Genetics of berry colour and anthocyanin content variation in grapevine (*Vitis vinifera* L. subsp. *vinifera*)

Silvana Coelho Cardoso

Dissertation presented in fulfillment of the requirements for the Degree of Doctor of Philosophy in Biology (Molecular Genetics) at the Instituto de Tecnologia Química e Biológica da Universidade Nova de Lisboa

Oeiras, January 2011

Financial support from Fundação para a Ciência e a Tecnologia, grant number SFRH / BD / 29379 / 2006 and ERA-PG 074B GRASP GRAPE WINE.

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Acknowledgements

I am grateful to Fundação para a Ciência e a Tecnologia for financial support (grant number SFRH/BD/29379/2006 and ERA-PG 074B GRASP GRAPE WINE).

I would like to thank my supervisors Pedro Fevereiro and José Eduardo Eiras Dias for all their valuable support.

Many people contributed to this work in different ways. I would like to show my gratitude to them all.

Antero Martins and Elsa Gonçalves for data on grapevine clones and help on field work.

José Miguel Martínez-Zapater for receiving me in his lab and for the advice on my work. Diego Lijavetzky and Rita Francisco for help with the analysis of microarray data, and Gema Bravo and Virginia Rodríguez for the support on the lab.

Nikolas Maniatis for his endless support, for receiving me in his group and taking the grapevine challenge. Winston Lau for his willingness to help with everything, especially with data handling and programming.

Isabel Spranger, Conceição Leandro, Sun Baoshan for their help with anthocyanin extraction. Amélia Soares and Deolinda Mota for all the help in the lab. Jorge Cunha for his help in field collection and Ana Margarida Teixeira Santos for data on viral infections and help with ELISA tests. Flávia Moreira and Stella Grando for SSR genotyping.

I also want to thank Mara, Inês, Rita, Duarte, Hugo, Heather, Chris and Winston for their cheerfulness and to all my lab colleagues. My family and Diogo for their unconditional support.

Table of Contents

List of Tables

List of Figures

List of Appendices

Summary

Anthocyanin content of grape berry skin determines the colour of grapes and wine. This trait has been widely studied due to its importance for grape and wine marketing and also due to the antioxidant properties of anthocyanins.

In this thesis the variation of this trait was investigated within and between cultivars. DNA sequence variation and differential gene expression were studied among clones of the cultivars Aragonez and Negra Mole. Grape colour phenotyping was explored using different phenotyping approaches. Association mapping was performed for a sample of 149 cultivars and association mapping methodologies considering structure and relatedness in the sample were discussed.

It was observed that no DNA sequence variation was present in the studied genomic regions between different clones of the same cultivar. Differential expression between Aragonez clones with contrasting values of skin total anthocyanin concentration was found to be very subtle not showing any significant results after correction for multiple testing and with two fold-change. However, relaxing statistical stringency and focusing on functional groups of interest (flavonoid metabolism and transcription factors) a list of 24 genes of interest was identified. This included two genes involved in the flavonoid metabolism, coding enzymes related with the glucosylation of flavonoids and transcription factors of the following families: Myb, Myc, zinc fingers, WRKY, DOF, GRAS, homeobox domain, YABBY, basic-leucine zipper, pathogenesis-related and plant homeodomain finger.

Characterisation of cultivars using data on anthocyanins showed the use of concentration measures to be different from relative abundance measures. Principal Component Analysis showed both measures to

separate cultivars according to anthocyanidin type and acylation pattern. However, they differed in the separation of cultivars according to anthocyanidin type. While concentration data separated cultivars based on methylation level, relative abundance separated them according to hydroxylation. The analysis based on concentration showed total skin anthocyanins concentration to be the strongest discriminative variable. Regression analysis showed that virus infection and maturity status of berries did not have an influence on total skin anthocyanin concentration in this sample. Diversity characterisation performed using SSR data was more accurate than that based on anthocyanin data. Characterisation based on SSR data was not correlated with the one based in anthocyanin concentration and mildly correlated with relative abundance distances. Visual characterisation showed skin colour to be more strongly related to relative abundance and pulp colour to concentration of anthocyanins.

From a methodological point of view, association analyses revealed relatedness measures based on the Proportion of Shared Alleles and on Ritland's kinship coefficient to be identical for association mapping purposes. The results from the complex model correcting for structure and relatedness and the simple model were also found to be similar for total skin anthocyanin concentration which raises questions about the advantages of using less parsimonious models and the risk of false negative results.

Association mapping results indicated three genes coding transcription factors (*MYB11*, MYC_B and *MYBCC*) to have a significant role in total skin anthocyanin concentration and specific anthocyanin content. *UFGT* and *MRP* genes involved in the pathway and transport, respectively, were also identified to be associated with specific anthocyanin types. Association tests taking into account statistical interactions between

SNPs in different genes revealed a high proportion of significant interactions between SNPs in three transcription factors coding genes (*MYB11*, MYC_B and *MYBCC*) and two genes of the biosynthetic pathway (*CHI* and *LDOX*). These interactions suggest biological interplay between these genes to regulate the biosynthesis of anthocyanins.

This thesis gives a strong contribution to the understanding of the colour phenotype in grape. It provides ground for further studies on gene expression, genetic association and functional assays.

Sumário

A videira cultivada é uma das espécies agrícolas de maior relevância económica. Pensa-se que as diferentes cultivares tenham resultado de vários processos incluindo múltiplos eventos de domesticação da parente selvagem (*Vitis vinifera* subsp. *sylvestris*), de cruzamentos entre esta e cultivares ou entre cultivares. A propagação vegetativa é a forma de propagação de cultivares uma vez que fixa as características desejadas. As plantas obtidas vegetativamente são designadas "clones" e são normalmente idênticas geneticamente. Contudo, diferenças fenotípicas comummente observáveis entre estas plantas. A presença de mutações somáticas é um das explicações mais frequentes para estas diferenças. Estas mutações ocorrem espontaneamente e acumulam-se ao longo de inúmeros ciclos de propagação vegetativa, sobretudo em variedades antigas. A caracterização de cultivares e de clones é extremamente importante nesta espécie já que a selecção e melhoramento exploram ambos os níveis de variabilidade. Esta caracterização tem sido feita usando dados fenotípicos e marcadores moleculares.

A cor dos bagos é uma das características de maior importância na videira uma vez que afecta a qualidade das uvas e do vinho. Esta característica é determinada pelo conteúdo de antocianinas na película e menos frequentemente na polpa dos bagos. As propriedades antioxidantes das antocianinas e consequentes benefícios para a saúde humana têm também contribuído para o interesse nestes compostos. A via metabólica de síntese das antocianinas está bastante estudada mas o seu controlo genético ainda não foi totalmente esclarecido. Vários autores demonstraram a influência de genes pertencentes a duas famílias de factores de transcrição (Myb e Myc) na regulação de genes da via metabólica das antocianinas.

A cor do bago descrita como presença *versus* ausência de pigmentação tem segregação Mendeliana e foi mapeada no cromossoma 2. A cor como uma característica quantitativa foi também mapeada no mesmo cromossoma, contudo a variação fenotípica observada entre cultivares pigmentados não foi ainda completamente explicada.

O mapeamento por associação é um método com grandes potencialidades na identificação dos factores genéticos que determinam o fenótipo. Este método tem a vantagem de permitir obter grande poder e resolução quando comparado com o mapeamento de ligação. Contudo esta abordagem ao mapeamento tem recebido críticas sobretudo devido à ocorrência de falsos positivos. Estes problemas têm sido habitualmente atribuídos à presenca de estrutura e parentesco na amostra utilizada. Inúmeros métodos foram propostos para colmatar este problema. Contudo, nao é ainda clara qual a melhor forma de avaliar a presença e a extensão da estrutura e do parentesco na amostra. Além disto, o rigor das correcções efectuadas e a proporção de falsos negativos resultantes é ainda desconhecida.

Com este trabalho procurou-se contribuir para uma melhor compreensão dos factores genéticos envolvidos na determinação da cor dos bagos de videira. Para isso investigou-se a diversidade entre clones ao nível da sequência de DNA e da expressão genética. Caracterizaram-se fenotipicamente diferentes cultivares e testou-se a presença de associações entre variantes genéticas em genes candidatos e vários fenótipos relacionados com a cor.

A variabilidade ao nível da sequência de DNA foi investigada entre clones das cultivares Aragonez e Negra Mole. A sequenciação de genes envolvidos na via metabólica das antocianinas, revelou a inexistência de polimorfismos entre clones do mesmo cultivar. Diferenças na expressão

genética foram avaliadas através da análise de *microarrays*. Para esta análise foram utilizados clones de Aragonez com concentrações contrastantes de antocianinas totais na película. Os resultados desta comparação mostraram que não existem diferenças evidentes entre os dois grupos estudados. Contudo, focando a atenção apenas em grupos funcionais considerados *a priori* relevantes para a cor (factores de transcrição e metabolismo de flavonóides) e em genes que mantêm os resultados para mais do que um teste estatístico, foi possível identificar um grupo de 24 genes de potencial interesse ($P < 0.05$, não corrigido para testes múltiplos). Este conjunto de genes incluíu dois genes que codificam enzimas relacionadas com a glucosilação de flavonóides e genes das famílias de factores de transcrição Myb, Myc, z*inc-fingers*, WRKY, DOF, GRAS, *homeobox*, YABBY, *basic-leucine zipper*, *homeodomain-fingers* e relacionados com a patogénese.

Por forma a compreender melhor o fenótipo da cor, caracterizou-se uma amostra de 149 cultivares usando diferentes abordagens. Utilizaram-se classificações visuais da cor, bem como a concentração de antocianinas totais e de tipos específicos de antocianinas, e a sua abundância relativa. A concentração e abundância relativa foram obtidas por RT-HPLC. Observou-se que apesar de uma certa sobreposição da informação fornecida pela concentração e a abundância relativa de antocianinas, estas não são completamente idênticas. A análise de componentes principais revelou que ambas a concentração e a abundância relativa permitem a separação de cultivares de acordo com o tipo de antocianidina e o tipo de acilação. Contudo, estas duas classificações separam de forma diferente os tipos de antocianidinas, sendo que os dados de concentração efectuaram uma separação baseada no grau de metilação enquanto que os dados de abundância relativa

distribuíram as cultivares de acordo com o grau de hidroxilação. A análise baseada na concentração indicou que as antocianinas totais têm o maior poder discriminativo. A caracterização da amostra baseada nos dados fenotípicos foi comparada com a caracterização feita com base em dados de 20 microssatélites (SSRs). Observou-se que a classificação baseada na concentração de antocianinas não se correlaciona com a baseada em SSRs. A abundância relativa mostrou uma correlação baixa mas significativa, apoiando a ideia de que os marcadores moleculares proporcionam maior rigor quando comparados com dados fenotípicos. Foi também analisado por regressão linear o efeito de infecções virais e do estado de maturação dos bagos na concentração total de antocianinas na película. Concluíu-se que nesta amostra estas variáveis não têm qualquer efeito neste fenótipo. No que diz respeito às caracterizações visuais da cor dos bagos mostrou-se que a cor da polpa reflecte sobretudo diferenças na concentração de antocianinas enquanto que a cor da película está sobretudo relacionada com diferenças ao nível da abundâncias relativa de tipos específicos de antocianinas.

A associação estatística entre 124 SNPs e os vários fenótipos relacionados com a cor foi testada numa amostra de 149 cultivares provenientes da colecção ampelográfica nacional portuguesa (Dois Portos, Portugal). Estes SNPs encontravam-se distribuídos ao longo de 15 genes candidatos. Quatro destes genes foram seleccionados com base nos resultados de expressão diferencial obtidos para os clones de Aragonez. Vários modelos estatísticos foram comparados para esta amostra. Do ponto de vista metodológico esta análise mostrou que a utilização das estimativas de parentesco baseadas na proporção de alelos partilhados e no coeficiente de Ritland (1996) são idênticas para os testes de associação genética. Foram também observados resultados semelhantes

entre o modelo que incluía a estructura e o parentesco e o modelo simples que não considerava estes factores, levantando questões acerca das vantagens da utilização de modelos menos parsimoniosos.

O estudo de associação indicou a associação entre três genes que codificam factores de transcrição (*MYB11*, MYC_B e *MYBCC*) e a concentração de antocianinas totais na película. Outros dois genes, um envolvido na via metabólica e outro no transporte de antocianinas para o vacúolo, mostraram associações com tipos específicos de antocianinas na película. A cor da polpa e da película e da polpa em conjunto mostraram associações com um elevado número de genes incluindo genes envolvidos na via metabólica, no transporte e que codificam factores de transcrição. Foram também identificadas interacções estatísticas entre uma grande proporção de SNPs em três genes que codificam factores de transcrição (*MYB11*, MYC_B e *MYBCC*) e entre dois destes e dois genes da via metabólica (*CHI* e *LDOX*).

Este trabalho constitui um importante contributo para a compreensão do fenótipo da cor dos bagos. As conclusões obtidas serão úteis para futuros trabalhos de expressão e associação genética e estudos funcionais.

CHAPTER I

GENERAL INTRODUCTION

General Introduction

In the present chapter the literature and accumulated knowledge on *Vitis vinifera* L., anthocyanins and association genetics are reviewed and the objectives of this thesis are stated.

1. *Vitis vinifera* L.

Vitis vinifera L. is a widely cultivated fruit crop with a harvested area above seven million hectares and more than 67 million tons of grapes produced per year (FAO, 2008).

Cultivated grapevine (*Vitis vinifera* subsp. *vinifera* L.) is the only member of the *Vitis* genus indigenous to Eurasia. It is suggested to have first appeared ~65 million years B.P. while domestication is thought to have occurred around 6000 to 9000 B.P. (de Saporta, 1879; Châtaignier, 1995; McGovern *et al.*, 1996; McGovern and Rudolph, 1996; Zohary, 1996; Zohary and Hopf, 2000). The number and geographic locations of domestication events are controversial. The restricted origin hypothesis states that one domestication event occurred in a single location with a limited number of founders and cultivars were later spread into other regions (Olmo, 1976). On the contrary, according to the multiple origin hypothesis, domestication consisted of a series of events through an extended time period along the whole area of distribution of the wild ancestor (*Vitis vinifera* subsp. *sylvestris* (C.C. Gmel., Hegi) (Arroyo-Garcia *et al.*, 2006; Grassi *et al.*, 2003; Mullins *et al.*, 1992).

Archaeological evidence in the Zagros mountains in the Near East supports the restricted origin hypothesis according to which domestication would have taken place in the Transcaucasus between the Black Sea and Iran. From here cultivars would have propagated south and westwards across all the Mediterranean basin, reaching the Iberian

Peninsula around 2800 BP (McGovern, 2003; McGovern and Rudolph, 1996; Zohary and Hopf, 2000).

On the other hand, the multiple origin hypothesis is supported by the existence of morphological differentiation among cultivars from different regions in the Near East and in West Mediterranean areas (Levadoux, 1956; Mullins, *et al.* 1992; Negrul, 1938). Recently, new data on genetic relationships between wild and cultivated grapevine have also suggested the existence of at least two origins, revealing an important contribution of *Vitis vinifera* subsp. *sylvestris* from both the Near East and Eastern Europe to modern cultivars germplasm (Arroyo-Garcia, 2006; Cunha *et al.*, 2009, 2010; Levadoux, 1956; Mullins *et al* . 1992; Negrul 1938).

Wild grapes are dioecious plants bearing black skinned berries with unpigmented flesh. They are forest climbers growing currently only along riverbank forests in dispersed populations across central and southern Europe, northern Africa, Middle East and southwestern Asia (Arnold *et al.*, 1998; Levadoux, 1956; Ocete *et al.*, 1999). Before the 19th century the wild vine had a broader habitat which was reduced due to the introduction of disease and pests from America (downy mildew, powdery mildew and *Phylloxera vastatrix*).

The domestication process involved several changes on morphological and biochemical traits, such as a shift from dioecious to hermaphroditic reproduction, increased uniformity of berry maturity within clusters, higher sugar content and a wider range of fruit colours (Levadoux, 1956; Olmo, 1995; Zohary and Spiegel-Roy, 1975). Traditionally viticulture was based on thousands of cultivars with very diverse characteristics (Einset & Pratt, 1975; Olmo, 1976). Alleweldt and Detweiler (1994) estimated the number of genotypes cultivated and in germplasm collections to be around 10 000. Nevertheless, modern viticulture has

reduced the number of cultivated genotypes, by focusing on few cultivars. This *et al.* (2007) argued that 5000 cultivars including many closely related may be a more realistic estimate of the number of grapevine genotypes currently available.

Although most modern cultivars are hermaphroditic and self-fertile, outbreeding through insect or wind pollination happens most often. Classical breeding of new cultivars is a very time consuming process, especially for species with a long life cycle such as grapevine. For wine cultivars, this process is additionally long due to winemaking and evaluation stages. Despite this, wine and table grape breeding has been effective in selecting cultivars to meet the demands on quality and resistance to pest and disease. Vegetative propagation has been also very important since it fixes desired phenotypes and avoids trait segregation (Zohary, 2004).

The different grapevine cultivars are thought to have resulted from several processes including multiple domestication events of *Vitis vinifera* subsp. *sylvestris* (Arroyo-Garcia *et al.*, 2006), crosses between wild plants and domesticated varieties, spontaneous crosses between cultivated varieties and controlled breeding programs (Pelsy, 2010). Due to the use of vegetative propagation of cultivars, the clone entity may be said to represent the simplest taxonomic unit for *Vitis vinifera* L. (Bisson, 1995). Clones are grouped in different cultivars when there are enough phenotypic differences between them to be grown for the production of different wines (Boursiquot and This, 1999).

Although clones of the same cultivar are usually identical, different phenotypes may be sometimes identified. These differences have been explained by phytopathological agents (Walter and Martelli, 1998), epigenetic modifications in response to environmental factors (Kaeppler

et al., 2000; Schellenbaum *et al.*, 2008) and somatic mutations (Hartman *et al.*, 1997). These mutations occur spontaneously and accumulate over many cycles of vegetative propagation. Its accumulation leads to phenotypic differences and the identification of different clones, especially among ancient varieties.

Grape apical meristems, as most dicotyledons, have a stratified arrangement of cells as a consequence of anticlinal divisions of the "tunica" (two outer cell layers, L1 and L2). The "corpus" cell layer is constituted by the cells underlying the tunica which divide in different planes (L3) (Schmidt *et al.*, 1924). Each layer gives rise to different plant tissues (Neilson-Jones, 1969). When a mutation occurs in a cell of one of the meristematic layers, it propagates by mitosis producing a mutated sector (D'Amato, 1977). Accordingly, chimeras may be classified in the following three types: mericlinal, sectorial and periclinal. Mericlinal chimeras comprise a mutation in just a part of one tissue layer. Sectorial chimeras contain a mutation in a section of several layers. Periclinal chimeras have a mutation in one or more entire layers (Dermen, 1960). Mericlinal and sectorial chimeras generate sectored organs. Periclinal chimeras are the most stable, maintained by vegetative propagation and depending on the mutated layer will derive different mutant organs (Franks *et al.*, 2002). A somatic mutation may also be sexually transmitted in case it occurs in a cell layer giving rise to gametophytic tissues (Neilson-Jones, 1969).

The occurrence of somatic mutations has influenced some major characteristics of grapevine such as flavour, seedlessness, colour, ripeness, size, compactness, canopy growth and productivity (Pelsy, 2010). Pinot cultivars phenotypic variation is an example of phenotypic diversity arising from clonal diversity (Franks *et al.*, 2002; Furiya *et al.*,

2009; Hocquigny *et al*, 2004; Walker *et al.*, 2006; Yakushiji *et al.*, 2006). Mutations causing phenotypic variation between clones have been identified in other grapevine cultivars and cultivar groups, such as Cabernet Sauvignon (Boss *et al.*, 1996; Walker *et al.*, 2006), Chardonnay (Duchene *et al.*, 2009; This *et al.*, 2007), Italia (Azuma *et al.*, 2009; Collet *et al.*, 2005; Kobayashi *et al.*, 2004), Muscat de Alexandria (Kobayashi *et al.*, 2004), Savagnis (Duchene *et al.*, 2009) and Ugni blanc (Fernandez *et al.*, 2006).

Due to the wide use of vegetative propagation, clonal selection has become the most important means to improve quality of grape cultivars. Therefore, the development of accurate methods for clonal characterisation is very important (Moreno *et al.*, 1998). Traditionally, this has been performed using ampelography and ampelometry. However, phenotypic differences due to the influence of environmental factors may lead to spurious identifications (Imazio *et al.*, 2002). As a consequence, much effort has been given to the development of methods for clonal discrimination based on molecular markers.

AFLP markers have been successfully used for clonal discrimination (Cervera *et al.*, 1998; Scott *et al.*, 2000; Sensi *et al.*, 1996; Vignani *et al.*, 2002). However, contrasting results have been obtained with SSR markers. Some authors were able to distinguish clones using these markers (Kozjak *et al.*, 2003; Moncada *et al.*, 2006; Regner *et al.*, 2000); while in other studies no variation was found between clones (Baneh *et al.*, 2009; Faria *et al.*, 2004; Imazio *et al.*, 2002; Loureiro *et al.*, 1998). The use of ISSR and RAPD markers has also been attempted unsuccessfully (Loureiro *et al.*, 1998; Moreno *et al.*, 1997). On the other hand, Faria *et al.* (2004) and Carcamo *et al.* (2010) have succeeded on clonal discrimination using respectively polymorphisms on the stilbene

synthase (*StSy*)–chalcone synthase (*CHS*) 5'untranslated genomic regions (*StSy–CHS* markers) and retrotransposon based markers.

Genotypes with three and four microsatellite alleles have been identified indicating chimerism in some grapevine cultivars, such as Primitivo (Franks *et al.*, 2002), Greco di Tufo and Corvina Veronese (Crespan, 2004), Cabernet Sauvignon (Moncada *et al.*, 2006), Pinot (Hocquigny *et al.*, 2004), Cabernet franc, Chenin, Grolleau, Riesling, Savagnin (Pelsy *et al.*, 2010), Chardonnay (Bertsch *et al.*, 2005) and Ugni blanc (Fernandez *et al.*, 2006). Recently, also Carcamo *et al.* (2010) identified a chimerical state in Tempranillo, synonym of Aragonez in Portugal (OIV, 2009). Nevertheless, only one clone showed this genotype, leaving the phenotypic variation among the remaining 27 clones unexplained.

1.1. Grapevine genetics and genomics

Vitis vinifera L. is a diploid species with nineteen chromosomes. A number of genetic linkage maps have been published for *Vitis* and in particular for *Vitis vinifera* L. aiming at the detection of particular traits or to serve as reference maps (Adam-Blondon *et al.*, 2004; Dalbo *et al.*, 2000; Doligez *et al.*, 2002, 2006; Doucleff *et al.* 2004; Fischer *et al.*, 2004; Fournier-Level *et al.*, 2009; Grando *et al.* 2003; Lodhi *et al.*, 1995; Riaz, 2004; Troggio *et al.*, 2007). Older linkage maps were based mainly on RAPD and AFLP markers. Recently these became mostly based on microsatellites produced by the international *Vitis* Microsatellites Consortium and a linkage map based in SNPs, SSRs and AFLPs was published by Troggio *et al.* (2007).

As a consequence of having an obligatory out-crossing ancestor and a long history of vegetative propagation, cultivated grapevine is highly

heterozygous and carries many deleterious recessive mutations (Olmo, 1976). The average heterozygosity has been estimated to be near 0.8 for microsatellite loci and 0.65 for SNP haplotypes (Lijavetzky *et al.,* 2007; Salmaso *et al.,* 2004). Inbreeding depression is very severe, such that after the second or third generation of selfing the descendants become sterile.

The French-Italian Public Consortium released the reference genome sequence for grapevine by sequencing the near homozygous Pinot Noir line (PN40024) in 2007 (Jaillon *et al.,* 2007). This has made it possible to quickly generate genetic analysis tools such as molecular markers based on sequence comparisons between variants and the reference sequence. Analysis of the near 500Mb genome also suggested the contribution of three ancestral genomes to the grapevine haploid content (Jaillon *et al.,* 2007). In the same year, Velasco *et al.* (2007) sequenced the heterozygous Pinot Noir, providing valuable information on single nucleotide polymorphisms (SNPs) at the genome level.

The accumulation of information on gene annotation as a result of whole genome sequencing, integrated genetic maps and expressed sequence tags (ESTs) databases has enabled large scale studies of gene expression profiling (Da Silva *et al.,* 2005; Jaillon *et al.,* 2007; Moser *et al.,* 2005; Peng *et al.,* 2007; Velasco *et al.,* 2007; Vezzulli *et al.,* 2008). Using these tools, several mRNA expression-profiling studies have been undertaken in grapevine, focusing mainly on plant development (Terrier *et al.,* 2005; Waters *et al.,* 2005) and biotic and abiotic interactions (Cramer *et al.,* 2007; Deluc *et al.,* 2007; Espinoza *et al.,* 2007; Grimplet *et al.,* 2007; Pilati *et al.,* 2007; Tattersall *et al.,* 2007).

Studies on *Vitis vinifera* L. proteomics and metabolomics have emerged in the last years, mainly concerning berry metabolism (Giribaldi

et al., 2007; Sarry *et al.*, 2004) and abiotic stress (Castro *et al.*, 2005; Cramer *et al.*, 2007; Deluc *et al.*, 2007; Figueiredo *et al.*, 2008; Grimplet *et al.*; Jellouli *et al.*, 2008; 2009; Vincent *et al.*, 2007).

2. Anthocyanins

Anthocyanins are phenolic compounds, belonging to a particular group named flavonoids. The word *anthocyanin* was coined by Marquart (1835) to designate blue pigments of flowers and derives from the Greek words *anthos* and *kyanos* meaning *flower* and *blue*, respectively. However, it was later realised that these compounds accumulate also in other organs and confer other colours besides blue (Markakis, 1982). These water-soluble pigments synthesised by higher plants accumulate in vacuoles (Harborne and Harborne, 1998), mostly in flowers but also in fruits and in other organs such as leaves and stems (Brouillard, 1982; Delgado-Vargas and Paredes-Lopez 2003). Anthocyanins confer colour to the tissues where they accumulate, ranging from magenta and red to blue, violet and purple. These compounds play several important roles in biological functions. Due to their colouration properties one of these functions is to attract pollinators and seed dispersers. However, they are also involved in pollen-tube growth and play protective roles against bacterial agents, insect attack and UV exposure (Harborne and Harborne, 1998; Winkel-Shirley, 2001).

In grape and wine industry these compounds are of utmost importance as their accumulation in berry skin and most rarely berry flesh, is responsible for the different colours of grapes. They also confer colour to red wine and contribute to other organoleptic characteristics due to interactions with other phenolic compounds, proteins and polysaccharides (Mazza and Miniati, 1993; Ribéreau-Gayon, 1982).

