td>
</tr>
<tr>
<td>W417</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.24</td>
<td>65.31</td>
<td>5.80E-20</td>
<td>0.50</td>
</tr>
<tr>
<td>R440</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.21</td>
<td>53.29</td>
<td>2.53E-17</td>
<td>0.45</td>
</tr>
<tr>
<td>S465</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.24</td>
<td>58.18</td>
<td>1.98E-18</td>
<td>0.47</td>
</tr>
<tr>
<td>Y754</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.12</td>
<td>16.69</td>
<td>4.54E-07</td>
<td>0.22</td>
</tr>
<tr>
<td>R762</td>
<td>SNP</td>
<td>Ex</td>
<td>NS</td>
<td></td>
<td>0.16</td>
<td>38.80</td>
<td>2.00E-13</td>
<td>0.40</td>
</tr>
<tr>
<td>K980</td>
<td>SNP</td>
<td>Ex</td>
<td>NS</td>
<td></td>
<td>0.27</td>
<td>87.08</td>
<td>3.19E-24</td>
<td>0.57</td>
</tr>
<tr>
<td>R984</td>
<td>SNP</td>
<td>Ex</td>
<td>NS</td>
<td></td>
<td>0.23</td>
<td>42.98</td>
<td>7.43E-15</td>
<td>0.40</td>
</tr>
<tr>
<td>R1020</td>
<td>SNP</td>
<td>Ex</td>
<td>Sy</td>
<td></td>
<td>0.23</td>
<td>39.31</td>
<td>7.00E-14</td>
<td>0.38</td>
</tr>
<tr>
<td>Indel1748</td>
<td>InDel</td>
<td>Ex</td>
<td>NS</td>
<td></td>
<td>0.21</td>
<td>65.01</td>
<td>2.31E-19</td>
<td>0.53</td>
</tr>
<tr>
<td>VVMycA2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VVNTM3</td>
<td>SSR</td>
<td></td>
<td></td>
<td></td>
<td>0.39</td>
<td>67.30</td>
<td>3.03E-18</td>
<td>0.56</td>
</tr>
<tr>
<td>Gret1</td>
<td>retro</td>
<td>Promo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K244</td>
<td>SNP</td>
<td>Ex</td>
<td>NS</td>
<td></td>
<td>0.25</td>
<td>51.00</td>
<td>1.79E-16</td>
<td>0.46</td>
</tr>
<tr>
<td>K259</td>
<td>SNP</td>
<td>Ex</td>
<td>NS</td>
<td></td>
<td>0.3</td>
<td>47.96</td>
<td>7.12E-16</td>
<td>0.44</td>
</tr>
<tr>
<td>M318</td>
<td>SNP</td>
<td>Ex</td>
<td>NS</td>
<td></td>
<td>0.19</td>
<td>45.31</td>
<td>2.87E-15</td>
<td>0.43</td>
</tr>
<tr>
<td>Y566</td>
<td>SNP</td>
<td>3'UTR</td>
<td></td>
<td></td>
<td>0.18</td>
<td>46.07</td>
<td>2.17E-15</td>
<td>0.44</td>
</tr>
<tr>
<td>R657</td>
<td>SNP</td>
<td>3'UTR</td>
<td></td>
<td></td>
<td>0.13</td>
<td>9.66</td>
<td>1.32E-04</td>
<td>0.14</td>
</tr>
<tr>
<td>Y620</td>
<td>SNP</td>
<td>3'UTR</td>
<td></td>
<td></td>
<td>0.04</td>
<td>13.69</td>
<td>3.31E-04</td>
<td>0.10</td>
</tr>
<tr>
<td>D660</td>
<td>SNP</td>
<td>3'UTR</td>
<td></td>
<td></td>
<td>0.04</td>
<td>6.74</td>
<td>7.42E-05</td>
<td>0.21</td>
</tr>
<tr>
<td>W811</td>
<td>SNP</td>
<td>3'UTR</td>
<td></td>
<td></td>
<td>0.13</td>
<td>9.90</td>
<td>1.10E-04</td>
<td>0.15</td>
</tr>
<tr>
<td>Y847</td>
<td>SNP</td>
<td>3'UTR</td>
<td></td>
<td></td>
<td>0.4</td>
<td>52.10</td>
<td>4.77E-16</td>
<td>0.52</td>
</tr>
<tr>
<td>VVMycA1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VVNTM4</td>
<td>SSR</td>
<td></td>
<td></td>
<td></td>
<td>0.39</td>
<td>33.29</td>
<td>2.47E-12</td>
<td>0.33</td>
</tr>
<tr>
<td>Y166</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M209</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.1</td>
<td>50.34</td>
<td>1.24E-16</td>
<td>0.44</td>
</tr>
<tr>
<td>M357</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.05</td>
<td>14.61</td>
<td>5.19E-11</td>
<td>0.38</td>
</tr>
<tr>
<td>W362</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.06</td>
<td>25.93</td>
<td>4.26E-10</td>
<td>0.29</td>
</tr>
<tr>
<td>R379</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.42</td>
<td>14.20</td>
<td>3.36E-06</td>
<td>0.20</td>
</tr>
<tr>
<td>K392</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.11</td>
<td>36.42</td>
<td>3.79E-13</td>
<td>0.36</td>
</tr>
<tr>
<td>K403</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.39</td>
<td>92.54</td>
<td>5.72E-25</td>
<td>0.58</td>
</tr>
<tr>
<td>R440</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.48</td>
<td>12.92</td>
<td>8.45E-06</td>
<td>0.17</td>
</tr>
<tr>
<td>R869</td>
<td>SNP</td>
<td>Promo</td>
<td></td>
<td></td>
<td>0.03</td>
<td>34.52</td>
<td>1.50E-12</td>
<td>0.35</td>
</tr>
</tbody>
</table>
Table 3: Structured Multivariate Regression Models. The models presented correspond to the best models for each cofactor number; after the *Gret1* model, all higher models are nested, allowing the classification of the identified QTNs from the least to the most important. The model in bold corresponds to the optimal model, i.e. minimizing AIC and BIC criteria. Nb Cof= Number of cofactors in the model, R²= percentage of the variance explained by the model, partial R²= percentage of the variance explained by individual polymorphism in the model, BIC= Bayesian (Schwarz) Information Criteria (the smaller the better), a is the additive component of the marker genetic effect, and d is the dominant component of the marker genetic effect.