The study of anthocyanins has received great attention due to their benefits to human health and potential applications in the food industry as natural food colourants (Giusti and Wrolstad, 2003). Anthocyanins have been shown to have antioxidant properties *in vitro* (Sun, 2009; Tedesco *et al.*, 2001; Tsuda *et al.*, 1994; Wang *et al.*, 1997) and *in vivo* (Ramirez-Tortosa *et al.*, 2001; Tsuda *et al.*, 2000). Several studies have shown an impact of anthocyanins consumption on cardiovascular disease prevention (Abuja *et al.*, 1998; Day *et al.*, 1997; Matsumoto *et al.* 2002; Tsuda *et al.*, 1996; Whitehead *et al.*, 1995) and anti-inflammatory activity (Rossi *et al.*, 2003; Seeram *et al.*, 2001; Wang *et al.*, 1999). Evidence of anticarcinogenic activity has been strongly based on *in vitro* evidence of the antiproliferative effect of anthocyanins on cancer cell lines (Jing *et al.*, 2008; Kamei *et al.*, 1995, 1998; Malik *et al.*, 2003; Yi et al, 2005; Zhang *et al.*, 2008; Zhao *et al.*, 2004). Consumption of anthocyanins has also been suggested to play a role on prevention of obesity, possibly by improving adipocyte function and preventing metabolic syndrome (Kwon *et al.*, 2007; Prior *et al.*, 2008; Tsuda *et al.*, 2003, 2005, 2008). Type 2 diabetes prevention has also been related to anthocyanins both by obesity control and by protecting β-cells from glucose-induced oxidative stress (Al-Awwadi *et al.* 2005; Sugimoto *et al.*, 2003). Eye vision improvement has been shown to be related with anthocyanins effects as well (Kramer, 2004).

Anthocyanins synthesis starts during veraison (onset of ripening) and accumulation accompanies berry ripening (Cholet and Darné, 2004; Fournand *et al.*, 2006; Pérez-Magarino and González-San José, 2004; Ryan and Revilla, 2003). However, it has also been shown that its concentration may decrease slightly before harvest (Ryan and Revilla, 2003) and with over-maturation (Fournand *et al.*, 2006). Its accumulation

in grape berries is influenced by environmental factors such as light exposure and temperature (Cortell and Kennedy, 2006; Downey *et al.*, 2006; Jeong *et al.*, 2004; Matus *et al.*, 2009), soil conditions (Gil and Yuste, 2004; Yokotsuka *et al.*, 1999), vine water status (Kennedy *et al.*, 2002; Ojeda *et al.*, 2002; Roby *et al.*, 2004) cultural practices (Esteban *et al.*, 2001; Orts *et al.*, 2005) and viral infections (Cabaleiro *et al.*, 1999; Goheen *et al.*, 1958; Lider *et al.*, 1975; Tomazic and Korosec-Koruza, 2003).

Anthocyanin accumulation also varies according to the cultivar (Cacho *et al.*, 1992; Pomar *et al.*, 2005; Ryan and Revilla, 2003). In fact, the qualitative anthocyanin composition of a cultivar has been shown to be a powerful tool for chemotaxonomical characterisation (Harborne and Harborne, 1998). Multivariate analysis of anthocyanins profile has been successfully used to distinguish cultivars (Carreño *et al.*, 1997; Ortega-Regules *et al.*, 2006). In *Vitis*, many works have been published on the comparison between species (Ribéreau-Gayon, 1959, 1964). Anthocyanidin diglucosides have been found frequently in American *Vitis* species but only traces may be identified in *Vitis vinifera* L. (Mazza, 1995; Ribéreau-Gayon, 1982; Stobiecki, 2000).

2.1. Chemical structure

As a group of flavonoids, anthocyanins are characterised by a C_6-C_3-C_6 carbon backbone, with linkage of the aromatic ring to the benzopyrano in position 2 (Harborne and Harborne, 1998). Anthocyanins are glycosides of polyhydroxy and polymethoxy derivatives of 2-phenylbenzopyrylium or flavylium salts (Eder, 2000).

It has been estimated that more than 635 different anthocyanins have been found in nature (Andersen and Jordheim, 2008). Individual

anthocyanins differ in the sugars, hydroxyl and methoxyl groups, and aliphatic or aromatic acids. Their aglycones called anthocyanidins are quite unstable and have been identified in nature in near 25 different types (Andersen and Jordheim, 2006). Only the following six of these forms, differing at the 3' and 5' positions of the B-ring, are common in higher plants: cyanidin (Cy), peonidin (Pn), pelargonidin (Pg), malvidin (Mv), delfinidin (Dp) and petunidin (Pt) (Eder, 2000; Kong *et al.*, 2003). In *Vitis vinifera* L. pelargonidin has not been observed.

The sugar residue of anthocyanidin glycosides is most often glucose, but rhamnose, xylose, galactose, arabinose, rutinose, sambubiose and other sugars may also occur. These glycosides are usually 3-monoglycosides and 3,5-diglycosides. In *Vitis vinifera* L., only 3-monoglucosides have been identified while 3,5-diglycosides are found in other *Vitis* species (Mazza, 1995; Ribéreau-Gayon, 1982; Stobiecki, 2000). Sugar residues may also be acylated with organic acids (Eder, 2000; Mazza and Miniati, 1993) such as derivatives of cinnamic acid (caffeic, *p*-coumaric, ferulic and sinapic acids) and aliphatic acids (acetic, malic, malonic, oxalic and succinic acids).

The number of methoxyl and hydroxyl groups influences the colour of anthocyanins. From red to blue shades, the number of methoxyl and hydroxyl groups increases as shown on Table I-1 (Delgado-Vargas and Paredes-Lopez, 2003; Heredia *et al.* 1998). In grapes, the colours range from magenta/red with cyanidin to purple/blue with delphinidin (Harborne and Harborne, 1967). However, the colour of anthocyanins is also influenced by other factors, such as glycosylation and acylation patterns of the molecule, pH of the solution, presence of co-pigments and cell shape (Grotewold, 1998; Stintzing *et al.*, 2002).

Table I-1 List of the most common anthocyanidins and the differences found on chemical structure, colour and maximum absorption.

Name	Substitution		Colour
	R1	R2	
Cy	OH	H	Magenta
Pn	OCH_3	H	Magenta
Pg	H	H	Red
Mv	OCH_3	OCH_3	Purple
Dp	OH	OH	Purple
Pt	OCH_3	OH	Purple

Vitis vinifera L. cultivars accumulate 3-monoglucosides and acetate, coumarate and caffeoate derivatives of delphinidin, cyanidin, peonidin, petunidin and malvidin (Mazza and Miniati, 1993).

The study of anthocyanin content was initially performed by paper chromatography and thin layer chromatography (Fong *et al.*, 1974; Hrazdina and Franzese, 1974; Koeppen and Basson, 1965). High performance liquid chromatography (HPLC) became later very popular for anthocyanin analysis due to its high sensitivity and strong ability to separate compounds (Goldy *et al.*, 1989; Morais *et al.*, 2002; Pomar *et al.*, 2005; Wulf and Nagel, 1978). Recently, mass spectrometry and nuclear magnetic resonance (NMR) became also important tools for identification of anthocyanins (Alcalde-Eon *et al.*, 2004; Bakker *et al.*, 1997; Mateus *et al.*, 2002; Revilla *et al.*, 1999).

2.2. Biosynthesis of anthocyanins

The biosynthetic pathway of anthocyanins is very well characterised as it has been widely studied in petunia, snapdragon and maize (Saito and Yamazaki, 2002). Figure I-1 shows a simplified schematic drawing of the reactions involved this pathway.

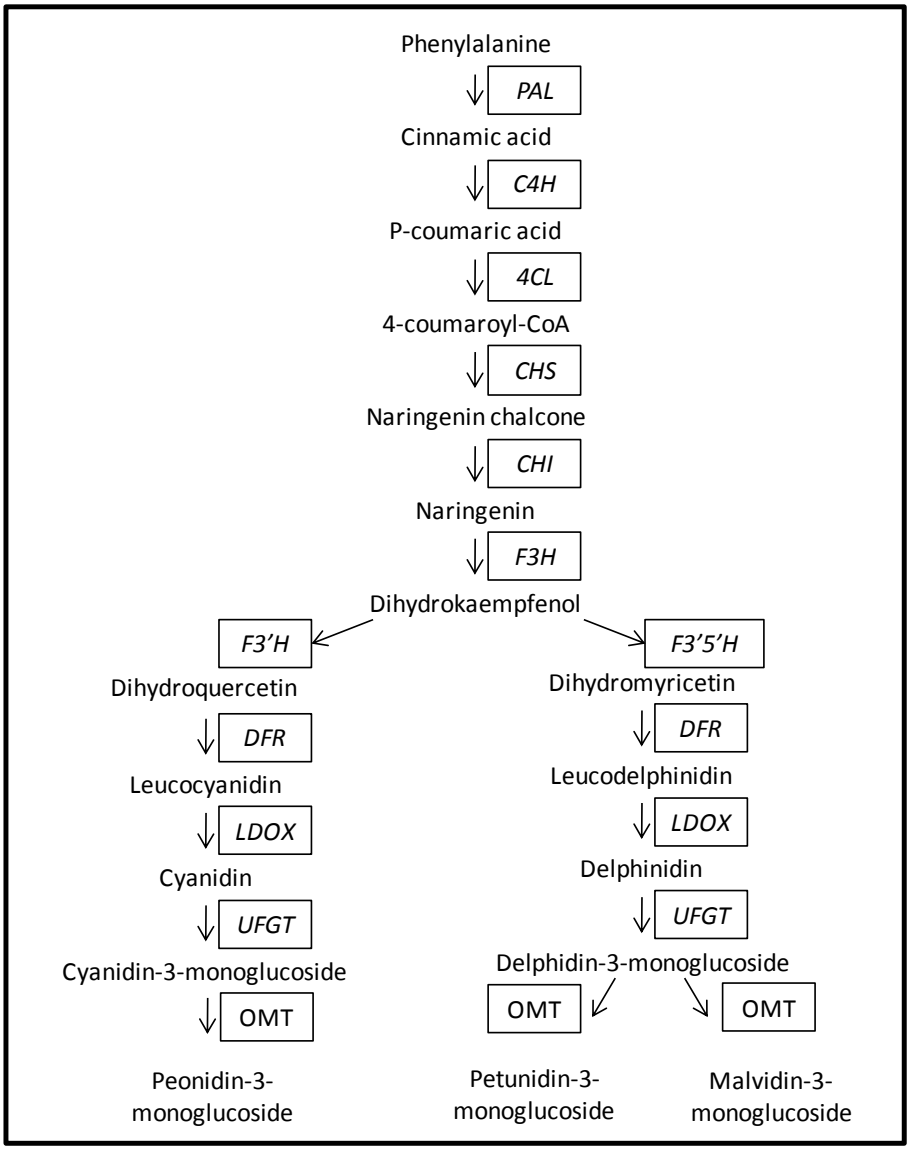


Figure I-1 Schematic drawing of the biosynthetic pathway of anthocyanins.

Abbreviations: PAL, phenylalanine ammonia lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate: coenzyme A ligase; CHS, chalcone synthase; CHI, chalcone isomerise; F3H, flavanone 3-hydroxylase; F3'H, flavonoid 3'-hydroxylase; DFR, dihydroflavonol reductase; LDOX, leucoanthocyanidin dioxygenase; UFGT, UDP-glucose: flavonoid 3-*O*-glucosyl transferase; OMT, O-methyltransferase.

This pathway may be divided in two parts. In the first part, integrated in the general phenylpropanoid pathway, phenylalanine is converted to 4-coumaroyl-CoA. Phenylalanine ammonia lyase (PAL) catalyzes the deamination of L-phenylalanine to *trans*-cinnamic acid. Cinnamate 4-hydroxylase (C4H) converts cinnamic acid into *p*-coumaric acid by hydroxylation. Finally, 4-coumarate: coenzyme A ligase (4CL) activates *p*-coumaric acid producing 4-coumaroyl-CoA (Verpoorte, 2000).

The second part, integrated in the general flavonoids pathway, is the conversion of 4-coumaroyl-CoA into anthocyanins. The first step is the condensation of one molecule of 4-coumaroyl CoA and three molecules of malonyl-CoA to form a 4,2',4',6'-tetrahydroxychalcone (naringenin chalcone), catalyzed by the enzyme chalcone synthase (CHS). The second step is the isomerisation of the naringenin chalcone into a naringenin flavanone. This isomerisation may be catalyzed by the enzyme chalcone isomerase (CHI) but it may also occur spontaneously (Grisebach, 1982; Verpoorte, 2000).

The obtained naringenin flavanone is then hydroxylated to yield dihydroflavonols, either dihydroquercetin or dihydromyricetin, precursors of cyanidin and delphinidin based anthocyanins, respectively. To produce dihydroquercetin, naringenin flavanone must be hydroxylated at the 3' position of the B-ring of the flavonoid by activity of flavonoid 3'-hydroxylase (F3'H) and at position 3 by F3H (flavanone 3-hydroxylase). To obtain dihydromyricetin, naringenin flavanone must be hydroxylated at the 3',5' positions of the B-ring by activity of flavonoid 3'5'-hydroxylase (F3'5'H) and at position 3 by F3H.

Dihydroflavonol 4-reductase (DFR) catalyzes the reduction of the dihydroflavonols at the position 4 to produce leucoanthocyanidins. This is followed by the conversion of leucoanthocyanidins into anthocyanidins

catalyzed by leucoanthocyanidin dioxygenase (LDOX) (Grotewold, 2006).

Addition of a sugar residue in position 3 of the anthocyanidin is catalyzed by UDP-glucose: flavonoid 3-*O*-glucosyltransferase (UFGT) producing anthocyanins. These anthocyanins are cyanidin-3-monoglucoside and delphinidin-3-monoglucoside. *O*-methyltransferase (OMT) catalyzes conversion of cyanidin-3-monoglucoside into peonidin-3-monoglucoside and of delphinidin-3-monoglucoside into petunidin-3-monoglucoside and malvidin-3-monoglucoside (Quattrochio *et al.*, 1993).

The genes coding enzymes involved in the biosynthetic pathway of anthocyanins have been characterised in several plant species (Dooner, 1991; Holton and Cornish, 1995). In grapevine, Sparvoli *et al.* (1994) cloned partial transcripts of genes encoding enzymes of the biosynthetic pathway of anthocyanins (*PAL, CHS, CHI, F3H, DFR, LDOX* and *UFGT*). Boss *et al.* (1996a,b) studied the expression of these genes in white and red cultivars. *UFGT* was found to be expressed in red cultivars only, while the other genes (*PAL*, *CHS*, *CHI*, *F3H*, *DFR*, *LDOX*) were expressed in both cultivars. These observations led to the conclusion that *UFGT* expression was critical for anthocyanin synthesis and grape skin colouration. Kobayashi *et al.* (2001) reached the same conclusion by studying red bud sports of white cultivars. They have found that the promoter and coding sequence of this gene were not different between the red and white sport, suggesting that a regulatory gene should be mutated causing the phenotypic difference.

2.3. Transport and accumulation of anthocyanins

Anthocyanin biosynthesis occurs by the action of a metabolon associated with the cytosolic surface of the endoplasmic reticulum

(Winkel-Shirley, 1999; Winkel, 2004). The transport to the vacuole has been suggested to follow two different models (Grotewold and Davies, 2008). The first model consists on a vesicular transport where vesicles carrying anthocyanins travel from the endoplasmic reticulum to the vacuole fusing with the tonoplast. The second model proposes a ligandin transport where ligandin binded anthocyanins are transported to the tonoplast entering the vacuole through membrane transporters.

Several studies have shown evidence of vesicular transport by observing the presence of vesicle like structures containing anthocyanins in the cytoplasm of *Arabidopsis thaliana* (Poustka *et al.*, 2007), *Lisianthus* (Grotewold *et al.*, 1998) and *Zea mays* (Zhang *et al.*, 2006). Nevertheless, there is also evidence supporting the model based on ligandin transport. Glutathione S-transferases (GSTs) are thought to bind anthocyanins through hydrofobic interactions and escort them to the tonoplast membrane (Springob *et al.*, 2003). These proteins have been found to play a role in vacuolar localization of anthocyanins in *Zea mays* (Conn *et al.*, 2008; Marrs, 1995) and to co-localise with transporter proteins in *Petunia hybrida* (Mueller *et al.*, 2000).

Two main different mechanisms have been proposed for transport across the tonoplast. Primary transport is mediated by ATP-binding cassette transporters (Goodman *et al.*, 2004; Lu *et al.*, 1998; Verrier *et al.*, 2008). Secondary transport involves a gradient of hydrogen ions (Martinoia *et al.*, 2007). Evidence of the involvement of multidrug resistance-associated protein (MRP)-type ABC transporters in vacuolar accumulation of anthocyanins and other phenolic compounds supports the primary transport model (Goodman *et al.*, 2004; Klein *et al.*, 2006; Verrier *et al.*, 2008). On the other hand, the vacuolar uptake of flavonoids was observed to be dependent on a proton gradient and Multidrug and

Toxic Extrusion (MATE) transporters were identified in the tonoplast (Debeaujon *et al.*, 2001; Hopp and Seitz, 1987; Klein *et al.*, 1996; Marinova *et al.*, 2007; Yazaki, 2005). These observations support the hypothesis of secondary transport mechanism for transport across the tonoplast.

In grapevine, a glutathione S-transferase (GST) has been shown to be involved in vacuolar accumulation of anthocyanins. Ageorges *et al.* (2006) verified an expression pattern of this gene matching grape berries colour development. Also, two different transporters were suggested to be involved in the transport of anthocyanins to the vacuole in this species. A translocator homologous to mammalian bilitranslocase and two MATE transporters were identified by Braidot *et al.* (2008) and Gomez *et al.* (2009), respectively.

2.4. Regulation of the biosynthetic pathway of anthocyanins

Transcriptional regulators of flavonoids have been widely studied in Arabidopsis, maize and petunia. Recently, there has been an increasing interest in transcription regulators in grapevine as well (Matus *et al.*, 2009).

Mutation studies on the biosynthetic pathway of anthocyanins have produced two different types of mutants. The first type includes mutations in the genes coding enzymes involved in the biosynthetic pathway of anthocyanins. The second type shows mutations in regulatory genes. In the latter, the expression of several genes involved in the biosynthetic pathway was found to be altered. In many plant species, anthocyanin biosynthesis has been shown to be regulated mainly by two families of transcription factors, the Myb and the Myc families (Baudry *et al.*, 2004; Borovsky *et al.* 2004; Dooner and Robbins, 1991; Holton

and Cornish, 1995; Matus *et al.*, 2010; Payne *et al.*, 2000; Ramsay *et al.* 2003; Robbins *et al.* 2003; Sainz *et al.* 1997; Schwinn *et al.*, 2006; Spelt *et al.* 2000).

Myb family proteins are characterised by a Myb-homologous DNA-binding domain. This domain is defined as the DNA-binding domain of the mammalian proto-oncogene Myb, a region of around 52 amino acids responsible for sequence specific DNA binding. This motif may be repeated three times (R1, R2 and R3) or only twice (R2 and R3), as is most common in plants. Each motif adopts a helix-turn-helix conformation to intercalate in the major groove of the target DNA (Lipsick, 1996; Martin and Paz-Ares, 1997).

Proteins of the plant Myc family have a DNA-binding domain similar to the DNA binding/dimerization domain of animal Myc oncogene (Ludwig *et al.*, 1989). The basic helix-loop-helix (bHLH) domain is composed by two subdomains. One domain is responsible for DNA binding, forming a helical structure which interacts with the major groove of the DNA molecule. The other domain is a HLH region formed by two helices separated by a loop which forms the interface for homo and heterodimerization (Ellenberger, 1994; Ferré-D'Amaré *et al.* 1993).

Other protein families, such as Tryptophan-aspartic acid repeat (WDR or WD40 repeat) family proteins and WRKY transcription factors, have been shown to play a role in the regulation of anthocyanin metabolism (Johnson *et al.*, 2002; Matus *et al.*, 2010; Vetten *et al.*, 1997).

In *Vitis vinifera* L., several *Myb* genes have been observed to be involved in the regulation of the flavonoid metabolism. Evidence has suggested *MybA1* and *MybA2* multiallelic mutations to control the biosynthetic step mediated by *UFGT* (Kobayashi, 2002, 2004; Walker *et al.*, 2007). The absence of anthocyanins has been shown to be

determined by the homozygous presence of a *MybA1* allele with a retrotransposon insertion (*Gret1*) in the gene promoter region (Fournier-Level, 2009; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; This *et al.*, 2007). However, rarely some white cultivars were observed without this insertion, suggesting that this same phenotype may be also influenced by other genes (This *et al.*, 2007).

The variation within coloured cultivars has not been completely understood yet. Four additional polymorphisms in *MybA1* have shown to be associated with pink/red cultivars (This *et al.*, 2007). Recently, Fournier-Level *et al.* (2009) identified four polymorphisms in *MybA1*, *MybA2* and *MybA3* accounting for 23 % of colour variance.

Myb5a, *Myb5b*, *MybPA1* and *MybPA2* were shown to affect expression of genes coding enzymes which catalyze early steps of the biosynthetic pathway of anthocyanins by promoter activation (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Matus *et al.*, 2009; Terrier *et al.*, 2009). Recently, Matus *et al.* (2008, 2010) have classified 108 members of the *Myb* gene family in grapevine according to their structure and similarity to Arabidopsis orthologs and showed the expression pattern of *WDR1* and *MYCA1* to be correlated with anthocyanin accumulation in *Vitis vinifera* L..

In many plant species, Myb, bHLH and WDR factors interact to determine the group of genes expressed (Baudry *et al.*, 2004; Matus *et al.*, 2010; Quattrocchio *et al.*, 1998; Walker *et al.*, 1999). In grapevine, cases have been observed where Myb factors interact with each other (Terrier *et al.*, 2009) or interact with bHLH factors (Bogs *et al.*, 2007).

Environmental effects on grape berries flavonoid composition has been widely studied (Downey and Robbins, 2006). It has been shown that light incidence on berries throughout ripening significantly increases the

expression of genes involved in the biosynthetic pathway of flavonoids (Downey *et al.*, 2004; Jeong *et al.*, 2004; Matus *et al.*, 2009) and their accumulation (Cortell and Kennedy, 2006). Temperature has also been observed to have an effect on flavonoid production and accumulation. High temperatures decrease expression of flavonoid biosynthetic genes and *MybA* genes and increase anthocyanins degradation (Mori *et al.*, 2005, 2007; Yamane *et al.*, 2006). Matus *et al.* (2009) has shown *Myb* genes regulating the biosynthetic pathway of anthocyanins to be affected by light exposure. However, the genes acting upon the final biosynthetic steps were more strongly affected than the ones controlling several steps. Other regulatory genes, included in the *bHLH* and *WDR* families have not been observed to respond in the same way.

2.5. Anthocyanin QTL mapping

According to International Organization of Vine and Wine (OIV) descriptors, colour of grape skin has been classified in B (white), N (black), Rs (rose), G (grey) and Rg (red). White cultivars do not accumulate anthocyanins in the berries' skin (Boss *et al.*, 1996). On the contrary, rose, red, grey and black cultivars accumulate anthocyanins in berries skin. Some cultivars also accumulate anthocyanins in berry flesh. The anthocyanins accumulated vary on type and on concentration, determining colour variation among grapevine cultivars (Mazza and Miniati, 1993). As a continuous trait, colour of berry skin is expected to be determined by several genes or gene variants with subtle contributions each.

Berry colour (with berries having coloured versus non coloured skin) has been observed to have Mendelian segregation (Fischer *et al.*, 2004; Salmaso *et al.*, 2008). This trait has been mapped to linkage group (LG) 2

(Doligez, 2006; Fischer *et al.*, 2004; Salmaso *et al.*, 2008). Colour as a quantitative trait was also mapped to LG2 (Fournier-Level *et al.*, 2009).

The genes involved in the biosynthetic pathway of anthocyanins have been mapped in different LGs. *CHI* was mapped to LG 13, *F3H* to LG4, *DFR* to LG18, *LDOX* to LG2 and *UFGT* to LG16. The gene coding the transcription factor *MybA1* and the colour dichotomous trait were mapped to the same locus in LG2 (Salmaso *et al.*, 2008).

3. Association mapping

Association genetics aims at establishing a correspondence between genotype and phenotype at the population level. Although linkage mapping has been successfully used to identify major genes (Corder *et al.*, 1993; Zielenski and Tsui, 1995) and QTL regions (Alpert and Tanksley, 1996; Stuber *et al.*, 1999), association mapping can provide higher power and resolution for the identification of genetic variants.

LD is the genetic basis for association mapping. Linkage disequilibrium (LD) is the statistical association of alleles at two loci in a population (Balding, 2006). LD is expected to be inversely proportional to the recombination rate as this is the main mechanism of LD break down. Nevertheless, other factors influence LD patterns, such as mutation, selection, genetic drift, population demography and breeding system. Due to the interaction of all these factors, LD is expected to vary between species and also between different regions of the genome. By measuring association between the phenotype of interest and a marker allele in LD with the allele influencing the trait, it is possible to map the locus influencing the phenotype.

Both association and linkage mapping use the co-inheritance of DNA variants to infer the position of genes influencing traits of interest.

However, linkage mapping explores the occurrence of recombination events over few generations in a pedigree, while association mapping uses the information on recombination events over many generations (Nordborg and Tavaré, 2002).

Although still less exploited than in human genetics, the interest in association mapping in plants has increased greatly with the growing gene discovery and availability of high throughput methods of polymorphism detection and genotyping.

3.1. Linkage disequilibrium

Several statistics have been proposed to measure LD. *D'* (Lewontin, 1964) and r^2 (Hill and Robertson, 1968) are the most commonly used measures of pairwise LD. *D'* measures only recombinational history while r^2 summarises recombinational and mutational history (Flint-Garcia, 2003). Several approaches have been suggested to summarise LD over a region. Pairwise values may be averaged or the region may be represented diagrammatically with values being represented by different colours. LD units maps (LDU) are a more comprehensive approach to represent local LD (Maniatis *et al.*, 2002). This method uses a decay parameter to calculate LDU distance. These units are strongly correlated with the recombination rate and reflect historical mutations occurrence (Balding, 2006).

In theory, plants domestication increases LD, since many mechanisms usually involved in this process, such as genetic drift, selection and admixture, contribute to high LD. However, considering the complexity of factors affecting LD pattern and the evolutionary history of each species, this is not always the case. Although some studies found high LD, up to over 10 cM and 50cM in sugarcane and wheat (Jannoo *et al.*,

1999; Maccaferri et al, 2005), other authors found decline of LD within few hundred base pairs in maize (Remington *et al.*, 2001; Tenaillon *et al.* 2001).

In *Vitis vinifera* L., evolutionary processes with contrary effects on LD have taken place. On one hand, as an outcrossing species, descending from obligatory outcrossing ancestors, it is expected to have low values of LD. On the other hand, high LD has been favoured by artificial selection and maintained by vegetative propagation (Barnaud *et al.*, 2006). Studies on grapevine have shown contrasting results on LD decay. Using 38 SSR markers across five linkage groups, Barnaud *et al.* (2006) observed a decrease in the average LD (r^2) to 0.1 within approximately 5cM/650-1080kb. However, using SNP data on target sequences, Lijavetzky *et al.* (2007) and This *et al.* (2007) observed strong LD decay within 100-200 bp and 700 bp, respectively. Recently, a study at the genome level also showed LD to be low at short ranges (average r^2 of 0.18 for a distance of 50 bp) but persistent above background levels up to 3kb (Myles *et al.*, 2010).

3.2. SNPs and SNP discovery approaches

Single Nucleotide Polymorphisms (SNPs) are one nucleotide base differences between two DNA sequences, where the least frequent variant has a frequency of 1 % or greater. These markers are usually bi-allelic although theoretically four different alleles could be observed.

SNPs are generated by mutations. These may be of two different types, transitions or transversions. Transitions occur when a pyrimidine base is replaced by another pyrimidine or a purine by another purine. Transversions result from the replacement of a pyrimidine by a purine or vice versa. Due to the higher number of possible replacements,

transversions would be expected to be more common than transitions. However, transitions occur at a much higher frequency than transversions (Vignal *et al.,* 2002). This bias is thought to be due to frequent spontaneous occurrence of 5-methyl cytosine deamination to thymine, especially in CpG dinucleotides (Cooper and Krawczak, 1989; Wang *et al.*, 1998).

The frequency of SNPs has been observed to vary according to the genomic region. This is influenced by both mutation occurrence and selective pressure. The mutation rate between two nucleotides is affected by the nucleotide base, the sequence surrounding it and the methylation status of the DNA (Edwards *et al.,* 2007). Selection effects cause SNPs under selective pressure to be either maintained or removed from the population (Bamshad and Wooding, 2003; Przeworski, 2002). Accordingly, it has been observed that SNPs are most common in non-coding regions of the genome (Edwards *et al.,* 2007). Within a coding region, SNPs may be synonymous or non-synonymous. Synonymous SNPs do not cause any change in the amino acid sequence and therefore, are also more common than non-synonymous ones (Edwards *et al.*, 2007).

In grapevine, re-sequencing projects have shown a high frequency of SNPs in *Vitis vinifera* L. In target genomic regions, Lijavetzky *et al.* (2007) observed one SNP per 64 bp. Velasco *et al.* (2007) identified four SNPs per kb at the genome level.

These markers are especially attractive as genetic markers in association studies because they are the most frequent type of genetic polymorphism, have a low mutation rate and are highly amenable for automation (Landegren *et al.*, 1998; Vignal *et al.*, 2002).

3.3. Preliminary analyses

To obtain reliable and meaningful results, it is of utmost importance to perform a careful process of preliminary analysis, assessing the quality of the data collected and LD. Quality control includes measuring Hardy-Weinberg equilibrium (HWE), minor allele frequency and genotype missingness. Deviations from HWE may be a consequence of inbreeding, selection, population stratification and association (Balding, 2006). Also genotyping errors such as a tendency to genotype heterozygotes as homozygotes may generate data in Hardy-Weinberg disequilibrium (Gomes *et al.*, 1999; Hosking *et al.*, 2004). Due to the difficulties on the diagnosis of the causes of deviation and the risks involved in using genotyping errors, deviates from HWE at a significance level of 10^{-3} or 10^{-4} are usually removed from the sample (Balding, 2006). Rare alleles are commonly defined as occurring at frequencies between 5 and 10 % (Barnaud *et al.*, 2006; Caldwell *et al.*, 2006; Rhoné *et al.*, 2007; Tenaillon *et al.*, 2001). It is common practice to remove rare alleles prior to association analysis since their presence reduces statistical power (WTCCC, 2007; Ziegler, 2009). Association between the trait of interest and the allele causing phenotypic variation relies on LD between this allele and the genotyped markers. Therefore, measuring LD on the genomic region of interest is key for the success of association mapping.

3.4. Experimental design

Depending on the species and the traits of interest, the design of association mapping studies may vary. Case-control design is suitable for dichotomous traits. In this approach, marker allele frequencies are expected to differ significantly between affected and unaffected individuals in the case of association. When the trait of interest is a

continuous trait, statistical association between the trait and the allelic variants is the evidence of genetic association.

The design of association may also vary on the genome area under study. Genome-wide association mapping looks for association between the trait of interest and variation across the whole genome. On the other hand, the candidate-gene approach tests for association between variation on specific genes or regions of the genome and the trait of interest. These candidate genes or regions are selected mainly considering regions associated with the trait of interest according to previous linkage or association mapping studies. Also known pathways and regulatory processes related with the trait of interest may justify candidate gene selection.

Studies of genetic association are often based on single marker tests but haplotype tests may also be used. Comparisons between the power achieved with each of these approaches have shown contradictory results. Long and Langley (1999), and Morris and Kaplan (2001) concluded that single SNP tests have higher power while Akey *et al.* (2001) reported haplotype tests to achieve higher power.

3.5. Type I and type II errors

The power of an association study is the probability of successfully detecting a true genetic effect. Power is influenced by sample size, linkage disequilibrium (LD) between the genotyped marker and the causal variant, effect size, and marker and causal variant frequencies. Many association study findings were impossible to replicate (Gambaro *et al.*, 2000; Weiss and Terwilliger, 2000). Several factors can lead to spurious association. For example, population structure, relatedness, poor study design and inaccurate phenotypic data. Nevertheless, population

stratification and more recently cryptic relatedness have received a great deal of attention. On the other hand, the inability to detect true effects has been much less debated. Besides poor study design this may be caused by pleiotropic and epistatic interactions and environment effects (Cardon and Bell, 2001).

3.5.1. Structure

Population structure, also called population stratification, is the case where a population comprises subgroups of individuals characterised by different allele frequencies (Cardon and Bell, 2001). Population structure may be caused by nonrandom mating between groups due to selection, geographic isolation followed by genetic drift and population admixture of populations with different allele frequencies (Hoggart *et al.*, 2004; Ziv and Burchard, 2003).

Population structure may give rise to false associations if the trait of interest is more common in one subpopulation and as a consequence associates with any allele with higher frequency in this subpopulation (Pritchard and Rosenberg, 1999). Many methods have been developed to address this problem. Devlin and Roeder (1999) developed *Genomic Control* (GC), a methodology to deal with population structure in association mapping. This method adjusts significance tests bias with an inflation factor calculated from random markers assuming that these are not associated with the phenotype and that structure has a similar effect on all loci. This approach is limited to biallelic markers.

Another method based on statistical correction using random markers was proposed by Pritchard (2000). This methodology uses a Bayesian clustering approach that estimates the proportion of each individual's variation that came from each subpopulation. This proportion is then used

as a covariate in association tests. This method can be used with various types of markers. However, it assumes that these loci are unlinked and is computationally very demanding. Thornsberry *et al.* (2001) extended this method to quantitative traits.

A method for dealing with structure problems with increasing popularity in genome-wide association studies uses principal component analysis (Price *et al.*, 2006). A classical principal component analysis is used on genotype data of random markers to infer continuous axes of variation explaining as much of the total variation as possible. These axes are then used to adjust phenotypes and genotypes for association tests. This approach requires a large number of markers, but provides a computationally effective way of handling structure appropriately (Price *et al.*, 2006).

3.5.2. Relatedness

The fact that some individuals in the sample may be close relatives, unbeknown to the researcher, may lead to false positive association results. This problem is of major concern in plant species where a certain degree of relatedness is expected due to selection and breeding history (Zhu *et al.*, 2008).

Recently, several studies suggest that correction for pairwise relatedness besides structure significantly decreases false positives and increases power (Kang *et al.*, 2008; Malosetti *et al.*, 2007; Yu *et al.*, 2006; Zhao *et al.*, 2007). These observations agree with the idea that structure and relatedness assessments capture different levels of variation (Yu *et al.*, 2006). Yu *et al.* (2006) developed a mixed model approach to account for population structure and cryptic relatedness while testing for association. In animal breeding studies, the mixed model has been

traditionally used for genetic evaluation of livestock. In these studies, pedigree records are available, what is usually not the case for genetic association studies. As a consequence, in the model used by Yu *et al.* (2006), relatedness as well as structure, are estimated using random unlinked molecular markers. Structure is estimated using the method presented by Pritchard *et al.* (2000). The estimated proportion of each individual's variation that came from each subpopulation is then included in the mixed model as a covariate. To account for relatedness, Yu *et al.* (2006) included in the model a relatedness matrix based on a Ritland's kinship coefficient (Ritland, 1996). This is estimated based on the probability of Identity by State (IBS) between two individuals adjusted to the average probability of IBS between random individuals in the population (Ritland, 1996; Yu *et al.*, 2006).

Several other pairwise methods of kinship inference using molecular markers have been developed (e.g. Li *et al.*, 1993; Loiselle *et al.*, 1995; Lynch & Ritland, 1999; Queller and Goodnight, 1989; Wang, 2002). The simplest method to assess genetic similarity using molecular markers is to calculate the proportion of alleles shared between two individuals over all genotyped loci, a measure first proposed by Chakraborty and Jin (1993). Kang *et al.* (2008) and Zhao *et al.* (2007) have shown this matrix to correct at least as effectively for relatedness among sampled individuals as the matrix based on Ritland's kinship.

3.5.3. Statistical interactions

Gene-gene interaction or epistasis has been often mentioned as a reason for non-replicable association studies results (Culverhouse *et al.*, 2002; Moore, 2003). Studies which do not account for interactions but examine genes only in isolation may miss true effects on phenotype if

these genes act through a complex mechanism of interactions with each other (Cordell, 2009). Many statistical methods are employed for testing epistasis. The most common are regression models; however, other methods have been proposed and are revised by Cordell (2009; Chanda *et al.*, 2007; Dong *et al.*, 2008; Kang *et al.*, 2008; Moore *et al.*, 2006; Yang *et al.*, 2008). Large sample sizes may create computational burden and multiple testing problems. One of the approaches to alleviate this is to do a preselection of locus to test based on a significance threshold or biologic interpretation (Cordell, 2009).

3.5.4. Correction for multiple testing

Association studies often involve large sample sizes and high numbers of markers genotyped. The number of tests performed quickly raises to very high numbers increasing the chances of obtaining false positive results.

Correcting appropriately for this is of utmost importance and is a compromise between reducing the risk of obtaining false-positive results and maintaining the ability to detect true associations. Bonferroni correction, assumes marker independence which is not verified when high LD is found between the genotyped markers. This correction will then often overcorrect for false-positives, reducing the power to detect true association (Cardon and Bell, 2001). The use of dataset permutations has been advised as the best method to correct for multiple testing (Cardon and Bell, 2001).

Despite the large number of studies on grape colour there is still no clear understanding on the genetics underlying this phenotype both between and within cultivars. Berry colour is one of the most important

traits of this crop. It influences grape and wine marketing ability and the anthocyanin concentration has a positive impact on human health. The lack of extensive phenotypic characterisation of large samples, namely including the cultivars present in germplasm collections, is one of the difficulties hampering the research on this and other traits. The relationships between the different approaches to the characterisation of the colour phenotype are largely unexplored. The studies performed to date at the clonal level have focused mainly on chimerism and on the categorical variation of colour. Also between cultivars, the works undertaken have concentrated on either presence or absence of colour, categorical variation of colour or total concentration of anthocyanins. Once the genetic factors controlling colour are known, marker assisted selection will greatly enhance the breeding processes for higher quality grapes and wine.

4. Objectives

The main objective of this study was to understand the genetic variation underlying grape colour. This objective may be divided in the following more specific objectives:

- Investigate the variation underlying phenotypic differences in anthocyanin concentration in berries skin between *Vitis vinifera* L. clones both at DNA sequence variation and at differential gene expression level;
- Characterise phenotypic diversity between different grapevine cultivars concerning colour, including different types of colour characterisation (visual assessment, anthocyanin concentration and relative abundance);

- Understand the relationships between the different forms of colour phenotypic characterisation;
- Analyse the impact of covariates, such as plants viral infection and berries maturity state on anthocyanin content for association mapping purposes;
- Identify DNA variation among candidate genes for colour and anthocyanin content;
- Find association between genetic variants and colour of grape berries defined by a thorough range of colour related phenotypes including visual assessment, anthocyanin concentration and relative abundance (RA).

5. References

Abuja, P.M., Murkovic, M. Pfannhauser, W. (1998). Antioxidant and prooxidant activities of elderberry (*Sambucus nigra)* extract in low-density lipoprotein oxidation. *Journal of Agricultural and Food Chemistry* **46**, 4091-96.

Adam-Blondon, A.-F., Roux, C., Claux, D., Butterlin, G., Merdinoglu, D. (2004). Mapping 245 SSR markers on the *Vitis vinifera* genome: a tool for grape genetics. *Theoretical and Applied Genetics* **109**, 1017-1027.

Ageorges, A., Fernandez, L., Vialet, S., Merdinoglu, D., Terrier, N., *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.

Akey, J., Jin, L., Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* **9**, 291-300.

Al-Awwadi, N.A., Araiz, C., Bornet, A., Delbosc, S., Cristol, J.P., *et al.* (2005). Extracts enriched in different polyphenolic families normalise increased cardiac NADPH oxidase expression while having differential effects on insulin resistance, hypertension, and cardiac hypertrophy in high-fructose-fed rats. *Journal of Agricultural and Food Chemistry* **53**, 151-57.

Alcalde-Eon, C., Saavedra, G., Pascual-Teresa, S.D., Rivas-Gonzalo, J.C. (2004). Liquid chromatography-mass spectrometry identification of anthocyanins of isla oca (*Oxalis tuberose*, Mol.) tubers. *Journal of Chromatography A*, **1054**, 211-215.

Alleweldt, G. and Dettweiler, E. (1994). The genetic resources of *Vitis*: world list of grapevine collections, 2nd edn. BAZ IRZ Geilweilerhof, Siebeldingen.

Alpert, K.B. and Tanksley, S.D. (1996). High-resoltuion mapping and isolation of a yeast artificial chromosome contig containing fw2.2: A major fruit weight quantitative trait locus in tomato. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 15503-15507.

Andersen, Ø.M. and Jordheim, M. (2006). The anthocyanins. In *Flavonoids: chemistry, biochemistry and applications.* Andersen, Ø.M. and Markham, K.R., Eds. CRC Press: Boca Raton, pp.471-552.

Andersen, Ø.M., Jordheim, M. (2008). In *Anthocyanin – food applications.* Proceedings of the 5th International Congress on Pigments Foods For Quality and Health, Helsinki, Finland, Aug 14 – 16.

Arnold, C., Gillet, F., Gobat, J.M. (1998). Situation de la vigne sauvage *Vitis vinifera* subsp. *Silvestris* en Europe. *Vitis* **37**, 159-170.

Arroyo-García, R., Ruiz-García, L., Bolling, L., Ocete, R., López, A., *et al.* (2006). Multiple origins of cultivated grapevine (*Vitis vinifera* L. ssp. *sativa)* based on chloroplast DNA polymorphisms. *Molecular Ecology* **15**, 3707-3714.

Bakker, J., Bridle, P., Honda, T., Saito, N., Kuwano, H., *et al.* (1997). Identification of an anthocyanin occurring in some red wines. *Phytochemistry* **44**, 1375-1382.

Balding, D. J. (2006). A tutorial on statistical methods for populataion association studies. *Nature Reviews Genetics* **7**, 781-791.

Ball, R.D. (2007). Statistical analysis and experimental design. In *Association Mapping in Plants.* Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E., De Silva, H.N., Eds. Springer Science: New York, pp 133-196.

Bamshad, M. and Wooding, S.P. (2003). Signatures of natural selection in the human genome. *Nature Reviews Genetics* **4**, 99-111.

Baneh, H.D., Mohammadi, S.A., Mahmoudzadeh, H., Mattia, F., Labra M. (2009). Analysis of SSR and AFLP markers to detect genetic diversity among selected clones of grapevine (*Vitis vinifera* L.) cv. Keshmeshi. *South African Journal of Enology and Viticulture* **30**, 38-42.

Barnaud, A., Lacombe, T., Doligez, A. (2006). Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L. *Theoretical and Applied Genetics* **112**, 708-716.

Baudry, A., Heim, M. A., Dubreucq, B., Caboche, M., Weisshaar, B., *et al.* (2004). TT2, TT8 and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant Journal* **39**, 366-380.

Bertsch, C., Kieffer, F., Maillot, P., Farine, S., Butterlin, G., *et al.* (2005). Genetic chimerism of *Vitis vinifera* cv Chardonnay 96 is maintained through organogenesis but not somatic embryogenesis. *BMC Plant Biology* **5**, 1-7.

Bisson, J. (1995). The principal ecogeographical groups in French grapevines assortment. *Journal International des Sciences de la Vigne et du Vin* **29**, 63-68.

Bogs, J., Jaffé, F.W., Takos, A.M., Walker, A.R., Robinson, S.P. (2007). The grapevine transcription factor *VvMYBPA1* regulates proanthocyanidin synthesis during fruit development. *Plant Physiology* **143**, 1347-1361.

Borovsky, Y., Oren-Shamir, M., Ovadia, R., De Jong, W., Paran, I. (2004). The A locus that controls anthocyanin accumulation in pepper encodes a MYB transcription factor homologous to *Anthocyanin2* of Petunia. *Theoretical and Applied Genetics* **109**, 23-29.

Boursiquot, J.M. and This, P. (1999). Essai de définition du cépage. *Progrès Agricole Viticole* **116**: 359–361.

Boss, P.K., Davies, C., Robinson, S.P. (1996). Analysis of the expression of anthocyanin pathway genes in developing *Vitis vinifera* L. cv Shiraz grape berries and the implications for pathway regulation. *Plant Physiology* **111**, 1059-1066.

Boss, P.K., Davies, C., Robinson, S.P. (1996). Expression of anthocyanin biosynthesis pathway genes in red and white grapes. *Plant Molecular Biology* **32**, 565-569.

Braidot, E., Petrussa, E., Bertolini, A., Peresson, C., Ermacora, P., *et al.* (2008). Evidence for a putative flavonoid translocator similar to mammalian bilitranslocase in grape berries (Vitis vinifera L.) during ripening. *Planta* **228**, 203-213.

Brouillard, R. (1982). Chemical structure of anthocyanins. In *Anthocyanins as food colors*. Markakis, P., Ed. Academic Press: New York, pp 1-38.

Cabaleiro, C., Segura, A., Garcia-Berrios, J.J. (1999). Effects of grapevine leafroll-associated virus 3 in the physiology and must of *Vitis vinifera* L. cv. Albarino following contamination in the field. *American Journal of Enology and Viticulture* **50**, 40-44.

Cacho, J., Fernandez, P., Ferreira, V., Castello, J.E. (1992). Evolution of five anthocyanin-3-glucosides in the skin of the Tempranillo, Moristel, and Garnacha grape varieties and influence of climatological variables. *American Journal of Enology and Viticulture* **43**, 244-248.

Caldwell, K.S., Russel, J., Langridge, P., Powell, W. (2006). Extreme population dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* **172**, 557-567.

Cardon, K.T. and Bell, J.I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics* **2**, 91-99.

Castro, A.J., Carapito, C., Zorn, N., Magne, C., Leize, E., *et al.* (2005). Proteomic analysis of grapevine (Vitis vinifera L.) tissues subjected to herbicide stress. *Journal of Experimental Botany* **56**, 2783-2795.

Carcamo, C., Procedo, I., Arroyo-García, R. (2010). Detection of polymorphism in ancient Tempranilo clones (*Vitis vinifera* L.) using microsatellite and retrotransposon markers. *Iraninan Journal of Biotechnology* **8**, 1-23.

Carreño, J., Almeida, L., Martínez, A., Fernandéz-López. (1997). Chemotaxonomical classification of red table grapes based on anthocyanin profile and external colour. *Lebensmittel-Wissenschaft und Technologie* **30**, 259-265.

Cervera, M.T., Cabezas, J.A., Sancha, J.C., Martínez de Toda, F., Martínez-Zapater, J.M. (1998). Application of AFLPs to the characterization of grapevine *Vitis vinifera* L. genetic resources. A case study with accessions from Rioja (Spain). *Theoretical and Applied Genetics* **97**, 51-59.

Cervera, M.T., Cabezas, J.A., Sanchez-Escribano, E., Cenis, J.L., Martinez-Zapater, J.M. (2000). Characterization of genetic variation within table grape varieties (*Vitis vinifera* L.) based on AFLP. *Vitis* **39**, 109-114.

Cervera, M.T., Rodriguez, I., Cabezas, J.A., Chavez, J., Martinez-Zapater, J.M., *et al.* (2001). Morphological and molecular characterization of grapevine accessions as Albillo. *American Journal of Enology and Viticulture* **52**, 127-135.

Chakraborty, R. and Jin, L. (1993). Determination of relatedness between individuals using DNA-fingerprinting. *Human Biology* **65**, 875-895.

Chanda, P., Zhang, A., Brazeau, D., Sucheston, L., Freudenheim, J.L., *et al.* (2007). Information-theoretic metrics for visualizing gene-environment interactions. *American Journal Human Genetics* **81**, 939-963.

Châtaignier, C. (1995). La Transcaucasie au Néolithique et au Chalcolitique. *British Archaeological Series* **624**, 1-240.

Cholet, C. and Darné, R. (2004). Evolution of the contents in soluble phenolic compounds, in proanthocyanic and in anthocyanins of shot grape berries of *Vitis vinifera* L. during their development. *Journal International des Sciences de la Vigne et du Vin* **38**, 171-180.

Collet, S.A.D., Collet, M.A., Machado, M.D.P.S. (2005). Differential gene expression for isozymes in somatic mutants of *Vitis vinifera* L. (*Vitaceae*). *Biochemical systematic and ecology* **33**, 691-703.

Conn, S., Curtin, C., Bezier, A., Franco, C., Zhang, W. (2008). Purification, molecular cloning, and characterization of glutathione S-transferases (GSTs) from pigmented *Vitis vinifera* L. cell suspension cultures as putative anthocyanin transport proteins. *Journal of Experimental Botany* **59**, 3621-3634.

Cooper, D.N. and Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Human Genetics* **83**, 181-188.

Cordell, H. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews* **10**, 392-404.

Corder, E. H., Saunder, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., *et al.* (1993). Gene dose of apolopoprotein-E type-4 allele and the risk of alzheimers-disease in late-onset families. *Science* **261**, 921-923.

Cortell, J. M. and Kennedy, J. A. (2006). Effect of shading on accumulation of flavonoid compounds in (*Vitis vinifera* L.) pinot noir fruit and extraction in a model system. *Journal of Agricultural and Food Chemistry* **54**, 8510-8520.

Cramer, G.R., Ergul, A., Grimplet, J., Tillet, R.L., Tattersall, E.A.R., *et al.* (2007). Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. *Functional & Integrative Genomics* **7**, 111-134.

Crespan, M. (2004). Evidence on the evolution of polymorphim of microsatellite markers in varieties of *Vitis vinifera* L. *Theoretical and Applied Genetics* **108**, 231-237.

Culverhouse, R., Suarez, B.K., Lin, J., Eich, T. A perspective on epistasis: limits of models displaying no main effect. *American Journal of Human Genetics* **70**, 461-471.

Cunha, J., Santos, M.T., Brazão, J., Carneiro, L.C., Veloso, M., *et al.* (2010). Genetic diversity in Portuguese native *Vitis vinifera* L. ssp. *vinifera* and ssp. *sylvestris*. *Czech Journal of Genetic Plant Breeding* **46**, S54-S56.

Cunha, J., Santos, M.T., Carneiro, L.C., Fevereiro, P., Eiras-Dias, J.E. (2009). Portuguese traditional grapevine cultivars and wild vines (*Vitis vinifera* L.) share morphological and genetic traits. *Genetic Resources and Crop Evolution* **56**, 975-989.

Dalbo, M.A., Ye, G.N., Weeden, N.F., Steinkellner, H., Sefc, K.M., *et al.* (2000). A gene controlling sex in grapevines placed on a molecular marker-based genetic map. *Genome* **43**, 333-340.

D'Amato, F. (1977). Nuclear cytology in relation to development. In *Developmental and cell biology series.* Abercrombie, M., Newth, D.R., Torrey, J.G., Eds. Cambridge University Press: Cambridge, pp 120-134.

Da Silva, F.G., Iandolino, A., Al-Kayal, F., Bohlmann, M.C.,Cushman, M.A. *et al.* (2005). Characterizing the grape transcriptome. Analysis of expressed seuqence tags from multiple *Vitis* species and development of a compendium of gene expression during berry development. *Genome Analysis* **139**, 574-597.

Day, A.P., Kemp, H.J., Bolton, C., Hartog, M., Stansbie, D. (1997). Effect of concentrated red grape juice consumption on serum antioxidant capacity and low-density lipoprotein oxidation. *Annals of Nutrition and Metabolism* **41**, 353-57.

Debeaujon, I., Peeters, A.J., Léon-Kloosterziel, K.M., Koornneed, M. (2001). The *TRANSPARENT TESTA12* gene of *Arabidopsis* encodes a multidrug secondary transporter-like protein required for flavonoids

sequestration in vacuoles of the seed coat endothelium. *The Plant Cell* **13**, 853-871.

Delgado-Vargas, F. and Paredes-Lopez, O. (2003). Anthocyanins and betalains. In *Natural colorants for food and nutraceutical uses.* Delgado-Vargas, F. and Paredez-Lopes, O., Eds. CRC Press: Boca Raton, pp 167-219.

Deluc, L.G., Grimplet, J., Wheatley, M.D., Tillet, R.L., Quilici, D.R., *et al.* (2007). Transcriptomic and metabolite analyses of Cabernet Sauvignon grape berry development. *BMC Genomics* **8**, 429-451.

Dermen, H. 1960. Nature of plant sports. *American Horticultural Magazine* **39**: 123–173.

De Saporta, G. (1879). *Le monde des plantes avant l'apparition de l'homme.* Masson, G., Ed. Libraire de L'Académie de Médecine: Paris, pp 416.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997-1004.

Doligez, A., Adam-Blondon, A. F., Cipriani, G., Di Gaspero, G., Laucou, V., *et al.* (2006). An integrated SSR map of grapevine based on five mapping populations. *Theoretical and Applied Genetics* **113**, 369-382.

Doligez, A., Bouquet, A., Danglot, Y., Lahogue, F., Riaz, S., *et al.* (2002). Genetic mapping of grapevine (*Vitis vinifera* L.) applied to the detection of QTLs for seedlessness and berry weight. *Theoretical and Applied Genetics* **105,** 780-795.