<table>
<thead>
<tr>
<th>Nb Cof</th>
<th>R²</th>
<th>BIC</th>
<th>Cofactors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>61</td>
<td>-42.2</td>
<td><em>Gret1</em> a</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
<td>-78.8</td>
<td><em>K244</em> a, <em>In1748a</em></td>
</tr>
<tr>
<td>3</td>
<td>81</td>
<td>-91.1</td>
<td><em>K244</em> a, <em>K980</em> d, <em>In1748a</em></td>
</tr>
<tr>
<td>4</td>
<td>82</td>
<td>-94.9</td>
<td><em>K244</em> a, <em>K980</em> d, <em>In1748a</em>, <em>K403</em> a</td>
</tr>
<tr>
<td>5</td>
<td>83</td>
<td>-96.4</td>
<td><em>Gret1</em> d, <em>K244</em> a, <em>K980</em> d, <em>In1748a</em>, <em>K403</em> a</td>
</tr>
<tr>
<td>6</td>
<td>84</td>
<td>-97.4</td>
<td><em>Gret1</em> a, <em>Gret1</em> d, <em>K244</em> a, <em>K980</em> d, <em>In1748a</em>, <em>K403</em> a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>partial R²: 58.6, 3.6, 2.8, 1.2, 3.2, 11.3</td>
</tr>
<tr>
<td>7</td>
<td>84</td>
<td>-96.6</td>
<td><em>Gret1</em> a, <em>Gret1</em> d, <em>K244</em> a, <em>K980</em> d, <em>In1748a</em>, <em>K403</em> a</td>
</tr>
<tr>
<td>8</td>
<td>84</td>
<td>-95.5</td>
<td><em>Gret1</em> a, <em>Gret1</em> d, <em>K244</em> a, <em>M318</em> a, <em>K980</em> d, <em>In1748a</em>, <em>K403</em> a</td>
</tr>
</tbody>
</table>
Figure Captions

Figure 1: Distribution of the anthocyanin content in berry skin for the year 2005, expressed in Ln of mg of anthocyanin/g of fresh skin. The quantity of anthocyanin displays a continuous variation from 0 to 33.2 mg of anthocyanin per g of fresh skin, with an overrepresented sample of white berried cultivars displaying no anthocyanin.

Figure 2: (a) Presentation of VvMybA gene cluster and SSR markers; (b) position of the Quantitative Trait Nucleotides (QTNs); (c) level of association between markers and total anthocyanin content of berry skin along scaffold 97 of the grape genome browser. Along the X-axes, the dashed lines correspond to non-linear scales. In the association tests, the microsatellite markers are presented in red, the genic polymorphisms in blue; for the genic polymorphisms, dots correspond to QTNs and diamonds to the other polymorphisms. The Bonferonni threshold is equal to 6.25E-4.