Dong, C., Chu, X., Wang, Y., Wang, Y., *et al.* (2008). Exploration of gene-gene interaction effects using entropy-based methods. *European Journal of Human Genetics* **16**, 877-905.

Dooner, H.K. and Robbins T.P. (1991). Genetic and developmental control of anthocyanin biosynthesis. *Annual Review of Genetics* **25**, 173-179.

Doucleff, M., Jin, Y., Gao, F., Riaz, S., Krivanek, A.F., *et al*. (2004). A genetic linkage map of grape, utilising *Vitis rupestris* and *Vitis arizonica. Theoretical and Applied Genetics* **109**, 1178-1187.

Downey, M. O., Dokoozlian, N. K., Krstic, M. P. (2006). Cultural practice and environmental impacts on the flavonoid composition of grapes and wine: a review of recent research. *American Journal of Enology and Viticulture* **57**, 257-268.

Downey, M.O., Harvey, J.S., Robinson, S.P. (2004). The effect of bunch shading on berry deveolpment and flavonoid accumulation in Shiraz grapes. *Australian Journal of Grape and Wine Research* **10**, 55-73.

Duchene, E., Butterlin, G., Claudel, P., Dumas, V., Jaegli, N., *et al.* (2009). A grapevine (*Vitis vinifera* L.) deoxy-d-xylulose synthase gene colocates with a major quantitative trait loci for terpenol content. *Theoretical and Applied Genetics* **118**, 541-552.

Eder, R. (2000). Pigments. In *Food Analysis by HPLC*. Nollet, L.M.L., Ed. Marcel Dekker: New York, pp 845-880.

Edwards, D., Forster, J.W., Chagné, D., Batley, J. What are SNPs? In *Association mapping in plants.* Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E., De Silva, H.N., Eds. Springer Science: New York, pp 133-196.

Einset, J. and Pratt, C. (1975). Grapes. In *Advances in Fruit Breeding.* Janick, J. and Moore, J.N., Eds. Purdue University Press: West Lafayette, pp 130-153.

Ellenberger, T. (1994). Getting a grip on DNA recognition – structures of the basic region leucine-zipper, and the basic region helix-loop-helix

DNA-binding domains. *Current Opinion in Structural Biology* **4**, 12-21.

Espinosa, C., Vega, A., Medina, C., Schlauch, K., Cramer, G., *et al.* (2007). Gene expression associated with compatible viral diseases in grapevine cultivars. *Functional & Integrative Genomics* **7**, 95-110.

Esteban, M.A., Villanueva, M.J., Lissarrague, J.R. (2001). Effect of irrigation on changes in the anthocyanin composition of the skin of cv. Tempranillo (*Vitis vinifera* L.) grape berries during ripening. *Journal of the Science of Food and Agriculture* **81**, 409-420.

Food and Agriculture Oganization (FAO) STAT database. http:// www.fao.org (accessed Sep 7, 2010).

Faria, M.A., Beja-Pereira, M., Martins, A., Ferreira, M.A., Nunes, M.E.S. (2004). Grapevine clones discriminated using stilbene synthase-chalcone synthase markers. *Journal of the Science of Food and Agriculture* **84**, 1186-1192.

Fernandez, L., Doligez, A., Lopez, G., Thomas, M.R., Bouquet, A., *et al.* (2006). Somatic chimerism, genetic inheritance, and mapping of the *fleshless berry (flb)* mutation in grapevine (*Vitis vinifera* L.). *Genome* **49**, 721-728.

Ferré-D'Amaré, A.R., Prendergast, G.C., Ziff, E.B., Burley, S.K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* **363**, 38-45.

Figueiredo, A., Fortes, A.M., Ferreira, S., Sebastiana, M., Choi, Y.H., *et al.* (2008). Transcriptional and metabolic profiling of grape (*Vitis vinifera* L.) leaves unravel possible innate resistance against pathogenic fungi. *Journal of Experimental Biology* **59**, 3371-3381.

Fischer, B. M., Salakhutdinov, I., Akkurt, M., Eibach, R., Edwards, K J., *et al.* (2004). Quantitative trait locus analysis of fungal disease

resistance factors on a molecular map of grapevine. *Theoretical and Applied Genetics* **108**, 501-515.

Flint-Garcia, S.A. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357-374.

Fong, R.A., Webb, A.D., Kepner, R.E. (1974). Acylated anthocyanins in a hybrid *Vitis* variety. *Phytochemistry* **13**, 1001-1004.

Fournand, D., Vicens, A., Sidhoum, L., Souquet, J., Moutounet, M., *et al.* (2006). Accumulation and extractability of grape skin tannins and anthocyanins at different advances physiological stages. *Journal of Agriculture and Food Chemistry* **54**, 7331-7338.

Fournier-Level, A., Le Cunff, L., Gomez, C., Doligez, A., Ageorges, A., *et al.* (2009). Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. ssp. *sativa*) berry: a quantitative trait locus to quantitative trait nucleotide integrated study. *Genetics* **183**, 1127-1139.

Franks, T., Botta, R and Thomas, M.R. (2002). Chimerism in grapevines: implications for cultivar identity, ancestry and genetics improvement. *Theoretical and Applied Genetics* **104**, 192-199.

Furiya, T., Suzuki, S., Sueta, T., Takayanagi, T. (2009). Molecular characterization of a bud sport of Pinot gris bearing white berries. *American Journal of Enology and Viticulture* **60**, 66-73.

Galet, P. (1988). *Cepages et Vignobles de France. Tome I, Les Vignes Americaines, 2nd edn*. Galet, P., Ed. Imprimerie Charles Déhan: Montpellier, pp 553.

Gambaro, G., Angiani, F., D'Angelo, A. (2000). Association studies of genetic polymorphisms and complex disease. *The Lancet* **355**, 308-311.

Gil, M. and Yuste, J. (2004). Phenolic maturity of Tempranillo grapevine trained as goblet, under different soil and climate conditions in the Duero valley area. *Journal International des Sciences de la Vigne et du Vin* **38**, 81-88.

Giribaldi, M., Perugini, L., Sauvage, F.X., Schubert, A. (2007). Analysis of protein changes during grape berry ripening by 2-DE and MALDI-TOF. *Proteomics* **7**, 3154-3170.

Giusti, M.M. and Wrolstad, R.E. (2003). Acylated anthocyanins from edible sources and their applications in food systems. *Biochemical Engineering Journal* **14**, 217-225.

Goheen, A.C., Harmon, F.N., Weinberger, J.H. (1958). Leafroll (white emperor disease) of grapes in California. *Phytopathology* **48**, 51-54.

Goldy, R.G., Maness, E.P., Stiles, H.D., Clark, J.R., Wilson, M.A. (1989). Pigment quantity and quality characteristics of some native *V. rotundifolia* Michx. *American Journal of Enology and Viticulture* **40**, 253-258.

Gomes, I., Collins, A., Lonjou, C., Thomas, N.S., Wilkinson, J., *et al.* (1999). Hardy-Weinberg quality control. *Annals of Human Genetics* **63**, 535-538.

Gomez, C., Terrier, N., Torregrossa, L., Vialet, S., Fournier-Level, A., *et al.* (2009). Grapevine MATE-type proteins act as vacuolar H^+-dependent acylated anthocyanin transporters. *Plant Physiology* **150**, 402-415.

Goodman, C.D., Casati, P., Walbot, V. (2004). A multidrug resistance-associated protein involved in anthocyanin transport in *Zea mays*. *The Plant Cell* **16**, 1812-1826.

Grando, M.S., Bellin, D., Edwards, K.J., Pozzi, C., Stefanini, M., *et al.* (2003). Molecular linkage maps of *Vitis vinifera* L. and *Vitis riparia* Mchx. *Theoretical and Applied Genetics* **106**, 1213-1224.

Grassi, F., Labra, M., Imazio, S., Spada, A., Sgorbati, S., *et al.* (2003). Evidence of a secondary grapevine domestication centre detected by SSR analysis. *Theoretical and Applied Genetics* **107**, 1315-1320.

Grimplet, J., Cramer, G.R., Dickerson, J.A., Mathiason, K., Hemert, J.V., *et al.* (2009). VitisNet: "omics" integration through grapevine molecular networks. *PLoS One* **4**, e8365.

Grimplet, J., Deluc, L.G., Tillet, R.L., Wheatley, M.D., Schlauch, K.A., *et al.* (2007). Tissue-specific mRNA expression profiling in grape berry tissues. *BMC Genomics* **8**, 187.

Grisebach, H. (1982). Biosynthesis of anthocyanins. In *Anthocyanins as food colors.* Markakis, P., Ed. Academic Press: New York, pp 69-92.

Grotewold, E. (2006). The genetics and biochemistry of floral pigments. *Annual Review of Plant. Biology* **57**, 761-780.

Grotewold, E., Chamberlin, M., Snook, M., Siame, B., Butter, L., *et al.* (1998). Engineering secondary metabolism in maize cells by ectopic expression of transcription factors. *Plant Cell* **10**, 721-740.

Grotewold, E. and Davies, K. (2008). Trafficking and sequestration of anthocyanins. *Natural Product Communications* **3**, 1251-1258.

Harborne, J.B. and Harborne, A.J. (1998). *Phytochemical methods: a guide to modern techniques of plant analysis.* Kluwer Academic Publishers: London, pp 226.

Heredia, F.J., Francia-Aricha, E.M., Rivas-Gonzalo, J.C., Vicario, I.M., Santos-Buelga, C. (1998). Chromatic characterization of anthocyanins from red grapes - I. pH effects. *Food Chemistry* **63**, 491-98.

Hill, W.G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226-231.

Hocquigny, S., Pelsy, F., Dumas, V., Kindt, S., Heloir, M.C., *et al.* (2004). Diversification within grapevine cultivars goes through chimeric states. *Genome* **47**, 579-589.

Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G., McKleigue, P.M. (2004). Design and analysis of admixture mapping studies. *American Journal of Human Genetics* **74**, 965-978.

Holton, T.A. and Cornish, E.C. (1995). Genetics and Biochemistry of Anthocyanin Biosynthesis. *The Plant Cell* **7**, 1071-1083.

Hopp, W. and Seitz, H.U. (1987). The uptake of acylated anthocyanin into isolated vacuoles from a cell suspension culture of *Daucus carota*. *Planta* **170**, 74.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., *et al.* (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics* **12**, 395-399.

Hrazdina, G. and Franzese, A.J. (1974). Structure and properties of the acylated anthocyanins in *Vitis* species. *Phytochemistry* **13**, 225-229.

Imazio, S., Labra, M., Grassi, F., Winfield, M. Bardini, M. and Scienza, A. (2002). Molecular tools for clone identification: the case of the grapevine cultivar 'Traminer'. *Plant Breeding* **121**, 531-535.

Jaillon, O., Aury, J-M., Noel, B., Policriti, A., Clepet, C. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467.

Jannoo, N., Grivet, L., Dookun, A., D'Hont, A., Glaszmann, J.C. (1999). Linkage disequilibrium among modern sugarcane cultivars. *Theoretical and Applied Genetics* **99**, 1053-1060.

Jellouli, N., Ben Jouira, H., Skouri, H., Ghorbel, A., Gourgouri, A., *et al.* (2008). Proteomic analysis of Tunisian grapevine cultivar Razegui under salt stress. *Journal of Plant Physiology* **165**, 471-481.

Jeong, S.T., Goto-Yamamoto, N., Kobayashi, S., Esaka, A. (2004). Effects of plant hormones and shading on the accumulation of anthocyanins and the expression of anthocyanin biosynthetic genes in grape berry skins. *Plant Science* **167**, 247-252.

Johnson, C.S., Kolevski, B., Smyth, D.R. (2002). TRANSPARENT TEST GLABRA2, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. *Plant Cell* **14**, 1359-1375.

Jing, P., Bomser, J.A., Schwartz, S.J., He, J., Magnuson, B.A., *et al.* (2008). Structure-function relationships of anthocyanins from various anthocyanin-rich extracts on the inhibition of colon cancer cell growth. *Journal of Agriculture and Food Chemistry* **56**, 9391-98.

Kaeppler, S.M., Kaeppler, H. F., Rhee, Y. (2000). Epigenetic aspects of somaclonal variation in plants. *Plant Molecular Biology* **43**, 179-188.

Kamei, H., Hashimoto, Y., Koide, T., Kojima, T., Hasegawa, M. (1998). Anti-tumor effect of methanol extracts from red and white wines. *Cancer Biotherapy and Radiopharmaceuticals* **13**, 447-52.

Kamei, H., Kojima, T., Hasegawa, M., Koide, T., Umeda, T., *et al.* (1995). Supression of tumor cell growth by anthocyanins in vitro. *Cancer Investigation* **13**, 590-94.

Kang, H.M., Zautlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., *et al.* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-1723.

Kennedy, J.A., Matthews, M.A., Waterhouse, A.L. (2002). Effect of maturity and vine water status on grape skin and wine flavonoids. *American Journal of Enology and Viticulture* **53**, 268-274.

Klein, M., Burla, B., Martinoia, E. (2006). The multidrug resistance-associated protein (*MRP*/ABCC) subfamily of ATP-binding cassette transporters in plants. *FEBS Letters* **580**, 1112-1122.

Kobayashi, S., Goto-Yamamoto, N., Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. *Science* **304**, 982.

Kobayashi, S., Ishimaru, M., Ding, C.K., Yakushiji, H., Goto, N. (2001). Comparison of UDP-glucose:flavonoid 3-*O*-glucosyltransferase (*UFGT*) gene sequences between white grapes (*Vitis vinifera*) and their sports with red skin. *Plant Science* **160**, 543-550.

Kobayashi, S., Ishimaru, M., Hiraoka, K., Honda, C. (2002). *Myb*-related genes of the Kyoho grape (*Vitis labruscana*) regulate anthocyanin biosyntesis. *Planta* **215**, 924-933.

Koeppen, B.H. and Basson, D.S. (1965). The anthocyanin pigments of Barlinka grapes. *Phytochemistry* **5**, 183-187.

Kong, J.M., Chia, L.S., Goh, N.K., Chia, T.F., Brouillard, R. (2003). Analysis and biological activities of anthocyanins. *Phytochemistry* **64**, 923-933.

Kozjak, P., Korosec-Koruza, Z, Javornik, B. (2003). Characteristics of cv. Refosk (*Vitis vinifera* L.) by SSR markers. *Vitis* **42**, 83-86.

Kramer, J.H. (2004). Anthocyanosides of *Vaccinium myrtillus* (bilberry) for night vision – a systematic review of placebo-controlled trials. *Survey of Ophtalmolology* **49**, 618-618.

Kwon, S.H., Ahn, I.S., Kim, S.O, Kong, C.S., Chung, H.Y., *et al.* (2007). Anti-obesity and hypolipidemic effects of black soybean anthocyanins. *Journal of Medicinal Food* **10**, 552-56.

Landegren, U., Kaiser, R., Sanders, J., Hood, L. (1988). A ligase-mediated gene detection technique. *Science* **241**, 1077-1080.

Levadoux, L. (1956). Les populations sauvages et cultivées de *Vitis vinifera* L. *Annales de l'Amélioration des Plantes* **6**, 59-117.

Lewontin, R.C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49-67.

Li, C.C, Weeks, D.E., Chakravarti, A. (1993). Similarity of DNA finger-prints due to chance and relatedness. *Human Heredity* **43**, 45-52.

Lider, L.A., Goheen, A.C., Ferrari, N.L. (1975). Comparison between healthy and leafroll-affected grapevine planting stocks. *American Journal of Enology and Viticulture* **26**, 144-147.

Lijavetzky, D. Cabezas, J.A., Ibáñez, A., Rodríguez, V., Martínez-Zapater, J.M. (2007). High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* **8**, 424-435.

Lipsick, J.S. (1996). One billion years of MYB. *Oncogene* **13**, 223-235.

Lodhi, M.A., Daly, M.J., Ye, G.N., Weeden, N.F., Reisch, B.I. (1995). A molecular marker based linkage map of *Vitis*. *Genome* **38**, 786-794.

Long, A. D. and Langley, C. H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* **9**, 720-731.

Loiselle, B.A., Sork, V.L., Nason, J., Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* **82**, 1420-1425.

Loureiro, M.D., Martinez, M.C., Boursiquot, J.M., This, P. (1998). Molecular marker analysis of *Vitis vinifera* 'Albarino' and some similar grapevine cultivars. *Journal of the American Society for Horticultural Science* **5**, 842-848.

Lu, Y.P., Li, Z.S., Drozdowicz, Y.M., Hortensteiner, S., Martinoia, E., *et al.* (1998). AtMRP2, an *Arabidopsis* ATP binding cassette transporter able to transport glutathione S-conjugates and chlorophyll catabolites: Functional comparisons with AtMRP1. *Plant Cell* **10**, 267-282.

Ludwig, S.R., Habera, L.F., Dellaporta, S.L., Wessler, S.R. (1989). *Lc*, a member of the maize R-gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the Myc-homology region. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 7092-7096.

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753-1766.

Maccaferri, M., Sanguineti, M.C., Noli, E., Tuberosa, R. (2005). Population structure and long-range linkage disequilibrium in a durm wheat alite collection. *Molecular Breeding* **15**, 271-289.

Malik, M., Zhao, C., Schoene, N., Guisti, M.M., Moyer, M.P., *et al.* (2003). Anthocyanin-rich extract from *Aronia meloncarpa* E induces a cell cycle block in colon cancer but not normal colonic cells. *Nutrition and Cancer* **46**, 186-96.

Malosetti, M., Van Der Linden, C. G., Vosman, B., Van Reuwijk, F.A. (2007). A mixed-model approach to association mapping using pedogree information with an illustration of resistance to *Phytophora infestans* in potato. *Genetics* **175**, 879-889.

Maniatis, N., Collins, A., Xu, C.F., McCarthy, L.C., Hewett, D.R., *et al.* (2002). The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 2228-2233.

Marinova, K., Kleinscgmidt, K., Weissenbock, G., Klein, M. (2007). Flavonoid biosynthesis in barley primary leaves requires the presence of the vacuole and controls the activity of vacuolar flavonoid transport. *Plant Physiology* **144**, 432-444.

Markakis, P. (1982). *Anthocyanins as food colours.* Markakis, P., Ed. Academic Press: London, pp 261.

Marquart, L. (1835). *Die farben der blüthen, eine chemischphysiologische abhandlung.* Kessinger Publishing: Bonn, pp 96.

Marrs, K.A., Alfenito, M.R., Lloyd, A.M., Walbot, V. (1995). A glutathione S-transferase involved in vacuolar transfer encoded by the maize gene *Bronze-2. Nature* **375**, 397-400.

Martin, C. And Paz-Ares, J. (1997). MYB transcription factors in plants. *Trends in Genetics* **13**, 67-73.

Martinoia, E., Maeshima, M., Neuhaus, E. (2006). Vacuolar transporters and their essential role in plant metabolism. *Journal of Experimental Biology* **58**, 83-102.

Mateus, N., Pascual-Teresa, S.D., Rivas-Gonzalo, J.C., Santos-Buelga, C., Freitas, V.D. (2002). Structural diversity of anthocyanin-derived pigments in port wines. *Food Chemistry* **76**, 335-342.

Matsumoto, H., Nakamura, Y., Hirayama, M. Yoshiki, Y. Okubo, K. (2002). Antioxidant activity of black currant anthocyanin aglycons and their glycosides measured by chemiluminescence in a neutral pH

region and in human plasma. *Journal of Agriculture and Food Chemistry* **50**, 5034-37.

Matus, J.T., Aquea, F., Arce-Johnson, P. (2008). Analysis of the grape *MYB R2R3* subfamily reveals expanded wine quality-related clades and conserved gene structure organization across *Vitis* and *Arabidopsis* genomes. *BMC Plant Biology* **8**, 83-98.

Matus, J. T., Loyola, R., Vega, A., Peña-Neira, A., Bordeu, E., Arce-Johnson, P. and Alcalde, J. A. (2009). Post-veraison sunlight exposure induces MYB-mediated transcriptional regulation of anthocyanin and flavonol synthesis in berry skins of *Vitis vinifers. Journal of Experimental Botany* **60**, 853-867.

Matus, J.T., Poupin, M.J., Cañón, P., Bourdeu, E., Alcalde, J.A., *et al.* (2010). Isolation of WDR and bHLH genes related to flavonoid synthesis in grapevine (*Vitis vinifera* L.). *Plant Molecular Biology* **72**, 607-620.

Mazza, G. (1995). Anthocyanins in grapes and grape products. *Critical Reviews in Food Science and Nutrition* **35**, 341.

Mazza, G. and Miniati, E. (1993). Grapes. In *Anthocyanins in Fruits, Vegetables and Grains.* Mazza, G. and Miniati, E. Ed. CRC Press: Boca Raton, pp 149-99.

McGovern, P.E. (2003). *Ancient Wine: the search for the origins of viniculture*. Princeton University Press: Princeton, pp 400.

McGovern, P.E., Glusker, D.L., Exner, L.J., Voigt, M.M. (1996). Neolithic resonated wine. *Nature* **381**, 480-481.

McGovern, P.E. and Rudolf, H.M. (1996). The analytical and archaeological challenge of detecting ancient wine: two case studies from the ancient Near East. In *The origins and ancient history of wine.*

McGovern, P.E., Fleming, S.J., Katz, S.H., Eds. Gordon and Breach: New York, pp 57-67.

Moncada, X., Pelsy, F., Merdinoglu, D., Hinrichsen, P. (2006). Genetic diversity and geographical dispersal in grapevine clones revealed by microsatellite markers. *Genome* **49**, 1459-1472.

Moore, J.H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity* **56**, 73-82.

Moore, J.H., Gilbert, J.C., Tsai, C-T., Chiang, F-T., Holden, T., *et al.* (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology* **241**, 252-261.

Morais, H., Ramos, C., Forgács, E., Cserháti, T., Oliveira, J. (2002). Influence of storage conditions on the stability of monomeric anthocyanins studied by reversed-phase high performance liquid chromatography. *Journal of Chromatography B* **770**, 297-301.

Moreno, S., Martín, J.P., Ortiz, J.M. (1998). Inter-simple sequence repeats PCR for characterization of closely related grapevine germplasm. *Euphytica* **101**, 117-125.

Mori, K., Goto-Yamamoto, N., Kitayama, M., Hashizume, K. (2007). Effect of high temperature on anthocyanin composition and transcription of flavonoids hydroxylase genes in 'Pinot noir' grapes (*Vitis vinifera*). *Journal of Horticultural Science & Biotechnology* **82**, 199-206.

Mori, K., Sugaya, S., Gemma, H. (2005). Decreased anthocyanin biosynthesis in grape berries grown under elevated night temperature. *Scientia Horticulturae* **105**, 319-330.

Morris, R.W. and Kaplan, N.L. (2001). When is haplotype analysis advantageous for linkage-disequilibrium mapping? *American Journal of Human Genetics* **69**, 181-181.

Moser, C., Segala, C., Fontana, P., Salakhudtinov, I., Gotto, P., *et al.* (2005). Comparative analysis of expressed sequence tags from different organs of *Vitis vinifera* L. *Functional & Integrative Genomics* **5**, 208-217.

Mueller, L.A., Goodman, C.D., Silady, R.A., Walbot, V. (2000). *AN9*, a petunia Glutathione S-transferase required for anthocyanin sequestration, is a flavonoid-binding protein. *Plant Physiology* **123**, 1561-1570.

Mullins, M.G., Bouquet, A., Williams, L.E. (1992). *Biology of the Grapevine.* Cambridge University Press: Cambridge, pp 239.

Myles, S., Chia, J-M., Hurwitz, B., Simon, C., Zhong, *et al.* (2010). Rapid genomic characterization of the genus *Vitis.PLoS One* **5**, e8219-e8219.

Negrul, A.M. (1938). Evolucija kuljturnyx form vinograda. *Doklady Akademii nauk SSSR* **8**, 585-585.

Neilson-Jones, W. (1969). *Plant chimeras, 2nd edn*. Methuen: London, pp 123.

Peng, U.K., Reid, F.Y., Liao, K.E., Schlosser, N., Lijavetzky, D. *et al.* (2007). Generation of ESTs in *Vitis vinifera* wine grape (Cabernet Sauvignon) and table grape (Muscat Hamburg) and discovery of new candidate genes with potential roles in berry development. *Gene* **402**, 40-50.

Nordborg, M. and Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**, 83-90.

Ocete, R., López Martínez, M.A., Pérez Izquierdo, A., Del Tío, M.R. (1999). *Las poblaciones españolas de vid Silvestre.* Monografias INIA. Ministerio de Agricoltura, pesca y alimentación: Madrid.

Organisation Internationale da la Vigne et du Vin (OIV). (2009). *International list of vine varieties and their synonyms.* OIV: Paris.

Ojeda, H., Andary, C., Kraeva, E., Carbonneau, A, Deloire, A. (2002). Influence of pre- and postveraison water deficit on synthesis and concentration of skin phenolic compounds during berry growth of *Vitis vinifera* cv. Shiraz. *American Journal of Enology and Viticulture* **53**, 261-267.

Olmo, H.P. (1976). Grapes. In *Evolution of Crop Plants*. Simmonds, N.W., Ed. Longman: London, pp 294-298.

Olmo, H.P. (1995). The origin and domestication of vinifera grape. In *The origin and ancient history of wine.* McGovern, P., Fleming, S.J., Katz, S.H., Eds. Gordon and Breach: Luxembourg, pp 31-43.

Ortega-Regules, A., Romero-Cascales, I., López-Roca, J.M., Ros-García, J.M., Gómez-Plaza, E. (2006). Anthocyanin fingerprint of grapes: environmental and genetic variations. *Journal of the Science of Food and Agriculture* **86**, 1460-1467.

Orts, M.L.D., Martinez-Cutillas, A., Roca, J.M.L., Perez-Prieto, L.J., Gomez-Plaza, E. (2005). Effect of deficit irrigation on anthocyanin content of Monastrell grapes and wines. *Journal International des Sciences de la Vigne et du Vin* **39**, 47-55.

Payne, C.T., Zhang, F., Lloyd, A.M. (2000). *GL3* encodes a bHLH protein that regulates trichome development in *Arabidopsis* through intercation with *GL1* and *TTG1*. *Genetics* **156**, 1349-1362.

Pelsy, F. (2010). Molecular and cellular mechanisms of diversity within grapevine cultivars. *Heredity* **104**, 331-340.

Pelsy, F., Hocquigny, S., Moncada, X., Barbeau, G., Forget, D., *et al.* (2010). An extensive study of the genetic diversity within seven French wine grape variety collections. *Theoretical and Applied Genetics* **120**, 1219-1231.

Pérez-Magarino, S. and Gonzalez-San José, M.L. (2004). Evolution of flavonols, anthocyanins, and their derivatives during the aging of red wines elaborated from grapes harvested at different stages of ripening. *Journal of Agriculture and Food Chemistry* **52**, 1181-1189.