Figure 3: LD plot based on R² values for the SNPs and InDels associated with the total anthocyanin content of berry skin on VvMybA1, A2, and A3 genes in the lower diagonal, and overall level of LD for the full genes in the upper diagonal. QTN selected through stepwise cofactor selection are framed in red. Presented R² values are estimated according to Remington et al. (2000)
Figure 1
Figure 2
Figure 3
<table>
<thead>
<tr>
<th>Amplicon</th>
<th>Forward</th>
<th>Reverse</th>
<th>Amplicon size</th>
<th>Number of polymorphic sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>PromoMybA1</td>
<td>TCCTCTCTCTAACCATGGCTAA</td>
<td>GAACCTTTTTTTGAAGTGTTGACT</td>
<td>900bp</td>
<td>0</td>
</tr>
<tr>
<td>PromoMybA1</td>
<td>GACGTAADAAATTGGTGCACGTG</td>
<td>GAACCTTTTTTTGAAGTGTTGACT</td>
<td>800bp</td>
<td>2</td>
</tr>
<tr>
<td>3'MybA1</td>
<td>TTCCAGGGAGGACTGCTAATGATG</td>
<td>TTGGCACCATAATTTTTCCATTTTC</td>
<td>1100bp</td>
<td>19</td>
</tr>
<tr>
<td>PromoMybA2</td>
<td>GTGAGGAGAGTACATTGTAGGA</td>
<td>GAACCTTTTTTTCAAGGTGTTGACC</td>
<td>1200bp</td>
<td>25</td>
</tr>
<tr>
<td>3'MybA2</td>
<td>TTCCAGGGAGGACTGCTAATGATG</td>
<td>GGTGAGGACTGCTAATGATG</td>
<td>950bp</td>
<td>3</td>
</tr>
<tr>
<td>Promo_MybA3</td>
<td>AACCCGTCATTGAATTGACAATAG</td>
<td>GAACCTTTTTTTCAAGGTGTTGACC</td>
<td>1400bp</td>
<td>19</td>
</tr>
<tr>
<td>3'MybA3</td>
<td>CTTGTATTGCGGGTAGGCTTC</td>
<td>GGGCCTCAAATGGAGAAGT</td>
<td>800bp</td>
<td>2</td>
</tr>
<tr>
<td>MybA4</td>
<td>TGTCGCAAGGAACATGGTAA</td>
<td>CATGCTTTGAGGATGGATGC</td>
<td>850bp</td>
<td>10</td>
</tr>
</tbody>
</table>

**Supplementary Table 1:** Description of the primers used for sequencing candidate genes
<table>
<thead>
<tr>
<th>SSR</th>
<th>Forward</th>
<th>Reverse</th>
<th>Amplicon size (min-max)</th>
<th>Type and Number of repeat on reference sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMC6B11</td>
<td>TGATTATGGCAATAATCACACC</td>
<td>TTGCTTACCCATCAAAAAGAAA</td>
<td>88-100bp</td>
<td>(CT)23</td>
</tr>
<tr>
<td>VVNTM1</td>
<td>CCACGCCACTATTGCTAAAC</td>
<td>TGCACCGTATCAAGATCATGTC</td>
<td>164-176bp</td>
<td>(AT)10</td>
</tr>
<tr>
<td>VVNTM2</td>
<td>TACCTGCTAACAATGCATTATG</td>
<td>TATTTGGTTTTTTCTAAATAGA</td>
<td>363-378bp</td>
<td>(AT)9</td>
</tr>
<tr>
<td>VVNTM3</td>
<td>TGCTGACCTGAATCATTITTCCTACTGATGTTCTGAGAGATGCTTTATC</td>
<td>254-299bp</td>
<td>(AT)9</td>
<td></td>
</tr>
<tr>
<td>VVNTM4</td>
<td>TTTGCATGACTGCTTTGTTGTTAT</td>
<td>CCCATTGCTAAACCTACTCCT</td>
<td>206-227bp</td>
<td>(CT)7</td>
</tr>
<tr>
<td>VVNTM5</td>
<td>AGGAGGAATCCACATCAAAAAGA</td>
<td>TGATTCAAGGAATAAATAACCATCA</td>
<td>273-298bp</td>
<td>(AT)11</td>
</tr>
<tr>
<td>VVNTM6</td>
<td>CTTTCTTGGACCCCATACAAA</td>
<td>TTCCCTATCAACAAAUCTTGAG</td>
<td>156-164bp</td>
<td>(AT)9</td>
</tr>
<tr>
<td>VVIU20</td>
<td>ACAACCTTAATGCTTCTACCAA</td>
<td>TCACCATGGAGATTTTCTGTAG</td>
<td>362-384bp</td>
<td>(TG)10</td>
</tr>
</tbody>
</table>

**Supplementary Table 2:** Description of the SSRs primers used to refine the association genetics study
Supplementary Figure Caption

Supplementary Figure 1: Cumulative distribution of P-values across different models. Cumulative Pareto(?) distribution of the test performed among 78 SNPs and total anthocyanin berry content across different models. SA is the simple structured association model, MLM SNP is the mixed linear model including the kinship calculated on 129 SNPs, and MLM µsat is the mixed linear model including the kinship calculated on 20 SSRs. At the threshold, the lowest(?) curve indicates the most conservative model.
Supplementary Figure 1