Pilati, S., Perazolli, M., Malossini, A., Cestaro, A., Dematte, L., *et al.* (2007). Genome-wide transcriptional analysis of grapevine berry ripening reveals a set of genes similarly modulated during three seasons and the occurrence of an oxidative burst at veraison. *BMC Genomics* **8**, 428-450.

Pomar, F., Novo, M., Masa, A. (2005). Varietal differences among the anthocyanin profiles of 50 red table grape cultivars studied by high performance liquid chromatography. *Journal of Chromatography* **1094**, 34-41.

Poustka, F., Irani, N.G., Feller, A., Lu, Y., Pourcel, L., *et al.* (2007). A trafficking pathway for anthocyanins overlaps with the endoplasmic reticulum-to-vacuole protein-sorting route in Arabidopsis and contributes to the formation of vacuolar inclusions. *Plant Physiology* **145**, 1323-1335.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., *et al.* (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909.

Prior, R.L., Wu, X., Gu, L., Hager, T.J., Hager, A., *et al.* (2008). Whole berries versus berry anthocyanins: interactions with dietary fat levels

in the C57BL/6J mouse model of obesity. *Journal of Agriculture and Food Chemistry* **56**, 647-53.

Pritchard, J.K., and Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **6**, 220-228.

Pritchard, J.K., Stephens, M., Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179-1189.

Quattrochio, F., Wing, J.F., Leppen, H.T.C., Mol, J.N.M., Koes, R.E. (1993). Regulatory genes controlling anthocyanin pigmentation are functionally conserved among plant species and have distinct setes of target genes. *The Plant Cell* **5**, 1497-1512.

Queller, D.C. and Goodnight, K.F. (1989). Estimating relatedness using genetic markers. *Evolution* **43**, 258-275.

Ramirez-Tortosa, C., Andersen, Ø.M., Gardner, P.T., Morrice, P.C., Wood, S.G., *et al.* (2001). Anthocyanin-rich extract decreases indices of lipid peroxidation and DNA damage in vitamine E-depleted rats. *Free Radical Biology and Medicine* **31**, 1033-37.

Ramsay, N.A., Walker, A.R., Mooney, M., Gray, J.C. (2003). Two basic-helix-loop-helix genes (MYC-146 and GL3) from Arabidopsis can activate anthocyanin biosynthesis in a white-flowered Matthiola incana mutant. *Plant Molecular Biology* **52**, 679-688.

Regner, F., Stadlbauer, A., Eisenheld, C., Kaserer, H. (2000). Genetic relationships among Pinots and related cultivars. *American Journal of Enology and Viticulture* **51**, 7-14.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., *et al.* (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Science of the USA* **98**, 11479-11484.

Revilla, I., Pérez-Magariño, S., González-SanJosé, M.L., Beltrán, S.J. (1999). Identification of anthocyanin derivatives in grape skin extracts and red wines by liquid chromatography with diode array and mass spectrometric detection. *Journal of Chromatography A* **847**, 83-90.

Rhoné, B., Rquin, A.-L., Goldringer, I. (2007). Strong linkage disequilibrium near the selected *Yr17* resistance gene in a wheat experimental population. *Theoretical and Applied Genetics* **114**, 787-802.

Riaz, S., Garrison, K.E., Dangl, G.S., Boursiquot, J.M. and Meredith, C.P. (2002). Genetic divergence and chimerism within ancient asesually propagated wine grape cultivars. *Journal of the American Society for Horticultural Science* **127**, 508-514.

Riaz, S., Dangl, G.S., Edwards, K.J., Meredith, C.J. (2004). A microsatellite marker based framework linkage map of *Vitis vinifera* L. *Theoretical and Applied Genetics* **108**, 864-872.

Ribéreau-Gayon, P. (1959). Recherches sur les anthocyanes des végétaux. Application au genre *Vitis.* Ph.D. Thesis, University of Bourdeux.

Ribéreau-Gayon, P. (1964). Les composés phénoliques du raisin et du vin. II. Les flavonosides et les anthocyanosides. *Annales de Physiologie Végétale* **6**, 211-242.

Ribéreau-Gayon, P. (1982). The anthocyanins of grapes and wines. In *Anthocyanins as food colors*. Markakis, P., Ed. Academic Press: New York, pp 209-242.

Ritland, K. (1996). A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* **50**, 1062-1073.

Robbins, M.P., Paolocci, F., Hughes, J.W., Turchetti, V., Allison, G., *et al.* (2003). Sn, a maize bHLH gene, modulates anthocyanin and condensed tannins pathways in Lotus corniculatus. *Journal of Experimental Botany* **54**, 239-248.

Roby, G., Harbertson, J., Adams, D., Matthews, M. (2004). Berry size and vine water deficits as factors in winegrape composition: anthocyanins and tannins. *Australian Journal of Grape Wine Research* **10**, 100-107.

Ryan, J.M. and Revilla, E. (2003). Anthocyanin composition of Cabernet Sauvignon and Tempranillo grapes at different stages of ripening. *Journal of Agriccultre and Food Chemistry* **51**, 3372-3378.

Sainz, M.B., Grotewold, E., Chandler, V.L. (1997). Evidence for direct activation of an anthocyanin promoter by the maize C1 protein and comparison of DNA binding by related Myb domain proteins. *Plant Cell* **9**, 611-625.

Saito, K., Yamazaki, M. (2002). Biochemistry and molecular biology of the late-stage of biosynthesis of anthocyanin: lessons from *Perilla frutescens* as a model plant. *New Phytologist* **155**: 9-23.

Salmaso, M., Faes, G., Segala, C., Stefanini, M., Salakhutdinov, L., *et al*. (2004). Genome diversity and gene haplotypes in the grapevine (*Vitis vinifera* L.), as revealed by single nucleotide polymorphisms. *Molecular Breeding* **14**, 385–395.

Salmaso, M., Malacarne, G., Troggio, M., Faes, G., Stefanini, M., *et al.* A grapevine (Vitis vinifera L.) genetic map integrating the position of

139 expressed genes. *Theoretical and Applied Genetics* **116**, 1129-1143.

Sarry, J.E., Sommerer, N., Sauvage, F.X., Bergoin, A., Rossignol, M., *et al.* (2004). Grape berry biochemistry revisited upon proteomic analysis of the mesocarp. *Proteomics* **4**, 201-215.

Saslowsky, D. and Winkel-Shirley, B. (2001). Localization of flavonoids enzymes in Arabidopsis roots. *Plant Journal* **27**, 37-48.

Schellenbaum, P., Mohler, V., Wenzel, G., Walker, B. (2008). Variation in DNA methylation of grapevine somaclones (*Vitis vinifera* L.). *BMC Plant Biology* **8**, 78-98.

Schmidt, E.D.L., Dejong, A.J., Devries, S.C. (1994). Signal molecules involved in plant embryogenesis. *Plant Molecular Biology* **26**, 1305-1313.

Schwinn, K., Venail, J., Shang, Y.J., Mackay, S., Alm, V., *et al.* (2006). A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus Antirrhinum. *The Plant Cell* **18**, 831-851.

Scott, K.D., Ablett, E.M., Lee, L.S., Henry, R.J. (2000). AFLP markers distinguishing an early mutant of Flame seedless grape. *Euphytica* **113**, 243-247.

Sensi, E., Vignani, R., Rohde, W., Biricolti, S. (1996). Characterization of genetic biodiversity with *Vitis vinifera* L. Sangiovese and Colorino genotypes by AFLP and ISTR DNA marker technology. *Vitis* **35**, 183-188.

Silvestroni, O., DiPietro, D., Intrieri, C., Vignani, R., Filippetti, I., *et al.* (1997). *Vitis* **36**, 147-150.

Sparvoli, F., Martin, C., Scienza, A., Gavazzi, G. and Tonelli, C. (1994). Cloning and molecular analysis of structural genes involved in

flavonoid and stilbene biosynthesis in grape (*Vitis vinifera* L.). *Plant Molecular Biology* **24**, 743-755.

Spelt, C., Quattrocchio, F., Mol, J.N.M., Koes, R. (2000). *anthocyanin1* of petunia encodes a basic helix-loop-helix protein that directly activates transcription of structural genes. *Plant Cell* **12**, 1619-1631.

Springob, K., Nakajima, J., Yamazaki, M., Saito, K. (2003). Recent advances in the biosynthesis and accumulation of anthocyanin. *Natural Product Reports* **20**, 288-303.

Stintzing, F.C., Stintzing, A.S., Carle, R., Frei, B., Wrolstad, R.E. (2002). Color and antioxidant properties of cyanidin-based anthocyanin pigments. *Journal of Agricultural and Food Chemistry* **50**, 6172-6181.

Stobiecki, M. (2000). Application of mass spectrometry for identification and structural studies of flavonoid glycosides. *Phytochemistry* **54**, 237.

Stuber, C.W., Polacco, M., Lynn, M. (1999). Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield. *Crop Science* **39**, 1571-1583.

Sugimoto, E., Igarashi, K., Kubo, K., Molyneux, J., Kubomura, K. (2003). Protective effects of boysenberry anthocyanins on oxidative stress in diabetic rats. *Food Science and Technology Research* **9**, 345-49.

Sun, B.S., Spranger, I., Yang, J.Y., Leandro, C., Guo, L., *et al.* (2009). Red wine phenolic complexes and their *in vitro* activity. *Journal of Agricultural and Food Chemistry* **57**, 8623-8627.

Tattersall, E.A.R., Grimplet, J., Deluc, L., Wheatley, M.D., Vincent, D. (2007). Transcript abundance profiles reveal larger and more complex responses of grapevine to chilling compared to osmotic and salinity stress. *Functional & Integrative Genomics* **7**, 317-333.

Tedesco, I., Russo, L.G., Nazzaro, F., Russo, M., Palumbo, R. (2001). Antioxidant effect of red wine anthocyanins in normal and catalase-inactive human erythrocytes. *Journal of Nutrition and Biochemistry* **12**, 505-11.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F., *et al.* (2001). Paterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *Mays* L.). *Proceedings of the. National Academy of Science of the USA* **98**, 9161-9166.

Terrier, N., Glissant, D., Grimplet, J., Barrieu, F., Abbal, P., *et al.* (2005). Isogene specific oligo arrays reveal multifaceted changes in gene expression during grape berry (*Vitis vinifera* L.) development. *Planta* **222**, 832-847.

Terrier, N., Torregrossa, L., Ageorges, A., Vialet, S., Verriès, C., *et al.* (2009). Ectopic expression of *VvMybPA2* promotes proanthocyanidin biosynthesis in grapevine and suggests sdditional targets in the pathway. *Plant Physiology* **149**, 1028-1041.

This, P., Lacombe, T., Cadle-Davidson, M., Owens, C. (2007). Wine grape (*Vitis vinifera* L.) color associates with allelic variation in the domestication gene *VvmybA1*. *Theoretical and Applied Genetics* **114**, 723-730.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., *et al.* (2001). *Dwarf8* polym orphisms associate with variation in flowering time. *Nature Genetics* **28**, 286-289.

Tomazic, I. and Korosec-Koruza, Z. (2003). Validity of phyllometric parameters used to differentiate local *Vitis vinifera* L. cultivars. *Genetic Resources and Crop Evolution* **50**, 773-778.

Troggio, M., Malacarne, G., Coppola, G., Segala, C., Cartwright, D., *et al.* (2007). A dense sigle-nucleotide polymorphism-based genetic

linkage map of grapevine (*Vitis vinifera* L.) anchoring pinot noir bacterial artificial chromosome contigs. *Genetics* **176**, 2637-2650.

Tsuda, T., Horio, F., Osawa, T. (2000). The role of anthocyanins as an antioxidant under oxidative stress in rats. *Biofactors* **13**, 133-39.

Tsuda, T., Horio, F., Uchida, K., Aoki, H., Osawa, T. (2003). Dietary cyanidin 3-*O*-*β*-D-glucoside-rich purple corn colour prevents obesity and ameliorates hyperglycemia in mice. *The Journal of Nutrition* **133**, 2125-2130.

Tsuda, T., Watanabe, M., Ohshima, K., Norinobu, S., Choi, S-W. *et al.* (2004). Antioxidative activity of the anthocyanin pigments cyanidin 3-O-beta-D-glucoside and cyanidin. *Journal of Agriculture and Food Chemistry* **42**, 2407-10.

Velasco, R., Zharkikh, A., Troggio, M. Cartwright, D.A., Cestaro, A. *et al.* (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* **12**, 1-18.

Verpoorte, R. (2000). Secondary Metabolism. In Metabolic Engineering of Plant Secondary Metabolism. Verpoorte, R, Alfermann, A.W., Eds. Springer – Verlag. Online version available at: http://knovel.com/web/portal/browse/display?_EXT_KNOVEL_DISPLAY_bookid=1153&VerticalID=0.

Verrier, P.J., Bird, D., Burla, B., Dassa, E., Forestier, C., *et al.* (2008). Plant ABC proteins – a unified nomenclature and updated inventory. *Trends in Plant Science* **13**, 151-159.

Vetten, N., Quattrochio, F., Mol, J., Koes, R. (1997). The *an11* locus controlling flower pigmentation in petunia encodes a novel WD-repeat protein conserved in yeast, plants, and animals. *Genes & Development* **11**, 1422-1434.

Vezzulli, S., Troggio, M., Coppola, G., Jermakow, A., Cartwright, D. *et al.* (2008). A reference integrated map for cultivated grapevine (*Vitis vinifera* L.) from three crosses, based on 283 SSR and 501 SNP-based markers. *Theoretical and Applied Genetics* **117**, 499-511.

Vignal, A., Milan, D., SanCristobal, M., Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection and Evolution* **3**, 275-305.

Vignani, R., Scali, M., Masi, E., Cresti, M. (2002). Genomic variability in *Vitis vinifera* L. "Sangiovese" assessed by microsatellite and non-radioactive AFLP test. *Electronic Journal of Biotechnology* **5**, 1-11.

Vincent, D., Ergül, A., Bohlman, M.C., Tattersall, E.A., Tillett, R.L., *et al.* (2007). Proteomic analysis reveals differences between *Vitis vinifera* L. Cv. Chardonnay and cv. Cabernet Chauvignon and their responses to water deficit and salinity. *Journal of Experimental Botany* **58**, 1873-1892.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nulceotide polymorphisms in the human genome. *Science* **280**, 1077-1082.

Walker, A.R., Davison, P.A., Bolognesi-Winfield, A.C., James, C.M., Srinivasan, N., *et al.* (1999). The *TRANSPARENT TESTA GLABRA1* locus, which regulates trichomes differentiation and anthocyanin biosynthesis in Arabidopsis, encodes a WD40 repeat protein. *The Plant Cell* **11**, 1337-1349.

Walker, A.R., Lee, E., Bogs, J., McDavid, D.A.J., Thomas, M.R., *et al.* (2007) White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant Journal* **49**: 772-785.

Walker, A.R., Lee, E., Robinson, S.P. (2006). Two new grape cultivars, bud sports of cabernet Sauvignon bearing pale-coloured berries, are the result of deletion of two regulatory genes of the berry colour locus. *Plant Molecular Biology* **62**, 623-635.

Walter, B. and Martelli, G.P. (1998). Considerations on grapevine selection and certification. *Vitis* **37**, 87-90.

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics* **160**, 1203-1215.

Wang, H., Cao, G., Prior, R.L. (1997). Oxygen radical absorbing capacity of anthocyanins. *Journal of Agriculture and Food Chemistry* **45**, 304-9.

Waters, D.L.E., Holton, T.A., Ablett, E.M., Lee, L.S., Henry, R.J. (2005). cDNA microarray analysis of developing grape (*Vitis vinifera* cv. Shiraz) berry skin. *Functional and Integrated Genomics* **5**, 40-58.

Weiss, K.M. and Terwilliger, J.D. (2000). How many diseases does it take to map a gene with SNPs? *Nature Genetics* **26**, 151-157.

Whitehead, T.P., Robinson, D., Allaway, S., Syma, J., Hale, A. (1995). Effect of red wine ingestion on the antioxidant capacity of serum. *Clinical Chemistry* **41**, 32-35.

Winkel, B.S.J. (2004). Metabolic channelling in plants. *Annual Review of Plant Biology* **55**, 85-107.

Winkel-Shirley, B. (1999). Evidence of enzyme compexes in the phenylpropanoid and flavonoid pathways. *Plant Physiology* **107**, 142-149.

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colourful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiology* **126**, 485-493.

Welcome Trust Case Control Consortium. (2007). Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature* **447**, 661-678.

Wulf, L.W. and Nagel, C.W. (1978). High pressure liquid chromatographic separation of anthocyanins of *V. Vinifera*. *American Journal of Enology and Viticulture* **29**, 42-49.

Yakushiji, H., Kobayashi, S., Goto-Yamamoto, N., Jeong, S.T., Sueta, T., *et al.* (2006). A skin colour mutation of grapevine, from black-skinned Pinot Blanc, is caused by deletion of the functional VvmybA1 allele. *Bioscience Biotechnology and Biochemistry* **70**, 1506-1508.

Yamane, T., Jeong, S.T., Goto-Yamamoto, N., Noshita, Y., Kobayashi, S. (2006). Effects of temperature on anthocyanin biosynthesis in grape berry skins. *American Journal of Enology and Viticulture* **57**, 54-59.

Yang, Y., Houle, A.M., Letendre, J., Richter, A. (2008). *RET* Gly691Ser mutation is associated with primary vesicoureteral reflux in the French-Canadian population from Quebec. *Human Mutation* **29**, 695-702.

Yazaki, K. (2005). Transporters of secondary metabolites *Current Opinion in Plant Biology* **8**, 301-307.

Yi, W., Fischer, J., Akoh, C.C. (2005). Study of anticancer activities of muscadine grape phenolics in vitro. *Journal of Agriculture and Food Chemistry* **53**, 8804-12.

Yokotsuka, K., Nagao, A., Nakazawa, K., Sato, M. (1999). Changes in anthocyanins in berry skin of Merlot and Cabernet Sauvignon grapes grown in two soils modified with limestone or oyster shell versus a nature soil over two years. *American Journal of Enology and Viticulture* **50**, 1-12.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Masanori, Y., *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208.

Zhang, Y., Seeram, N.P., Lee, R., Feng, L., Heber, D. (2008). Isolation and identification of strawberry phenolics with antioxidant and human cancer cell antiproliferative properties. *Journal of Agriculture and Food Chemistry* **56**, 670-75.

Zhang, H., Wang, L., Deroles, S., Bennett, R., Davies, K. (2006). New insight into the structures and formation of anthocyanin vacuolar inclusions in flower petals. *BMC Plant Biology* **6**, 29-43.

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., *et al.* (2007). An Arabidopsis example of association mapping in structured samples. *PLoS Genetics* **3**, 71-82.

Zhao, C., Giusti, M.M., Malik, M., Moyer, M.P., Magnuson, B.A. (2004). Effects of commercial anthocyanin-rich extracts on colonic cancer and nontumorigenic colonic cell growth. *Journal of Agriculture and Food Chemistry* **52**, 6122-28.

Zhu, C., Gore, M., Bucker, E.S., Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* **1**, 5-20.

Ziegler, A. (2009). Genome-wide association studies: quality control and population-based measures. *Genetic Epidemiology* **33**, s45-s50.

Zielenski, J. and Tsui, L. (1995). Cystic fibrosis - genotypic and phenotypic variations. *Annual Review of Genetics* **29,** 777–807.

Ziv, E. and Burchard, E.G. (2003). Human population structure and genetic association studies. *Pharmacogenomics* **4**, 431-441.

Zohary, D. (1996). The mode of domestication of the founder crops of the Southwest Asian agriculture. In *The origin and spread of*

agriculture and pastoralism in Eurasia. Harris, D.R., Ed. University College London Press: London, pp 142-158.

Zohary, D. (2004). Unconscious selection and the evolution of domesticated plants. *Economic Botany* **58**, 5-10.

Zohary, D. and Hopf, M. (2000). "Domestication of plants in the Old World", Third Ed. Oxford University Press, New York.

Zohary, D. and Hopf, M. (1993). Domestications of plants in the old world. Clarendon Press: Oxford, pp 150.

CHAPTER II

SEQUENCE VARIATION AND DIFFERENTIAL GENE EXPRESSION UNDELYING BERRY COLOUR VARIATION AMONG *VITIS VINIFERA* L. CLONES

A shorter version of this chapter will be shortly submitted to *Vitis*:

Cardoso, S., Lijavetzky, D., Eiras Dias, J., Zapater, J., Fevereiro, P. (To submit). Sequence variation and differential gene expression underlying berry colour variation among *Vitis vinifera* L. clones.

Contributions to this chapter:

- Designed the study: Cardoso, S., Eiras Dias, J.E., Fevereiro, P., Zapater, J.
- Performed experiments: Cardoso, S.
- Analysed data: Cardoso, S., Lijavetzky, D.

Summary

Vegetative propagation has a very important role in grapevine cultivars. The obtained plants are called “clones” and are usually genetically identical. However, phenotypic variation is often observed among clones of *Vitis vinifera* L. Berry colour is one of the traits where this is observed. Somatic mutations are usually one of the causes of this diversity. However, in many cases these variants have not yet been identified.

In this study the presence of DNA sequence variation among clones of two cultivars (Aragonez and Negra Mole) was investigated. Differential gene expression was also explored among clones of Aragonez cultivar. The results showed some variation in the DNA sequence level between the two cultivars Aragonez and Negra Mole but not between clones of the same cultivar in the studied candidate genes. Concerning gene expression, the differences identified were very subtle. A total of 106 probesets (104 genes) were differentially expressed ($P < 0.01$, not corrected for multiple testing). Focusing on groups of interest for the colour phenotype, such as flavonoid metabolism and transcription factors, and on probesets significant for more than one statistical test, a group of 24 genes were identified ($P < 0.05$, not corrected for multiple testing). This included two genes involved in the flavonoid metabolism, coding enzymes related with the glucosylation of flavonoids, an important step in anthocyanin biosynthesis. Transcription factors showing differential expression included members of the Myb and Myc families known to be important regulators of the anthocyanin biosynthesis pathway. Genes of the transcription families zinc finger, WRKY and homeobox domain, showed differential expression as well and have previously been found involved in proanthocyanidin and

anthocyanin regulation. Differential expression was also observed for members of other transcription factor families such as DOF, GRAS, YABBY, basic-leucine zipper, pathogenesis-related and plant homeodomain finger which have no previous indication of relevance for anthocyanin or flavonoids metabolism.

Despite the risks of false positives, these results indicate possible genes of interest for further studies on gene expression, genetic association and functional assays.

1. Introduction

Several processes are thought to have originated the available diversity of grapevine cultivars available today. These processes include multiple domestication events of *Vitis vinifera* subsp. *sylvestris* (Arroyo-Garcia *et al.,* 2006), crosses between wild plants and domesticated varieties, spontaneous crosses between cultivated varieties and controlled breeding programs (Pelsy, 2010).

Since domestication, vegetative propagation has been a widespread practice to fix desired traits. The plants obtained by this method are said to be clones. Usually, clones within a cultivar are identical, but sometimes different phenotypes are identified. Differences between clones have been pointed as a consequence of infection by phytopathological agents (Walter and Martelli, 1998), epigenetic modifications in response to environmental factors (Kaeppler *et al.*, 2000; Schellenbaum, 2008) and mutations (Hartman *et al.,* 1997). Somatic mutations may accumulate over time, especially among older cultivars where pruning and vegetative propagation has occurred many times, separating mutant from wild-type cells. Somatic mutations may lead to the formation of chimeras, and be further propagated vegetative and/or sexually depending of which meristematic layer is affected (Franks *et al.*, 2002). Chimeras and mutation causing phenotypic changes have been described for several grapevine cultivars, as for example Pinot, Cabernet Sauvignon and Chardonnay (Bertsch *et al.*, 2005; Boss, *et al.,* 1996; Crespan, 2004; Fernandez *et al.,* 2006; Franks *et al.,* 2002; Hocquigny *et al.,* 2004; Moncada *et al.*, 2006; This, *et al.,* 2007). Recently, also Carcamo *et al.* (2010) identified a chimerical state in Tempranillo, synonym of Aragonez in Portugal (OIV, 2009). Nevertheless, only one

clone was found with this genotype, remaining to be explained the phenotypic variation among the other 27 clones studied.

Clonal selection is an important means of grape quality improvement. Clonal characterisation is therefore of utmost importance (Moreno *et al.*, 1998). Ampelographic methods have been traditionally used to discriminate clones but with insufficient accuracy (Imazio *et al.*, 2002). As a consequence several attempts have been made to distinguish clones using molecular markers. ISSR and RAPD markers were not successfully on clone distinction (Moreno *et al.*, 1997; Loureiro *et al.*, 1998). Successful clonal discrimination has been performed with AFLP and SSR markers (Baneh *et al.*, 2009; Cervera *et al.*, 1998, 2000, 2001; Imazio *et al.*, 2002; Kozjak *et al.*, 2003; Moncada *et al.*, 2006; Regner *et al.*, 2000; Scott *et al.*, 2000; Sensi *et al.*, 1996). Despite the successful cases with SSR markers, several authors have failed to distinguish clones with these markers (Baneh *et al.*, 2009; Faria *et al.*, 2004; Imazio *et al.*, 2002; Loureiro *et al.*, 1998). Faria *et al.* (2004) have used a method based on polymorphisms on the *stilbene synthase* (*StSy*)*–chalcone synthase* (*CHS*) 5'untranslated genomic regions (*StSy–CHS* markers) which succeeded on clonal discrimination. Also, Carcamo *et al.* (2010), was able to distinguish between clones using retrotransposon based markers.

Over the last years genomic resources for *Vitis vinifera* L. increased considerably. Availability of expressed sequence tags (ESTs) databases (da Silva *et al.*, 2005; Peng *et al.*, 2007), the whole genome sequence (Jaillon *et al.*, 2007; Velasco *et al.*, 2007) and integrated genetic maps (Vezzulli *et al.*, 2008) has increased the information on gene annotation and also made possible large scale gene expression studies. The access to all these genomic resources on this species provides means to search for a better understanding of the features underlying phenotypic variation

among clones and to find tools for clonal characterisation at several levels, namely DNA sequence variation and differential gene expression. Single Nucleotide Polymorphisms (SNPs) are a very attractive kind of genetic markers as they are the most frequent type of genetic polymorphism and highly amenable for automation. At intraspecific level SNPs have been observed to occur at high frequencies in grapevine (Lijavetzky *et al.*, 2007; Velasco *et al.*, 2007). Accumulation of somatic mutations among clones may be tested using these markers.

The colour of grape berries is one of the most important traits of grapevine. Anthocyanins are the natural pigments that confer colour to grapes and also contribute to other organoleptic properties, playing an important role in grape and wine marketing. Also the antioxidant properties of anthocyanins and their benefits for human health have raised the interest in its study (Giusti and Wrolstad, 2003).

Anthocyanins biosynthetic pathway has been well characterised in petunia, snapdragon and maize (Martin and Gerats, 1993). The first part included in the phenylpropanoid pathway consists on the conversion of phenylalanine into 4-coumaroyl Co-A. The second part is included in the flavonoids pathway and converts 4-coumaroyl Co-A into anthocyanins. The genes coding the enzymes involved in this pathway have been characterised in different plant species (Dooner and Robbins, 1991; Holton and Cornish, 1995). In grapevine, Sparvoli *et al.* (1994) have cloned the cDNA of several genes encoding enzymes involved on the biosynthetic pathway of anthocyanins (*PAL, CHS, CHI, F3H, DFR, LDOX, UFGT*).

Anthocyanin biosynthesis has been shown to be regulated in grapevine by genes belonging to two major families, *Myb* and *β helix-loop-helix* (*bHLH*) (also known as *Myc*). These genes have been shown

to interact with each other and to affect the expression of genes involved on the biosynthetic pathway of anthocyanins (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; Matus *et al.*, 2009, 2010; Terrier *et al.*, 2009; This *et al.*, 2007). In other species, other protein families, such as tryptophan-aspartic acid repeat (WDR or WD40 repeat) (Vetten *et al.*, 1997) and WRKY transcription factors (Johnson *et al.*, 2002) have also been shown to play a role in anthocyanin regulation. Recently, Matus *et al.* (2010) found a correlation between the expression of *WDR1* and anthocyanin accumulation in *Vitis vinifera* L. as well.

The biosynthetic step mediated by UFGT (UDP-glucose: flavonoid 3-*O*-glucosyltransferase) has been shown to be controlled by *MybA1* and *MybA2* (Ageorges *et al.*, 2006; Kobayashi *et al.*, 2002; Walker *et al.*, 2007). Absence of anthocyanins in grape berries skin has been shown to be determined by the homozygous presence of a *MybA1* allele with a retrotransposon insertion (*Gret1*) in the gene promoter region (Fournier-Level *et al.*, 2009; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; This *et al.*, 2007). However, variation within coloured cultivars has not been so well understood yet. This *et al.* (2007) identified four polymorphisms in *MybA1* associated with pink/red cultivars and recently, Fournier-Level *et al.* (2009) identified other four polymorphisms in *MybA1*, *MybA2* and *MybA3* accounting for 23 % of colour variance. Transient expression experiments have shown *Myb5a*, *Myb5b*, *MybPA1* and *MybPA2* to affect the expression of genes coding enzymes which catalyze early steps of the anthocyanins biosynthetic pathway by promoter activation (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Matus *et al.*, 2009; Terrier *et al.*, 2009).

Environmental effects, such as light and temperature, throughout grape berry ripening significantly influence flavonoids production and accumulation. Flavonoid biosynthetic genes and *MybA* genes expression were shown to be affected by these factors (Downey *et al.*, 2006; Jeong *et al.*, 2004; Matus *et al.*, 2009; Mori *et al*., 2005; Yamane *et al.*, 2007). Also biotic factors, such as viruses, have been reported to affect anthocyanins content of grape berries, namely grapevine leafroll associated viruses (GLRaVs) (Brar *et al.*, 2008; Guidoni *et al.*, 1997; Lee and Martin, 2009).

The genetics underlying the variation of grape colour between clones of the same cultivar is very much unknown. Studies on clonal variation have focused mostly on the categorical variation of colour. This variation has been found to be a consequence of chimerism in some cases. Variation of anthocyanin concentration in berries skin, SNP marker variation and differential gene expression have not been explored so far.

The aim of this chapter is to investigate the variation underlying phenotypic differences in total skin anthocyanin concentration between *Vitis vinifera* L. clones. Two levels of variation were investigated, DNA sequence variation and differential gene expression.

2. Material and methods

2.1. Variation at DNA sequence level

2.1.1. Plant material

Two grapevine cultivars were used to study intracultivar variation, Negra Mole and Aragonez. These cultivars were chosen for this study considering the phenotypic characterisation of the clones total skin anthocyanin (TSA) concentration provided by Instituto Superior de

Agronomia (ISA) (Antero Martins, pers. comm.). Clones of these cultivars were grown on the same field divided on 3 blocks. Each block had three plants of each clone. TSA in berries was assessed by collecting 20 berries per plant. Measurements were taken for Aragonez cultivar over two years, 2002 and 2004 and for Negra Mole over four years, 1996, 1997, 1998 and 2000. As shown on Table II-1, these cultivars showed contrasting average values and a wide range of TSA concentration. Cultivar Aragonez shows on average high levels of TSA concentration while Negra Mole shows very low levels. Young leaves were collected and stored at -80ºC for DNA extraction. A total of 40 clones of Negra Mole and 50 of Aragonez were collected (Appendix 1).

Table II-1 Phenotypic data for the two cultivars used on DNA sequence variation search.

	Aragonez (skin anthocyanin concentration mg/l)	**Negra Mole** (skin anthocyanin concentration mg/l)
Mean	679.3	98.5
Median	688.6	96.2
Standard deviation	162.8	20.0
Maximum	993.7	149.9
Minimum	394.2	69.5

Data provided by Instituto Superior de Agronomia (Antero Martins, pers. comm.).

2.1.2. Selected genes

The Institute for Genomic Research (TIGR) database, based on Expressed Sequence Transcripts (ESTs), was searched for the genes involved on the anthocyanin biosynthetic pathway. This search was performed before the release of the Grapevine Genomic Sequence (Jaillon *et al.*, 2007). Genes with complete consensus sequences and reported SNPs were selected. These criteria were met for genes encoding chalcone synthase (*CHS*), chalcone isomerase (*CHI*), dihydroflavonol 4-reductase (*DFR*) and leucoanthocyanidin dioxygenase (*LDOX*) as shown on Table II-2.

Table II-2 List of genes used to search for sequence variation.

Chr.	Scaffold[1]	Gene ID[1]	Code	Coded protein name	Function
14	9	GSVIVT00037969001	*CHS_B*	Chalcone synthase family	Involved in anthocyanins biosynthetic pathway. Catalyzes the condensation of one molecule of 4-coumaroyl CoA and three molecules of malonyl-CoA into a naringenin chalcone.
13	48	GSVIVT00029513001	*CHI*	Chalcone isomerase	Involved in anthocyanins biosynthetic pathway. Catalyzes the isomerisation of the naringenin chalcone into a naringenin flavanone.
18	1	GSVIVT00014584001	*DFR*	Dihydroflavonol reductase	Involved in anthocyanins biosynthetic pathway. Catalyzes the reduction of the dihydroflavonols into leucoanthocyanidins.
2	112	GSVIVT00001063001	*LDOX*	Leucoanthocyanidin dioxygenase	Involved in anthocyanins biosynthetic pathway. Catalyzes the conversion of leucoanthocyanidins into anthocyanidins.

[1]Scaffold and gene IDs are according to Genoscope, sequencing version 8x coverage. *Chr.* Stands for chromosome.

2.1.3. PCR and sequencing

DNA was extracted from *Vitis vinifera* L. young leaves with approximately 100mg fresh weight. Quiagen Mini Kit (Quiagen Inc, Hilden, Germany) was used following mortar and pestle grinding with sterile quartz sand. Quantification was done spectrophotometricaly.

Primers were designed using Primer3 software (Rozen and Skaletsky, 2000) to amplify DNA fragments including regions where SNPs were reported to occur on the genes of interest according to the Institute for Genomic Research database. Six primer pairs were tested (Appendix 2). Amplifications were performed in a 25 μl final volume containing 1X PCR buffer, 0.2 mM of each dNTP, 0.5 μM of each primer, 75 ng of genomic DNA as template and 1 U of Taq DNA Polymerase (Invitrogen, Groningen, Netherlands). A touchdown cycling strategy was adopted using Biometra Thermocycler (Biometra, Göttingen, Germany). The thermocycler was programmed as follows: an initial denaturing step of 1 minute at 98 ºC, 35 cycles and a final extension of 10 minutes at 72 ºC. Each cycle consisted on a denaturing step of 10 seconds at 98 ºC, an annealing step of 30 seconds at temperatures depending on each primer pair decreasing each cycle by 0.5 ºC along 15 cycles and an extension step of 30 seconds at 72 ºC. Fragments were checked by electrophoresis in a 2 % agarose gel. Automated sequencing was performed by STAB Vida (Portugal).

Primer design was based on expressed sequences from TIGR database. Fragment sizes ranged from 300 bp to 1650 bp. The successfully amplified and sequenced regions were analysed on 90 clones. Amplification specificity was confirmed by BLAST with NCBI database. SNPs were identified using CodonCode Aligner software (Codon Code Corp.).

2.2. Variation on gene expression

2.2.1. Plant material

Clones of the cultivar Aragonez were used to perform the gene expression study. Figure II-1 shows a schematic drawing of the experimental design used. As described on section 2.1.1., Aragonez cultivar clones were grown on the same field, divided in three blocks with three plants of each clone per block.

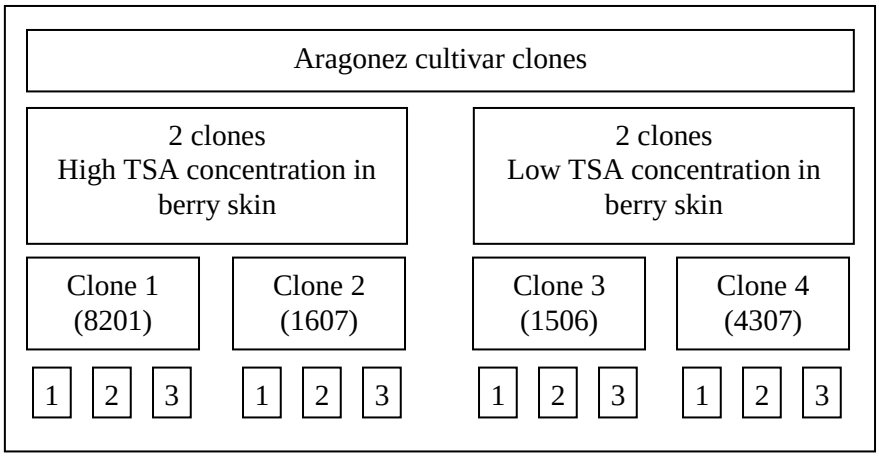


Figure II-1 Experimental design scheme.

Berries were collected from four clones of Aragonez. These clones were selected based on their contrasting TSA concentration, two clones with high average TSA concentration and two with low average TSA concentration (Table II-3).

Berries were collected at a maturity state corresponding to 13 to 14 % of probable alcohol. This was estimated by a non-destructive density measurement method by comparison of each berry density with a range of concentrated NaCl solutions. Collection was made between 8 a.m. and 11 a.m., followed by immediate freezing in liquid nitrogen and storage at

-80ºC. Three biological replicates of each clone, with ten berries each, were collected (Figure II-1).

Table II-3 Phenotypic data for clones used on differential expression analysis.

Clone	Average TSA concentration (mg/l)	Standard deviation	Maximum	Minimum
8201	475.17	82.77	588	375
1607	412.51	133.41	574	246
1506	835.00	194.45	1069	591
4307	830.83	219.60	1081	520

Data provided by Instituto Superior de Agronomia (Antero Martins, pers. comm.).

A number of previous studies reported associations between decreased anthocyanins concentration on berry skin and grapevine plants infection with grapevine leafroll associated viruses (GLRaVs) (Brar *et al.*, 2008; Guidoni *et al.*, 1997; Lee and Martin, 2009). In order to avoid bias in expression analysis due to this variable, ELISA tests were performed on the plants available for collection, testing for infection with GLRV 1, 2, 3 and 7. Most plants showed negative results for these viruses. Results for GLRaV3 on the clones 1607-2 and 1506-2 were borderline.

2.2.2. RNA extraction and microarray hybridisation

Total RNA was extracted from the skin of five berries per biological replicate according to the adapted protocol of Zeng and Yang (2002). The extracted RNA was then purified using RNeasy Mini Quiagen Kit (Quiagen Inc., Hilden, Germany), including DNase treatment. Quantification was performed spectrophotometricaly and quality was checked on 0.8 % agarose gel.

The Genomics Service of the Centro Nacional de Biotecnologia of Madrid performed labelled probe synthesis, RNA hybridisation and array

scanning. The array used was the custom made Affymetrix GrapeGen GeneChip™ (Lijavetzky *et al.*, *in preparation*) which consists of 23096 probesets. The publicly available Unigen information at the National Center for Biotechnology Information (by July 2006; http://www.ncbi.nlm.nih.gov) was used to design the probesets in this array. The probeset sequences included in GrapeGen GeneChip™ were annotated by manual and automated database searches on UniProtKB/Swiss-Prot plant protein databases (The UniProt Consortium, 2007; Pontin *et al.*, 2010).

2.2.3. Microarray data quality control, processing and analysis

Microarray data analysis was performed using the Gene Expression Profile Analysis Suite (GEPAS; Montaner *et al.*, 2006). Raw intensity values were first processed by Robust Multi-array Average (RMA) (Irizarry *et al.*, 2003). Expression values were computed from *CEL* files by applying the RMA model of probe-specific correction of perfect match (PM) (Li and Wong, 2001). Quantile array standardisation (Bolstad *et al.*, 2003) was used to normalise probe values and median polish procedure (Tukey, 1977) was applied to compute one expression measure from all probe values. Expression values were then $\log_2$-transformed.

To determine genes differentially expressed between clones with low and high TSA concentration, Student's *t*-tests were performed on the expression values. This analysis was performed on the total number of probesets. *T*-tests were carried using two clones of low and two clones of high TSA concentration and also pairwise comparisons using one clone of each condition. Table II-4 shows the list of *t*-tests performed.

Table II-4 List of *t*-tests performed for analysis of differential gene expression.

***T*-test**	**Clones included**	**Comparison description**
1	Clones 1 and 2 (high TSA concentration) versus Clones 3 and 4 (low TSA concentration)	Overall
2	Clone 1 (high TSA concentration) versus Clone 3 (low TSA concentration)	Pairwise
3	Clone 1 (high TSA concentration) versus Clone 4 (low TSA concentration)	Pairwise
4	Clone 2 (high TSA concentration) versus Clone 3 (low TSA concentration)	Pairwise
5	Clone 2 (high TSA concentration) versus Clone 4 (low TSA concentration)	Pairwise

Multiple testing adjustments were performed by computing the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) on the *P*-values of the *t*-tests performed. The FDR procedure consists of ordering n *P*-values of the tests. The threshold for rejection is obtained by finding the largest integer *i* such that $Pi \leq i\alpha/n.$ Where *Pi* is the *P*-value of the integer *i* and α is the significance level.

Gene lists were analyzed considering FDR corrected *P*-values, nominal *P*-values at different significance levels ($P < 0.05$ and $P < 0.01$) and fold-change level. For these analyzes, the ability to replicate results using both clones of each condition (high TSA concentration versus low TSA concentration) and one clone per condition, was also considered.

From the final gene list, a small group of genes were selected as candidate genes for genetic association analysis (Chapter IV). This selection was based on the transcription family, on the ability to replicate

results for different *t*-tests, on the *P*-values and on the ability to design good quality primer pairs. *Myb* and *Myc* families were prioritised since genes encoding these transcription factors have been previously shown to affect expression of genes involved in the anthocyanin biosynthetic pathway (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; Matus *et al*., 2009, 2010; Terrier *et al*., 2009; This *et al*., 2007).

2.2.4. Microarray validation

Quantitative real-time RT-PCR was performed to validate the expression profiles obtained by microarray analysis. A set of eight genes was used to examine transcript abundance (Appendix 3). Gene-specific primers were designed based on the corresponding probe sequences on the GrapeGen GeneChip™. Ubiquitin was selected as the reference gene which was confirmed to be expressed at a constant level in the current experimental conditions using GENORM (Vandesompele *et al.*, 2002).

RNA samples of one clone per condition studied in the microarray experiment, including three biological replicates of each, were used to synthesise cDNA. Reverse transcription was performed using the ImProm-II™ Reverse Transcription System (Promega) with Oligo(dT) primers. Three replicates of each reverse transcription were performed for each sample starting with 800ng of total RNA per reaction. The three cDNAs obtained were pooled. The obtained cDNA was quantified spectrophotometricaly.

Quantitative real-time PCR reactions were performed on an iQ5 Multicolor iCycler using iQ™ SYBR® Green Supermix (Bio-Rad Laboratories). Amplifications were performed following manufacturer's instructions in a total volume of 20µl with 1ng/µl of cDNA. The

thermocycler was programmed as follows: an initial denaturing step of 3 minutes at 95 ºC, 40 cycles and a final extension of 10 minutes at 72 ºC. Each cycle consisted on a denaturing step of 30 seconds at 95ºC, an annealing step of 30 seconds at 60 ºC and an extension step of 30 seconds at 72 ºC. Amplification of single products and absence of primer dimers were checked on agarose gel at 2 % and by melting curve analysis. Analysis of relative gene expression was determined for three technical replicates and two biological replicates. Raw Ct values were imported into the qBase Plus software (Biogazelle, Ghent, Belgium). This software was used to analyse relative expression of the genes taking into account gene-specific efficiency and normalizing to the reference gene Ubiquitin (Hellemans *et al.*, 2007). Serial template dilutions were used to assess amplification efficiencies of target and reference genes.

Correlation between the results obtained with quantitative real-time RT-PCR and microarray for this gene set was performed to validate microarray expression profiles. The data used for correlation over all genes were the $\log_2$ ratio value of the average of the replicates for each gene. The correlation values per gene were obtained for the $\log_2$ ratio across the three replicates.

3. Results

3.1. Sequence variation

The four candidate genes selection was based on availability of complete consensus sequences and reported SNPs on TIGR database. These were sequenced in two cultivars which show contrasting average values and a wide range of TSA concentration. Table II-5 shows the base pair lengths sequenced on this study. Since primer design was based on expressed sequences, the obtained fragments sizes showed a wide range

due to different sizes of introns. Fragment sizes ranged from 300bp to 1650bp. PCR amplification or sequencing did not succeed for few individuals or regions. These problems were mostly common for long DNA fragments, genes that were part of gene families or regions of repeated heterozygous INDELs. Overall, a total of 3380bp were sequenced, 1031bp on expressed sequences and 2349bp on intronic regions. *DFR* was the gene with a longest sequenced area (2429bp), comprising mostly intronic regions (2072bp) (Table II-5). The remaining genes were sequenced across smaller areas, from 181bp to 534bp, mostly in exonic regions (Table II-5).

Table II-5 Sequenced regions (bp) in clones of Aragonez and Negra Mole cultivars.

	DFR	LDOX	CHS	CHI	Overall
Exon	357	363	178	133	1031
Intron	2072	171	58	48	2349
Total	2429	534	236	181	3380

Table II-6 presents the identified polymorphisms between Negra Mole and Aragonez cultivars (using 50 Aragonez clones and 40 Negra Mole clones). This table shows also the polymorphisms location, base substitution and allele frequency. Information on polymorphism frequency per base pair sequenced is shown on Table II-7. The analysis of the sequenced regions between Aragonez and Negra Mole cultivars showed eight polymorphisms, seven SNPs and one INDEL. The gene showing the highest number of polymorphisms was *DFR* (Table II-6). However, polymorphisms frequency was higher in *CHS*, since the sequenced area was smaller (Table II-7). No polymorphisms were identified in the *LDOX* sequenced regions.

The number of polymorphisms identified was higher in intronic regions than in exons (Table II-6). However, overall the frequency was

Table II-6 Polymorphisms detected between Aragonez and Negra Mole cultivars.

Gene	Polymorphism ID	Chr.	Scaf.[1]	Position in scaffold	Type	Region	Amino acid change	Missing (%)	Allele 1	Allele 1 frequency	Allele 2	Allele 2 frequency	MAF
DFR	1_2949550_S	18	1	2949550	SNP	exon	syn	4.44	C	0.43	T	0.57	0.43
DFR	1_2949596_S	18	1	2949596	SNP	intron	_	3.33	C	0.78	A	0.22	0.22
DFR	1_2949618.5_I	18	1	2949618.5	INDEL	intron	_	3.33	N	0.78	_	0.22	0.22
DFR	1_2949647_S	18	1	2949647	SNP	intron	_	3.33	G	0.78	A	0.22	0.22
CHS_B	9_1279490_S	14	9	1279490	SNP	exon	syn	6.67	A	0.42	G	0.58	0.42
CHS_B	9_1279323_S	14	9	1279323	SNP	intron	_	5.56	T	0.74	C	0.26	0.26
CHS_B	9_1279306_S	14	9	1279306	SNP	intron	_	5.56	G	0.74	A	0.26	0.26
CHI	48_2128696_S	13	48	2128696	SNP	exon	syn	12.22	A	0.73	G	0.27	0.27

Chr. means chromosome and *Scaf.* means scaffold. Scaffold IDs are according to Genoscope (http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/), sequencing version 8x coverage. In the amino acid change column, *syn* means synonymous mutation.

higher in coding regions since the sequenced area was smaller. All the exon polymorphisms caused synonymous amino acid replacements (Table II-6). Transitions were more common than transversions. Six transitions, base substitutions where one pyrimidine base is replaced with another or one purine with another purine, were observed. Only one transversion was identified, where a pyrimidine is replaced with a purine or the other way around. Although theoretically, transversions should be twice as frequent as transitions, it has been observed in various species that transitions occur at a much higher frequency than transversions (Vignal et al., 2002). No polymorphisms were identified between the different clones of Aragonez or Negra Mole cultivars.

Table II-7 Frequency of polymorphisms between Aragonez and Negra Mole cultivars in the studied genomic regions.

	Frequency (1 polymorphism per bp)		
	Overall	**Coding regions**	**Non-coding**
Overall	422.5	343.7	469.8
DFR	607.3	357.0	690.7
LDOX	0.0	0.0	0.0
CHS	78.7	178.0	29.0
CHI	181.0	133.0	0.0

3.2. Differential gene expression

In this section the results for the differential gene expression analyses are presented. These are organised by showing first the overall results for Test 1, where expression values for Clones 1 and 2 were compared with Clones 3 and 4, using a range of threshold criteria. This is followed by a more detailed account of the functional categories with differentially expressed probesets for Test 1 ($P < 0.01$). Then the results are shown only for the differential expressed probesets encoding enzymes involved in the flavonoid metabolism and transcription factors. These results are

shown for two different significance levels and for the probesets significant for Test 1 and another pairwise test performed.

3.2.1. General outline of the results for Test 1

Table II-8 shows a summary of the number of probesets selected after applying different filtering thresholds to the results of the *t*-test performed between two clones with high TSA concentration and two with low TSA concentration (Test 1). This analysis showed 716 probesets (corresponding to 681 genes) significantly differentially expressed with $P < 0.05$ and 106 probesets (104 genes) with $P < 0.01$ (Table II-8, Appendix 4). After adjusting for FDR (Benjamini and Hochberg, 1995), no genes were differentially expressed between dark and light skinned clones ($P < 0.05$) (Table II-8). When a fold change of two was used as threshold only ten genes were differentially expressed for $P < 0.05$ (Appendix 5). At a significance of $P < 0.01$, this number was reduced to only one gene (Table II-8, Appendix 6).

Table II-8 Number of probesets listed after filtering *t*-test results for Test 1.

Filtering threshold	Total number of probesets
$P < 0.05$	716
$P < 0.01$	106
$P < 0.05$ and FC > 2	10
$P < 0.01$ and FC > 2	1
FDR $P < 0.05$	0

FC stands for fold-change.

3.2.2. Differentially expressed genes for Test 1 ($P < 0.01$)

Table II-9 shows the list of categories represented on the 106 probesets showing differential expression between clones with high and low TSA concentration ($P < 0.01$). Among the 106 significant probesets (corresponding to 104 genes) ($P < 0.01$), 50 were downregulated in the

lighter clones and 56 were upregulated. These probes included a high number of hypothetical proteins and of unclassified probesets (Table II-9). Also there were twelve genes involved in nucleic acid metabolism coding transcription factors, nine genes involved in protein metabolism, six in signal transduction, five in stress response, four in cell wall metabolism and two coding enzymes involved on the secondary metabolism (Table II-9).

Table II-9 List of probesets showing significant differential expression for Test 1 ($P < 0.01$).

Functional category	Number of probesets
Unclassified	25
Hypothetical protein	19
Nucleic acid metabolism	12
Protein metabolism and modification	9
Signal transduction	6
Stress response	5
Cell wall metabolism	4
Carbohydrate metabolism	3
Coenzyme and prosthetic group	3
Metabolite transport facilitation	3
Cell structure and motility	2
Lipid, fatty acid, steroid metabolism	2
Metabolism	2
Secondary metabolism	2
Storage protein	2
Amino acid metabolism	1
Cell growth and death	1
Cytochrome P450	1
Hormone metabolism	1
Myo-Inositol metabolism	1
Nitrogen metabolism	1
Pentatricopeptide repeat	1
Total	106

3.2.3. Functional categories of interest (flavonoid metabolism and transcription factors)

Differentially expressed genes for Test 1 (P < 0.01)

Table II-10 presents the probe annotation of all the probes significant for Test 1 ($P < 0.01$) that are part of the functional groups of transcription factors and flavonoid metabolism. Only one of the genes involved on secondary metabolism is involved in flavonoids metabolism (Table II-10). The 12 genes involved in nucleic acid metabolism included transcription factors of the YABBY, GRAS, Myb, WRKY and basic-leucine zipper families (Tables II-9, II-10).

Differentially expressed genes for Test 1 (P < 0.05)

The functional groups of transcription factors and flavonoid metabolism are especially interesting for the study of grape skin colour and were therefore further explored. Among the results for Test 1, the threshold for *P*-values was relaxed to 0.05 and only the functional categories of transcription factors and flavonoid metabolism were explored. By applying these criteria, a list of 39 probesets (38 genes) was obtained, including four involved on flavonoid biosynthesis and 35 transcription factors. This list is shown on Table II-11.

Differentially expressed genes for Test 1 and one of the remaining tests (P < 0.05)

A total of 24 genes involved in flavonoids metabolism and coding transcription factors were significant for Test 1 and more than one pairwise test ($P < 0.05$; Tests 2 to 5). These included two genes involved on the flavonoid metabolism (Q2LAM6; Q59J80), coding enzymes related with the glucosylation of flavonoids, an important step in anthocyanin biosynthesis. Table II-12 shows the 22 genes coding

Table II-10 List of probesets significantly differentially expressed for Test 1 (*P* < 0.01) and included in the functional categories of flavonoid metabolism and transcription factors.

Probeset ID	*P*-value	Fold change	UniProt ID	Annotation	Function	
VVTU165_at	$6.97X10^{-04}$	1.21	Q1S9P5	cAMP response element binding (CREB) protein related cluster	Transcription factor	Basic-leucine zipper (bZIP)
VVTU11499_at	$9.13X10^{-03}$	-1.12	Q1SRF6	GRAS transcription factor related cluster	Transcription factor	GRAS transcription factor
VVTU5365_at	$1.81X10^{-03}$	-1.16	O23063	A_IG005I10.6 protein related cluster	Transcription factor	General transcription factor
VVTU2631_at	$8.49X10^{-03}$	1.22	Q1XAN1	Sucrose responsive element binding protein related cluster	Transcription factor	Myb transcription factor
VVTU35012_at	$2.93X10^{-03}$	1.15	Q9ATD1	GHMyb9 related cluster	Transcription factor	Myb transcription factor
VVTU34340_s_at	$3.72X10^{-03}$	1.41	Q9FXS1	WRKY transcription factor NtEIG-D48 related cluster	Transcription factor	WRKY
VVTU1501_at	$2.00X10^{-03}$	-1.13	Q6SRZ8	YABBY2-like transcription factor YAB2 related cluster	Transcription factor	YABBY transcription factor
VVTU35538_at	$7.20X10^{-03}$	1.17	Q1S835	2OG-Fe(II) oxygenase related cluster	Flavonoid metabolism	_

transcription factors that were differentially expressed between low and high TSA clones. These included 12 families of transcription factors. Four genes encoded Myb transcription factors. Three genes were part of the basic helix-loop-helix transcription factor family. These two families have been pointed out as important in anthocyanin regulation in other species and in grapevine (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Dooner, 1991; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; Matus *et al.*, 2009; Terrier *et al.*, 2009; This *et al.*, 2007). Genes encoding transcription factors from other families such as WRKY, zinc finger and homeobox were also differentially expressed. Transcription factors of WRKY and zinc finger families have been shown to influence proanthocyanidin synthesis (Johnson *et al.*, 2002; Sagasser *et al.*, 2002) and Kubo *et al.* (1999) identified a homeobox gene to be involved in anthocyanin accumulation in *Arabidopsis*.

From the last gene list, a small group of genes were selected as candidate genes for genetic association analysis. This group of genes consisted of three *Myb* family genes with the UniProt codes Q2LME9, Q9ATD1 and Q3LHL3; and one *bHLH* family gene with UniProt code Q700B9 (see Chapter IV).

3.2.4. Array validation

Quantitative real time RT-PCR assays were performed on a subset of eight genes to validate expression profiles obtained with the microarray. Three replicates of clones two and four where used on this validation. Correlation between $\log_2$ ratio observed with quantitative real time RT-PCR and microarray for this group of genes was performed. Correlation coefficients obtained between the $\log_2$ ratio values for each gene under the two techniques are shown on Table II-13. A total of five genes

Table II-11 List of probesets functionally annotated as involved in flavonoid metabolism and transcription factors and significantly differentially expressed for Test 1 ($P < 0.05$).

Probeset ID	*P*-value	FC	UniProt ID	Function		Annotation
VVTU35538_at	$7.16x10^{-03}$	1.17	Q1S835	Flavonoid metabolism	_	2OG-Fe(II) oxygenase related cluster
VVTU226_s_at	$3.87x10^{-02}$	-1.38	Q9FEA1	Flavonoid metabolism	Anthocyanin biosynthesis	Anthocyanin 1 related cluster
VVTU2442_x_at	$3.60x10^{-02}$	1.24	Q2LAM6	Flavonoid metabolism	Plant flavonol 3-*O*-glucosyltransferase related	UDP-D-apiose UDP-D-xylose synthase related cluster
VVTU1972_at	$1.87x10^{-02}$	1.18	Q59J80	Flavonoid metabolism	Plant flavonol 3-*O*-glucosyltransferase related	Cyclo-DOPA 5-*O*-Glucosyltransferase related cluster
VVTU13141_at	$2.07x10^{-02}$	1.13	Q700B9	Transcription factor	Basic helix-loop-helix (bHLH)	Myc transcription factor related cluster
VVTU6674_at	$3.01x10^{-02}$	1.27	Q1S089	Transcription factor	Basic helix-loop-helix (bHLH)	Helix-loop-helix DNA-binding related cluster
VVTU34392_at	$3.77x10^{-02}$	1.19	Q41102	Transcription factor	Basic helix-loop-helix (bHLH)	Phaseolin G-box binding protein PG2 related cluster
VVTU14552_s_at	$3.90x10^{-02}$	1.24	Q1S089	Transcription factor	Basic helix-loop-helix (bHLH)	Helix-loop-helix DNA-binding related cluster
VVTU165_at	$6.97x10^{-04}$	1.21	Q1S9P5	Transcription factor	Basic-leucine zipper (bZIP)	cAMP response element binding (CREB) protein related cluster
VVTU3488_s_at	$3.89x10^{-02}$	-1.13	O22208	Transcription factor	Basic-leucine zipper (bZIP)	BZIP family transcription factor related cluster
VVTU3691_at	$1.55x10^{-02}$	1.50	Q8LJC3	Transcription factor	DOF	Zinc finger protein-like related cluster
VVTU31051_at	$1.81x10^{-02}$	1.43	Q0GLD0	Transcription factor	DOF	Dof21b related cluster
VVTU13044_at	$4.24x10^{-02}$	1.09	Q2XTB9	Transcription factor	General transcription factor	Transcription factor NF-Y CCAAT-binding-like protein-like related cluster
VVTU11499_at	$9.13x10^{-03}$	-1.12	Q1SRF6	Transcription factor	GRAS transcription factor	GRAS transcription factor related cluster
VVTU28168_s_at	$2.32x10^{-02}$	1.84	Q1SFJ9	Transcription factor	Homeobox domain	Leucine zipper, homeobox-associated, homeodomain-related related cluster
VVTU4060_at	$2.81x10^{-02}$	-1.42	Q8LLE1	Transcription factor	Homeobox domain	BEL1-related homeotic protein 14 related cluster
VVTU3022_at	$4.55x10^{-02}$	-1.30	Q1SIG3	Transcription factor	Homeobox domain	POX, homeodomain-related related cluster
VVTU5365_at	$1.81x10^{-03}$	-1.16	O23063	Transcription factor	General transcription factor	A_IG005I10.6 protein related cluster
VVTU32235_at	$1.25x10^{-02}$	-1.10	Q1SAU7	Transcription factor	General transcription factor	Paired amphipathic helix related cluster
VVTU35012_at	$2.93x10^{-03}$	1.15	Q9ATD1	Transcription factor	Myb transcription factor	GHMyb9 related cluster
VVTU2631_at	$8.49x10^{-03}$	1.22	Q1XAN1	Transcription factor	Myb transcription factor	Sucrose responsive element binding protein related cluster
VVTU39809_at	$2.21x10^{-02}$	-1.10	O49021	Transcription factor	Myb transcription factor	Myb-like DNA-binding domain protein related cluster
VVTU121_at	$2.28x10^{-02}$	-1.19	Q2LME9	Transcription factor	Myb transcription factor	Myb11 related cluster
VVTU3165_s_at	$4.14x10^{-02}$	1.09	Q3LHL3	Transcription factor	Myb transcription factor	Myb-CC type transfactor related cluster
VVTU16122_at	$4.40x10^{-02}$	-1.12	Q9LFL3	Transcription factor	Myb transcription factor	TOM (target of Myb1)-like protein related cluster
VVTU35624_at	$4.70x10^{-02}$	-1.17	Q0PJL6	Transcription factor	Myb transcription factor	Myb56 related cluster
VVTU22207_at	$1.75x10^{-02}$	1.21	Q1SMR9	Transcription factor	Pathogenesis-related transcription factor	Pathogenesis-related transcriptional factor and ERF related cluster
VVTU36189_s_at	$3.02x10^{-02}$	1.07	O81488	Transcription factor	Plant homeodomain (PHD) finger	PHD finger protein At5g26210 related cluster
VVTU3149_at	$3.09x10^{-02}$	-1.11	Q70MT1	Transcription factor	TCP transcription factor	Putative transcription factor related cluster
VVTU34340_s_at	$3.72x10^{-03}$	1.41	Q9FXS1	Transcription factor	WRKY transcription factor	WRKY transcription factor NtEIG-D48 related cluster
VVTU21525_at	$1.93x10^{-02}$	1.16	Q5JM93	Transcription factor	WRKY transcription factor	Putative WRKY DNA-binding protein 49 related cluster
VVTU1501_at	$1.96x10^{-03}$	-1.13	Q6SRZ8	Transcription factor	YABBY transcription factor	YABBY2-like transcription factor YAB2 related cluster
VVTU6691_at	$2.17x10^{-02}$	-1.13	Q3S345	Transcription factor	Zinc finger transcription factor	Zinc finger protein-like protein related cluster
VVTU10773_at	$3.98x10^{-02}$	1.15	Q1RYL8	Transcription factor	Zinc finger transcription factor	Zinc finger, DHHC-type related cluster
VVTU22486_at	$4.21x10^{-02}$	1.08	Q1S4D5	Transcription factor	Zinc finger transcription factor	FAR1, zinc finger, SWIM-type related cluster
VVTU28005_at	$1.82x10^{-02}$	-1.15	Q5JNB3	Transcription factor	Zinc finger, C3HC4-type	Zinc finger protein-like related cluster
VVTU5145_at	$1.93x10^{-02}$	1.16	Q2TE73	Transcription factor	Zinc finger, C3HC4-type	Ring zinc finger protein related cluster
VVTU12888_at	$4.70x10^{-02}$	1.23	Q1T0E0	Transcription factor	Zinc finger, CCHC-type	Zinc finger, CCHC-type related cluster
VVTU8889_at	$4.77x10^{-02}$	1.13	Q94AD9	Transcription factor	Zinc finger, CCCH-type	Zinc finger CCCH domain-containing protein ZFN-like 4 related cluster

FC stands for fold-change.

Table II-12 List of probesets of the functional groups involved on transcription factors, significantly differentially expressed for Test 1 ($P < 0.05$) and also significant for at least another *t*-test (2-5) ($P < 0.05$).

Probeset ID	Test 1		Test 2		Test 3		Test 4		Test 5		UniProt ID	Function
	P-value	FC	*P*-value	FC	*P*-value	FC	*P*-value	FC	*P*-value	FC		
VVTU13141_at	$2.07x10^{-02}$	1.13	n.s.	-	$7.35x10^{-03}$	1.16	n.s.	-	$8.65x10^{-03}$	1.14	Q700B9	Basic helix-loop-helix (bHLH)
VVTU6674_at	$3.01x10^{-02}$	1.27	n.s.	-	$2.17x10^{-02}$	1.53	n.s.	-	$3.10x10^{-02}$	1.39	Q1S089	Basic helix-loop-helix (bHLH)
VVTU34392_at	$3.77x10^{-02}$	1.19	$4.20x10^{-02}$	1.17	n.s.	-	n.s.	-	n.s.	-	Q41102	Basic helix-loop-helix (bHLH)
VVTU165_at	$6.97x10^{-04}$	1.21	n.s.	-	$8.04x10^{-03}$	1.29	n.s.	-	$1.68x10^{-03}$	1.23	Q1S9P5	Basic-leucine zipper (bZIP)
VVTU3691_at	$1.55x10^{-02}$	1.50	n.s.	-	n.s.	-	$4.55x10^{-02}$	1.64	n.s.	-	Q8LJC3	DOF
VVTU11499_at	$9.13x10^{-03}$	-1.12	n.s.	-	n.s.	-	n.s.	-	$4.10x10^{-02}$	-1.13	Q1SRF6	GRAS transcription factor
VVTU3022_at	$4.55x10^{-02}$	-1.30	n.s.	-	$4.49x10^{-02}$	-1.57	n.s.	-	n.s.	-	Q1SIG3	Homeobox domain
VVTU28168_s_at	$2.32x10^{-02}$	1.84	n.s.	-	$2.06x10^{-02}$	2.30	n.s.	-	n.s.	-	Q1SFJ9	Homeobox domain
VVTU5365_at	$1.81x10^{-03}$	-1.16	$3.83x10^{-02}$	-1.17	$8.58x10^{-03}$	-1.23	n.s.	-	$3.06x10^{-02}$	-1.15	O23063	General transcription factor
VVTU35012_at	$2.93x10^{-03}$	1.15	n.s.	-	$1.64x10^{-02}$	1.21	n.s.	-	$1.83x10^{-02}$	1.18	Q9ATD1	Myb transcription factor
VVTU121_at	$2.28x10^{-02}$	-1.19	$1.67x10^{-02}$	-1.22	$7.42x10^{-03}$	-1.40	n.s.	-	n.s.	-	Q2LME9	Myb transcription factor
VVTU35624_at	$4.70x10^{-02}$	-1.17	n.s.	-	$1.12x10^{-02}$	-1.27	n.s.	-	$1.06x10^{-02}$	-1.37	Q0PJL6	Myb transcription factor
VVTU3165_s_at	$4.14x10^{-02}$	1.09	$3.60x10^{-02}$	1.09	n.s.	-	n.s.	-	n.s.	-	Q3LHL3	Myb transcription factor
VVTU22207_at	$1.75x10^{-02}$	1.21	$2.93x10^{-02}$	1.38	$1.24x10^{-02}$	1.19	n.s.	-	n.s.	-	Q1SMR9	Pathogenesis-related transcription factor
VVTU36189_s_at	$3.02x10^{-02}$	1.07	n.s.	-	$3.58x10^{-02}$	1.06	n.s.	-	$1.30x10^{-02}$	1.13	O81488	Plant homeodomain (PHD) finger
VVTU34340_s_at	$3.72x10^{-03}$	1.41	n.s.	-	$2.33x10^{-02}$	1.62	$1.33x10^{-02}$	1.22	$9.09x10^{-03}$	1.63	Q9FXS1	WRKY transcription factor
VVTU21525_at	$1.93x10^{-02}$	1.16	n.s.	-	n.s.	-	$4.74x10^{-02}$	1.18	n.s.	-	Q5JM93	WRKY transcription factor
VVTU1501_at	$1.96x10^{-03}$	-1.13	$4.85x10^{-02}$	-1.08	$3.00x10^{-02}$	-1.16	$3.10x10^{-02}$	-1.10	$2.25x10^{-02}$	-1.18	Q6SRZ8	YABBY transcription factor
VVTU22486_at	$4.21x10^{-02}$	1.08	$7.08x10^{-03}$	1.15	n.s.	-	$1.65x10^{-02}$	1.09	n.s.	-	Q1S4D5	Zinc finger transcription factor
VVTU10773_at	$3.98x10^{-02}$	1.15	n.s.	-	n.s.	-	n.s.	-	$5.58x10^{-05}$	1.28	Q1RYL8	Zinc finger transcription factor
VVTU6691_at	$2.17x10^{-02}$	-1.13	n.s.	-	n.s.	-	n.s.	-	$4.29x10^{-02}$	-1.17	Q3S345	Zinc finger transcription factor
VVTU28005_at	$1.82x10^{-02}$	-1.15	n.s.	-	$4.60x10^{-02}$	-1.16	n.s.	-	n.s.	-	Q5JNB3	Zinc finger transcription factor

FC stands for fold-change and *ns* for non-significant *P*-values.

showed high correlation coefficients between the two platforms (r > 0.85; Table I-13).

However, the remaining genes included three showing lower correlation coefficients (r≈0.4) and a negative correlation (r = -0.5) (Table II-13). The correlation coefficient over all eight genes was r = 0.74 (*P* = 0.04). Excluding the gene *LOB1* as an outlier, a high correlation coefficient is obtained r = 0.89 (*P* < 0.007; Table II-13). Figure II-2 shows a plot of the gene expression ratios between quantitative real-time RT-PCR and microarray analyses excluding *LOB1*.

Other studies where correlations ranged between -0.48 and 0.94 are found in the literature (Beckman *et al.*, 2004; Etienne *et al.*, 2004; Larkin *et al.*, 2004). The utility of quantitative real-time RT-PCR to validate microarray studies has been widely debated since the two methods use very different normalisation procedures and both have inherent pitfalls (Morey *et al.*, 2006).

Table II-13 Correlation coefficients for each gene.

Genes	**Correlation Coefficient**	***P*-value**
BAG6	0.9997	0.0150
CCR2	0.9980	0.0407
LOB1	-0.5027	0.6647
MYB5	0.8680	0.3308
PAL1	0.8899	0.3016
RSGTA	0.9954	0.0608
RSGTC	0.4104	0.7308
TCP9	0.4245	0.7209
Over 8 genes	0.7393	0.0361
Over 7 genes, excluding *LOB1*	0.8900	0.0073

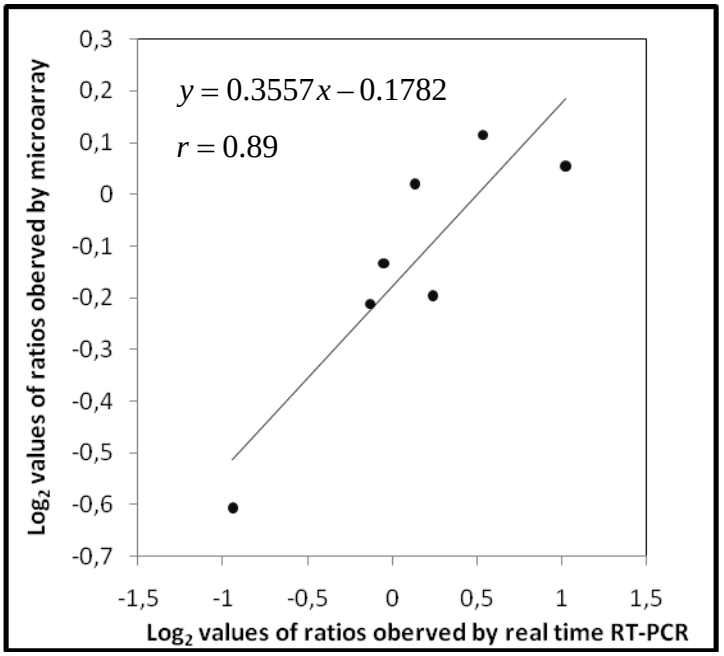


Figure II-2 Regression between gene expression ratios obtained by quantitative real-time RT-PCR of seven transcripts and microarray.

4. Discussion

Eight polymorphisms, one INDEL and seven SNPs were identified between Aragonez and Negra Mole cultivars; however, no sequence variations were found between different clones of each cultivar. Although grapevine clones have been successfully discriminated using SSR (Kozjak *et al.*, 2003; Moncada *et al.*, 2006; Regner *et al.*, 2000; Silvestroni *et al.*, 1997) and AFLP markers (Baneh *et al.*, 2009; Cervera *et al.*, 2000, 2001; Imazio *et al.*, 2002; Scott *et al.*, 2000; Sensi *et al.*, 1996), many other authors have failed to identify clonal polymorphisms in grapevine with SSR and ISSR (Baneh *et al.*, 2009; Faria *et al.*, 2004; Imazio *et al.*, 2002; Moreno *et al.*, 1998). A possible reason for the inability to detect polymorphisms may be the study of a small fraction of

the genome including only genes involved in the biosynthetic pathway of anthocyanins.

The gene expression analysis performed between clones with high and low anthocyanin concentration in berries skin showed subtle differences. After multiple testing adjustments, none of the genes showed significant differential expression (FDR $P < 0.05$). Therefore, further interpretation of these results must be carried carefully and consider the possibility of false positive occurrence. Due to this, attention was mainly focused on genes involved on functional groups *a priori* considered relevant for the phenotype of interest, TSA concentration. These functional groups consisted on genes coding enzymes involved in the flavonoid metabolism and transcription factors. To obtain a final list of probesets, the ability to replicate results was also considered, i.e. the fact that the same probeset showed significant differential expression for more than one of the tests performed. A group of 24 genes, including 22 transcription factors and two involved on the flavonoid metabolism met these criteria for $P < 0.05$. Two genes involved in the flavonoid metabolism, coding enzymes related with the glucosylation of flavonoids, an important step in anthocyanin biosynthesis, were differentially expressed between lighter and darker clones of Aragonez. The differentially expressed transcription factors included three *basic helix-loop-helix* genes and four *Myb* family genes, previously reported as important in anthocyanin biosynthesis regulation in grapevine and other species (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Dooner, 1991; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; Matus *et al.*, 2009; Terrier *et al.*, 2009; This *et al.*, 2007). Genes of other transcription factor families like zinc finger, WRKY and homeobox showed also differential expression and have been previously shown to be involved in proanthocyanidin and anthocyanin regulation (Johnson *et*

al., 2002; Kubo *et al.*, 1999; Sagasser *et al.*, 2002). Genes of other transcription factor families such as DOF, GRAS, YABBY, basic-leucine zipper, pathogenesis-related and plant homeodomain finger, showed differential expression as well. These genes should be further investigated to assess their influence on grape skin colour. Four genes of this list were selected as candidate genes for association mapping (*Q2LME9, Q9ATD1, Q3LHL3* and *Q700B9*).

These results showed that variation at the DNA sequence level influencing TSA concentration among the studied clones is not in the genomic regions sequenced and analysed for the presence of SNPs, although this may be found in other genomic regions. The use of a genome-wide approach would be a good design for this purpose. Nevertheless, some polymorphisms were identified between Aragonez and Negra Mole cultivars. These results also suggest that phenotypic differences in berry skin colour between clones may be related with subtle differences in gene expression, involving mostly genes coding transcription factors. Further investigation must be performed in order to confirm the role of these genes in skin colour variation between clones. The importance of some of these genes on anthocyanin content has been supported by association mapping results obtained in Chapter IV.

5. References

Ageorges, A., Fernandez, L., Vialet, S., Merdinoglu, D., Terrier, N. *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.

Baneh, H.D., Mohammadi, S.A., Mahmoudzadeh, H., Mattia, F., Labra, M. (2009). Analysis of SSR and AFLP Markers to Detect Genetic

Diversity Among Selected Clones of Grapevine (*Vitis vinifera* L.) cv. Keshmeshi. *South African Journal of Enology and Viticulture* **30**, 38-42.

Beckman, K.B., Lee, K.Y., Golden, T., Melov, S. (2004). Gene expression profiling in mitochondrial disease: assessment of microarray accuracy by high-throughput Q-PCR. *Mitochondrion* **4**, 453-470.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289-300.

Bertsch, C., Kieffer, F., Maillot, P., Farine, S., Butterlin, G., *et al.* (2005). Genetic chimerism of *Vitis vinifera* cv. Chardonnay 96 is maintained through organogenesis but not somatic embryogenesis. *BMC Plant Biology* **5**, 1-7.

Bisson, J. (1995). The principal ecogeographical groups in French grapevines assortment. *Journal International des Sciences de la Vigne et du Vin* **29**, 63-68.

Bogs, J., Jaffé, F. W., Takos, A. M., Walker, A. R., Robinson, S. P. (2007). The grapevine transcription factor VvMYBPA1 regulates proanthocyanidin synthesis during fruit development. *Plant physiology* **143**, 1347-1361.

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193.

Brar, H.S., Singh, Z. Swinny, E., Cameron, I. (2008). Girdling and grapevine leafroll associated viruses affect berry weight, colour development and accumulation of anthocyanins in 'Crimson Seedless' grapes during maturation and ripening. *Plant Science* **175**, 885-897.

Cervera, M.-T., Cabezas, J.A., Sancha, J.C., Martínrz de Toda, F. and Martínez-Zapater, J.M. (1998). Application of AFLPs to the characterization of grapevine *Vitis vinifera* L. genetic resources. A case study with accessions from Rioja (Spain). *Theoretical and Applied Genetics* **97**, 51-59.

Cervera, M.T., Cabezas, J.A., Sanchez-Escribano, E., Cenis, J.L., Martinez-Zapater, J.M. (2000). Characterization of genetic variation within table grape varieties (*Vitis vinifera* L.) based on AFLP. *Vitis* **39**, 109-114.

Cervera, M.T., Rodriguez, I., Cabezas, J.A., Chavez, J., Martinez-Zapater, J.M., *et al.* (2001). Morphological and molecular characterization of grapevine accessions as Albillo. *American Journal of Enology and Viticulture* **52**, 127-135.

Crespan, M. (2004). Evidence on the evolution of polymorphim of microsatellite markers in varieties of *Vitis vinifera* L. *Theoretical and Applied Genetics* **108**, 231-237.

Da Silva, F.G., Iandolino, A., Al-Kayal, F., Bohlmann, M.C.,Cushman, M.A. *et al.* (2005). Characterizing the grape transcriptome. Analysis of expressed seuqence tags from multiple *Vitis* species and development of a compendium of gene expression during berry development. *Genome Analysis* **139**, 574-597.

Deluc, L., Barrieu, F., Marchive, C., Lauvergeat, V., Decendit, A., Richard, T., *et al.* (2006). Characterization of a grapevine R2R3-MYB transcription factor that regulates the Phenylpropanoid pathway. *Plant Physiology* **140**, 499-511.

Deluc, L., Bogs, J., Walker, A. R., Ferrier, T., Decendit, A., *et al.* (2008). The transcription factor VvMYB5b contributes to the regulation of

anthocyanin and proanthocyanidin biosynthesis in developing grape berries. *Plant Physiology* **147**, 2041-2053.

Dooner, H.K. and Robbins T.P. (1991). Genetic and developmental control of anthocyanin biosynthesis. *Annual Review of Genetics* **25**, 173-179.

Downey, M. O., Dokoozlian, N. K. and Krstic, M. P. (2006). Cultural practice and environmental impacts on the flavonoid composition of grapes and wine: A review of recent research. *American Journal of Enology and Viticulture* **57**, 257-268.

Etienne, W., Meyer, M.H., Peppers, J., Meyer, R.A.Jr. (2004). Comparison of mRNA gene expression by RT-PCR and DNA microarray. *Biotechniques* **36**, 618-621.

Faria, M.A., Beja-Pereira, M., Martins, A., Ferreira, M.A., Nunes, M.E.S. (2004). Grapevine clones discriminated using stilbene synthase-chalcone synthase merkers. *Journal of the Science of Food and Agriculture* **84**, 1186-1192.

Fernandez, L., Doligez, A., Lopez, G., Thomas, M.R., Bouquet, A., Torregrosa, L. (2006). Somatic chimerism, genetic inheritance, and mapping of the *fleshless berry (flb)* mutation in grapevine (*Vitis vinifer* L.). *Genome* **49**, 721-728.

Fournier-Level, A., Le Cunff, L., Gomez, C., Doligez, A., Ageorges, A., et al. (2009) Quantitative Genetic Bases of Anthocyanin Variation in Grape (*Vitis vinifera* L. ssp. sativa) Berry: A Quantitative Trait Locus to Quantitative Trait Nucleotide Integrated Study. *Genetics* **183**: 1127-1139.

Franks, T., Botta, R and Thomas, M.R. (2002). Chimerism in grapevines: implications for cultivar identity, ancestry and genetics improvement. *Theoretical and Applied Genetics* **104**, 192-199.

Giusti, M.M. and Wrolstad, R.E. (2003). Acylated anthocyanins from edible sources and their applications in food systems. *Biochemical Engineering Journal* **14**, 217-225.

Guidoni, S., Mannini, F., Ferrandino, A., Argamante, N., Di Stefano, R. (1997). The effect of grapevine leafroll and rugose wood sanitation on agronomic performance and berry and leaf phenolic content of Nebbiolo clone (*Vitis vinifera* L.). *American Journal of Enology and Viticulture* **48**, 438-442.

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biology* **8**, R19.1-R19.14.

Hocquigny, S., Pelsy, F., Dumas, V., Kindt, S., Heloir, M-C., *et al.* (2004). Diversification within grapevine cultivars goes through chimeric states. *Genome* **47**, 579-589.

Holton, T.A. and Cornish, E.C. (1995). Genetics and Biochemistry of Anthocyanin Biosynthesis. *The Plant Cell* **7**, 1071-1083.

Imazio, S., Labra, M., Grassi, F., Winfield, M. Bardini, M., e*t al.* (2002). Molecular tools for clone identification: the case of the grapevine cultivar 'Traminer'. *Plant Breeding* **121**, 531-535.

Irizarry, R.A., Hobbs, B., Collin, F. Beazer-Barclay, Y.D., Antonellis, K.J., *et al.* (2003). Exploration, normalization, and summaries of high density oligonuceotide array probe level data. *Biostatistics* **2**, 249-264.

Jaillon, O., Aury, J-M., Noel, B., Policriti, A., Clepet, C. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467.

Jeong, S. T., Goto-Yamamoto, N., Hashizume, K. and Esaka, M. (2006). Expression of the flavonoid 3'-hydroxylase and flavonoid 3',5'-

hydroxylase genes and flavonoid composition in grape (*Vitis vinifera*). *Plant Science* **170**, 61-69.

Johnson, C.S., Kolevski, B., Smyth, D.R. (2002). *TRANSPARENT TESTA GLABRA2*, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. *The Plant Cell* **14**, 1359-1375.

Kaeppler, S.M., Kaeppler, H. F., Rhee, Y. (2000). Epigenetic aspects of somaclonal variation in plants. *Plant Molecular Biology* **43**, 179-188.

Kobayashi, S., Ishimaru, M., Hiraoka, K. and Honda, C. (2002). *Myb*-related genes of the Kyoho grape (*Vitis labruscana*) regulate anthocyanin biosyntesis. *Planta* **215**, 924-933.

Kobayashi, S., Goto-Yamamoto, N., Hirochika, H. (2004). Retrotransposon-Induced Mutations in Grape Skin Color. *Science* **304**, 982.

Kozjak, P., Korosec-Koruza, Z., Javornik, B. (2003). Characteristics of cv. Refosk (*Vitis vinifera* L.) by SSR markers. *Vitis* **42**, 83-86.

Kubo, H., Peeters, A.J.M., Aarts, M.G.M., Pereira, A., Koornneef, M. (1999). *ANTHOCYANINLESS2*, a homeobox gene affecting anthocyanin distribution and root development in Arabidopsis. *The Plant Cell* **11**, 1217-1226.

Larkin, J.E., Frank, B.C., Gaspard, R.M., Duka, I., Gavras, H. *et al.* (2004). Cardiac transcriptional response to acute and chronic angiotensin II treatments. *Physiology and Genomics* **18**, 152-166.

Lee, J. and Martin, R. (2009). Influence of grapevine leafroll associated viruses (GLRaV-2 and -3) on the fruit composition of Oregon *Vitis vinifera* L. Cv. Pinot noir: Phenolics. *Food Chemistry* **112**, 889-896.

Li, C. and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Science of USA* **98**, 31-36.

Lijavetzky, D., Ruiz-García, L., Cabezas, J. A., De Andrés, M. T., Bravo, *et al.* (2006). Molecular genetics of berry colour variation in table grape. *Molecular Genetics and Genomics* **276**, 427-435.

Lijavetzky, D. Cabezas, J.A., Ibáñez, A., Rodríguez, V., Martínez-Zapater, J.M. (2007). High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* **8**, 424-435.

Loureiro, M.D., Martinez, M.C., Boursiquot, J.M., This, P. (1998). Molecular marker analysis of *Vitis vinifera* 'Albarino' and some similar grapevine cultivars. *Journal of the American Society for Horticultural Science* **5**, 842-848.

Martin, C. and Gerats, T. (1993). Control of pigment biosynthesis genes during petal development. *Plant Cell* **5**, 1253-1264.

Matus, J. T., Loyola, R., Vega, A., Peña-Neira, A., Bordeu, E., *et al.* (2009). Post-veraison sunlight exposure induces MYB-mediated transcriptional regulation of anthocyanin and flavonol synthesis in berry skins of *Vitis vinifers. Journal of Experimental Botany* **60**, 853-867.

Matus, J.T., Poupin, M.J., Cañón, P., Bourdeu, E., Alcalde, J.A., *et al.* (2010). Isolation of WDR and bHLH genes related to flavonoid synthesis in grapevine (*Vitis vinifera* L.). *Plant Molecular Biology* **72**, 607-620.

Moncada, X., Pelsy, F., Merdinoglu, D., Hinrichsen, P. (2006). Genetic diversity and geographical dispersal in grapevine clones revealed by microsatellite markers. *Genome* **49**, 1459-1472.

Montaner, D., Tarraga, J., Huerta-Cepas, J., Burguet, J., Vaquerizas, J.M. *et al.* (2006). Next station in microarray data analysis: GEPAS. *Nucleic Acids Research* **34**, W486-A491.

Moreno, S., Martín, J.P. and Ortiz, J.M. (1998). Inter-simple sequence repeats PCR for characterization of closely related grapevine germplasm. *Euphytica* **101**, 117-125.

Morey, J.S., Ryan, J.C., Van Dolan, F.M. (2006). Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biology Procedings Online* **8**, 175-193.

Mori, K., Goto-Yamamoto, N., Kitayama, M., Hashizume, K. (2007). Loss of anthocyanins in red-wine grape under high temperature. *Journal of Experimental Botany* **58**, 1-11.

Mulcahy, D.L., Cresti, M., Sansavini, S., Douglas, G.C., Linskens, H.F. *et al.* (1993). The use of random amplified polymorphic DNAs to fingerprint apple genotypes. *Scientia Horticulturae* **54**, 89-96.

Organisation Internationale da la Vigne et du Vin (OIV). (2009). *International list of vine varieties and their synonyms*. OIV: Paris.

Pelsy, F. (2010). Molecular and cellular mechanisms of diversity within grapevine cultivars. *Heredity* **104**, 331-340.

Peng, F.Y., Reid, K.E., Liao, N., Schlosser, J., Lijavetzky, D. *et al.* (2007). Generation of ESTs in *Vitis vinifera* wine grape (Cabernet Sauvignon) and table grape (Muscat Hamburg) and discovery of new candidate genes with potential roles in berry development. *Gene* **402**, 40-50.

Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research* **9**, e45.

Pontin, M.A., Piccolo, P.N., Francisco, R., Bottini, R., Martinez-Zapater, J.M., Lijavetzky, D. (2010). Transcriptome changes in grapevine

(*Vitis vinifera* L.) cv. Malbec leaves induces by ultraviolet-B radiation. *BMC Plant Biology* **10**, 224.

Regner, F., Stadlbauer, C., Eisenheld, C., Kaserer, H. (2000). Genetic relationships among Pinots and related cultivars. *American Journal of Enology and Viticulture* **51**, 7-14.

Riaz, S., Garrison, K.E., Dangl, G.S., Boursiquot, J.M., Meredith, C.P. (2002). Genetic divergence and chimerism within ancient asesually propagated wine grape cultivars. *Journal of the American Society for Horticultural Science* **127**, 508-514.

Sagasser, M., Lu, G., Hahlbrock, K., Weisshaar, B. (2002). *A. thaliana* TRANSPARENT TESTA 1 is involved in seed coat development and defines the WIP subfamily of plant zinc finger proteins. *Genes and Development* **16**, 138-149.

Schellenbaum, P., Mohler, V., Wenzel, G., Walker, B. (2008). Variation in DNA methylation of grapevine somaclones (*Vitis vinifera* L.). *BMC Plant Biology* **8**, 78-98.

Scott, K.D., Ablett, E.M., Lee, L.S., Henry, R.J. (2000). AFLP markers distinguishing an early mutant of Flame seedless grape. *Euphytica* **113**, 243-247.

Sensi, E., Vignani, R., Rohde, W., Biricolti, S. (1996). Characterization of genetic biodiversity with *Vitis vinifera* L. Sangiovese and Colorino genotypes by AFLP and ISTR DNA marker technology. *Vitis* **35**, 183-188.

Silvestroni, O., DiPietro, D., Intrieri, C., Vignani, R., Filippetti, I., *et al.* (1997). *Vitis* **36**, 147-150.

Sparvoli, F., Martin, C., Scienza, A., Gavazzi, G., Tonelli, C. (1994). Cloning and molecular analysis of structural genes involved in

flavonoid and stilbene biosynthesis in grape (*Vitis vinifera* L.). *Plant Molecular Biology* **24**, 743-755.

Terrier, N., Torregrossa, L., Ageorges, A., Vialet, S., Verriès, C., *et al.* (2009). Ectopic expression of VvMybPA2 promotes proanthocyanidin biosynthesis in grapevine and suggest additional targets in the pathway. *Plant Physiology* **149**, 1028-1041.

The UniProt Consortium. (2007). The Universal Protein Resources (UniProt). *Nucleic Acids Res.* **35**, D193-D197.

This, P., Lacombe, T., Cadle-Davidson, M. and Owens, C. (2007). Wine grape (*Vitis vinifera* L.) color associates with allelic variation in the domestication gene *VvmybA1*. *Theoretical and Applied Genetics* **114**, 723-730.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Massachusetts: Addison-Wesley Publishing, pp 688.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., *et al.* (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology* **7**, research0034.1-0034.11.

Velasco, R., Zharkikh, A., Troggio, M. Cartwright, D.A., Cestaro, A. *et al.* (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* **12**, 1-18.

Vetten, N., Quattrochio, F., Mol, J., Koes, R. (1997). The *an11* locus controlling flower pigmentation in petunia encodes a novel WD-repeat protein conserved in yeast, plants, and animals. *Genes & Development* **11**, 1422-1434.

Vezzulli, S., Troggio, M., Coppola, G., Jermakow, A., Cartwright, D. *et al.* (2008). A reference integrated map for cultivated grapevine (*Vitis*

vinifera L.) from three crosses, based on 283 SSR and 501 SNP-based markers. *Theoretical and Applied Genetics* **117**, 499-511.

Vignal, A., Milan, D., SanCristobal, M., Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection and Evolution* **3**, 275-305.

Walker AR, Lee E, Bogs J, McDavid DAJ, Thomas MR, et al. (2007) White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant Journal* **49**: 772-785.

Walter, B. and Martelli, G.P. (1998). Considerations on grapevine selection and certification. *Vitis* **37**, 87-90.

Yamane, T., Jeong, S.T., Goto-Yamamoto, N., Koshita, Y., Kobayashi, S. (2006). Effects of temperature on anthocyanin biosynthesis in grape berry skins. *American Journal of Enology and Viticulture* **57**, 54-59.

Zeng, Y. and Yang, T. (2002). RNA isolation from highly viscous samples rich in polyphenols and polysaccharides. *Plant Molecular Biology Reporter* **20**, 417a-417e.

6. Acknowledgements

I would like to acknowledge Antero Martins and Elsa Gonçalves for the phenotypic data and plant material on Aragonez and Negra Mole cultivars. I would also like to thank José Miguel Martínez Zapater and Diego Lijavetzky for the advice on microarray experimental design and interpretation, Gema Sanz, Virginia Rodriguez and Rita Francisco for assistance on laboratory work and data analysis.

CHAPTER III

GRAPEVINE CULTIVAR CHARACTERISATION USING GENOTYPIC AND PHENOTYPIC DATA ON BERRY COLOUR AND ANTHOCYANIN COMPOSITION

A shorter version of this chapter will be shortly submitted to *Theoretical and Applied Genetics*:

Cardoso, S., Maniatis, N., Spranger, I., Moreira, F., Eiras Dias, J., Fevereiro, P. (To submit). Grapevine cultivar characterisation using genotypic and phenotypic data on berry colour and anthocyanin composition.

Contributions to this chapter:

- Designed the study: Cardoso, S., Eiras Dias, J.E., Fevereiro, P., Maniatis, N.
- Performed experiments: Cardoso, S., Moreira, F, Spranger, I.
- Analysed data: Cardoso, S., Maniatis, N.

Summary

Anthocyanin content determines grape and wine colour. This is a trait of utmost importance for grape and wine marketing. Characterisation of large collections of grapevine cultivars for this trait is scarce and has often been done based only on categorical visual assessments of colour. In this study, the anthocyanins profile of 149 cultivars in the Portuguese germplasm collection was characterised using RT-HPLC. Multivariate analysis tools were used to characterise the phenotypic diversity using anthocyanins concentration and relative abundance data. It was observed that although overlapping slightly, characterisation of cultivars using concentration and relative abundance of specific anthocyanins is not identical. Principal Component Analysis showed both data to separate cultivars based on anthocyanidin type and acylation pattern. However, they differed as concentration separated cultivars by methylation level and relative abundance by hydroxylation. For the concentration classification, total skin anthocyanin concentration was the strongest discriminating variable. Cultivars characterisation based on phenotypes was compared with the classification based on SSR markers. The classification based on concentration data was found to be uncorrelated with SSR classification. The one based on relative abundance showed a significant but weak correlation with SSR. The SSR classification was found to be more accurate. Visual characterisation showed skin colour to be more strongly related to relative abundance while pulp colour reflects mainly concentration of anthocyanins.

This work presents data of major importance for characterisation of the anthocyanin profile of cultivars and will be a basis for further work on anthocyanin profile studies and association mapping.

1. Introduction

Vitis vinifera L. is a widely cultivated crop with great economic importance. The optimal management of the species genetic resources depends heavily on a deep knowledge of its genotypic and phenotypic diversity. The incomplete characterisation of the existing germplasm is a limiting factor for the conservation and utilisation of grapevine genetic resources. Further difficulties arise also from the use of synonymous and homonymous cultivar designations.

The combined use of phenotypic data and molecular markers has been argued as the best option for diversity analysis and management of plant genetic resources. Phenotypic data provides valuable information on germplasm evaluation despite environmental factors influence on the observed variation. On the other hand, molecular markers provide a direct assessment of genetic diversity.

Traditionally, grape cultivars identification has been based on ampelography. However, this approach is limited because very similar cultivars cannot be easily differentiated by visual comparison (Aradhya *et al.*, 2003). On the other hand intracultivar clones may differ phenotypically despite being genetically identical (Franks *et al.*, 2002; Riaz *et al.*, 2002; Vignani *et al.*, 1996). As a result, molecular markers have become a popular alternative for the characterisation and identification of grapevine cultivars (Bowers and Meredith, 1996; Tessier *et al.*, 1999, Cervera *et al.* 1998; Aradhya *et al.* 2003). Particularly microsatellite (SSR) markers have been favoured due to their codominant nature and reproducibility (Sefc *et al.*, 2001).

Another successful approach to surpass the limitations of ampelographic analysis has been the use of chemical markers (Caló *et al.*, 1994; Benin *et al.*, 1988). Anthocyanins are one of the most important

metabolic compounds used for this purpose (Ribéreau-Gayon, 1959, 1964; Caló *et al.*, 1994; Wenzel *et al.*, 1987; Arozarena *et al.*, 2002). Anthocyanins are a group of compounds included in the flavonoids family. These compounds are responsible for grape skin, pulp and wine colouration. It has been generally accepted that relative abundance of anthocyanins is primarily determined by genetic factors (Ribéreau-Gayon, 1978; Caló *et al.*, 1994; Mazza and Miniati, 1993; Mazza, 1999). It is however known that maturation, temperature and light exposure influence anthocyanin content (Cacho *et al.*, 1992; Mori *et al.*, 2007). Viral infections can reduce anthocyanin content but the effects vary widely among cultivars, virus strains and with other environmental factors (Guidoni *et al.*, 2000; Lider *et al.*, 1975, Goheen, 1958; Cabaleiro *et al.*, 1999). Despite the influence of environmental factors in anthocyanins content, the use of these chemical markers has been considered more accurate than ampelographic methods, and has been successfully used for cultivar characterisation (Ryan and Revilla, 2003; Carreño *et al.*, 1997; Arozarena *et al.*, 2002).

Anthocyanins are not only important in colour determination but also influence grapes and wine organoleptic properties due to their ability to interact with other compounds such as proteins, polysaccharides and other phenolic compounds (Mazza and Miniati, 1993). Other important aspects of these compounds are their antioxidant properties and benefits for human health. Therefore, anthocyanins are not only useful as chemical markers for cultivar classification, but are also an important trait for grape and wine marketing.

Vitis vinifera L. coloured cultivars have only five anthocyanidins (delphinidin, cyanidin, petunidin, peonidin and malvidin) in the form of 3-monoglucosides and acetate, coumarate and caffeoate derivatives. The

colour of anthocyanins is a consequence of the number of hydroxyl, methoxyl groups and glycosylation and acylation patterns of the molecule (Grotewold *et al.*, 1998). Differences in enzymatic activities influence the acylation pattern and anthocyanidins that are found in a cultivar. Acyl transferase activity determines the presence of acyl derivatives. Flavonoid 3' hydroxylase and flavonoid 3'5' hydroxylase activities affect the proportion of dihydroxylated (cyaniding and peonidin) and trihydroxylated (delphinidin, petunidin and malvidin) anthocyanins. Methyl transferases determine the presence of methoxylated anthocyanins (peonidin, petunidin and malvidin).

The aims of this chapter are to:

a) Characterise grapevine cultivars based on anthocyanin content;
b) Compare cultivar characterisation using anthocyanins content data with molecular marker (SSR) data analysis;
c) Study the relationships between relative abundance and concentration of anthocyanins and visual colour classification of grape berries;
d) Analyse the impact of plants viral infection and berries maturity state on anthocyanin content for association mapping purposes.

2. Material and methods

2.1. Plant material

A sample with 149 cultivars was collected from the same vineyard in Dois Portos, Portugal, where the national ampelographic collection is established. These cultivars are listed in Appendix 7.

For phenotypic characterisation fifty healthy berries from different parts of the plant and bunch were collected from each cultivar. Probable alcohol percentage which was used as an indicator of berries maturity

state at harvest was measured using a hand Atago refractometer (Atago, Madrid). The collected berries had approximately 8 % probable alcohol. It must be considered however that this is a destructive technique and therefore some variation among berries maturation state was expected. The berries were stored at -20ºC. Young leaves were collected from the same plants as it was done for phenotypic characterisation and stored at -20ºC. These leaves were later used for DNA extraction for molecular analysis.

2.2. DNA extraction and genotyping

Genomic DNA was extracted from 100 mg of leaf fresh weight using Quiagen Mini Kit (Quiagen Inc, Hilden, Germany) with mortar and pestle grinding with sterile quartz sand. Quantification was done spectrophotometricaly.

Data on 20 SSR loci scattered across 18 different chromosomes for 149 cultivars were provided by the Istituto Agrario San Michele all'Adige (IASMA). Although two pairs of markers were located on the same chromosome, the distance between them according to previous linkage studies was 5 cM (Doligez *et al.*, 2006; Troggio *et al.*, 2007). These markers were independent and had a high Polymorphic Information Content (PIC), with an average of 0.7 (Appendix 8).

2.3. Anthocyanin extraction

Each sample collected was divided in two replicates and each replicate included 25 berries. The berries were peeled manually and the skins grind in mortar and pestle with liquid nitrogen. The grinded frozen skins were dispersed in 50ml 0.1 % hydrochloric acid in methanol and stored at -

20ºC for one hour. The skins were further agitated for 30 minutes at 35 rpm in 25 ml of renewed solvent three times.

2.4. Anthocyanins identification

The anthocyanins extracted were analysed using reversed phase high performance liquid chromatography (HPLC) using a Waters chromatograph (Waters Scientific, Mississauga, Ontario) equipped with a photodiode array detector. Extracts were passed through a 0.2 μm filter (Waters Scientific, Mississauga, Ontario). Samples of 20 μl were injected onto a reversed phase C_{18} column. Flow rate was 0.7 ml/min and the mobile phase consisted of formic acid and water (5:95, v/v) as solvent A, acetonitrile/water/formic acid 5 % (30:65:5, v/v) as solvent B, acetonitrile/water (75:25, v/v) as solvent C and methanol as solvent D. Data were collected by Millenium[32] software. Chromatograms were acquired at 525 nm and photodiode array spectra were recorded between 250 and 600 nm.

Individual anthocyanins identification was based on retention times and spectral properties (Table III-1). The typical observed chromatogram showed 18 peaks. Considering retention time and spectral properties was possible to identify 17 anthocyanin compounds and one isomer. The isomer presence was confirmed by exposure to UV light and analysis of the changed peak areas. Figure III-1 shows the identified anthocyanins and the corresponding retention times.

2.5. Anthocyanins quantification

Concentrations were calculated using a calibration curve obtained by regression through the origin of HPLC peak areas on concentration (in mg/l) of an external pattern of malvidin-3-O-glucoside chloride (Hoffman-La Roche, Switzerland). Concentrations were expressed in the

following three different ways: milligrams of anthocyanins per litre of extract, milligrams of anthocyanins per berry and milligrams of anthocyanins per kilogram of berries.

Table III-1 Spectral characteristics and retention times of the chromatographic peaks identified.

Peak number	Retention time (min.)	λ maximum (nm)	Identification	Symbol used
1	20.6	542;278[1]	Delphinidin-3-monoglucoside	Df
2	26.1	530;282[1]	Cyanidin-3-monoglucoside	Cy
3	30.4	540;278[1]	Petunidin-3-monoglucoside	Pt
4	36.0	528;280[1]	Peonidin-3-monoglucoside	Pn
5	39.3	538;278[1]	Malvidin-3-monoglucoside	Mv
6	43.6	542;280[1]	Delphinidin-3-monoglucoside-acetate	Dfac
7	48.8	530;500; 280; 270[1]	Cyanidin-3-monoglucoside-acetate	Cyac
8	53.5	540; 280[1]	Petunidin-3-monoglucoside-acetate	Ptac
9	60.4	527; 281[1]	Peonidin-3-monoglucoside-acetate	Pnac
10	61.5	542; 282[1]	Delphinidin-3-monoglucoside-*p*-coumarate	Dfcoum
11	62.8	538; 280[1]	Malvidin-3-monoglucoside-acetate	Mvac
12	66.7	522[2]	Peonidin-3-monoglucoside-caffeoate	Pncaff
13	68.0	532; 283[1]	Cyanidin-3-monoglucoside-*p*-coumarate	Cycoum
14	68.4	538; 328; 283[1]	Malvidin-3-monoglucoside-caffeoate	Mvcaff
15	70.4	541; 283[1]	Petunidin-3-monoglucoside-*p*-coumarate	Ptcoum
16	71.8	538; 280[2]	Cis-Malvidin-3-monoglucoside-*p*-coumarate	Cmvcoum
17	77.4	528; 283[1]	Peonidin-3-monoglucoside-*p*-coumarate	Pncoum
18	79.1	538; 284[1]	Malvidin-3-monoglucoside-*p*-coumarate	Mvcoum

[1] In 0.01% hydrochloric acid in methanol (Wulf and Nagel, 1978).
[2] In 5% acid in methanol (Nuñez *et al.*, 2003).

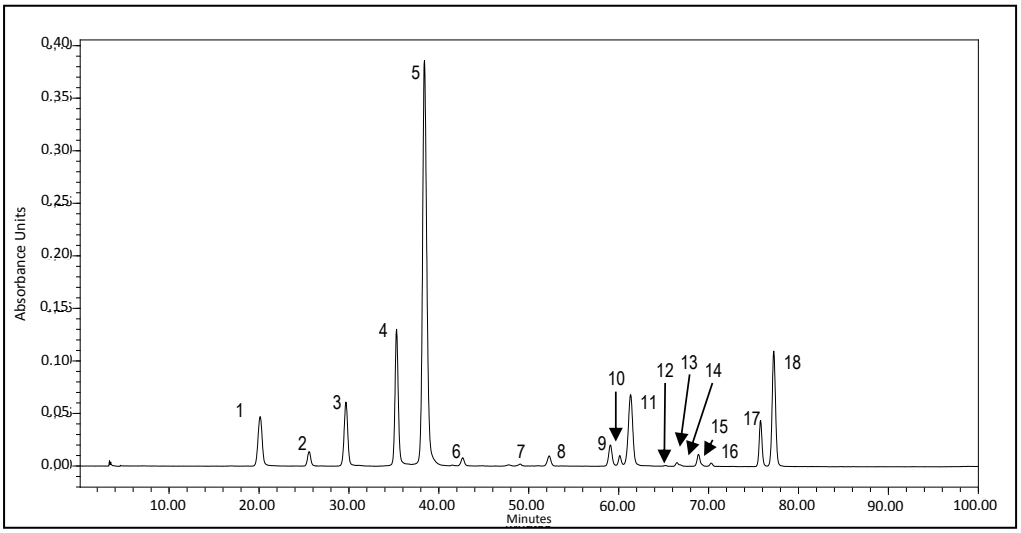


Figure III-1 Typical HPLC chromatogram (peaks identified on Table III-1).

2.6. Berry colour visual characterisation

Different visual characterisations of berries colour were used and are described in Table III-2. Pulp colour (PC) and skin colour (SC) were used. PC was characterised as a dichotomous trait (coloured versus white pulp) following descriptor number 230 by the International Organization of Vine and Wine (OIV, 1983). SC was classified according to OIV descriptor number 225 (OIV, 2009a). This descriptor establishes the following five categories: rose, red, grey, dark red violet and blue black.

OIV descriptor 225 (OIV, 2009a) has a certain degree of subjectivity and PC is likely to influence to some extent the classification of SC. Since these visual phenotypes were aimed to be used for genetic association analysis (Chapter IV), two new classifications were tested targeting higher accuracy. These classifications were named SPC and SPC′ and were obtained by joining pulp and skin colour classifications (Table III-2). For SPC only three categories were established, the first including cultivars with rose and red skin berries with white pulp, the

second for grey, dark red violet and blue black skin cultivars with white pulp and the third for red pulp cultivars. For SPC′ a sixth category for red pulp cultivars was added to the five OIV225 categories (OIV, 2009a). Table III-2 shows the different visual characterisations of berry colour used in this study.

Table III-2 List of visual colour characterisations of grape skin and pulp considered.

Categories
Pulp Colour (PC)
Coloured pulp
White pulp
Skin colour (SC)
Rose
Red
Grey
Dark red violet
Blue black
Skin and pulp 1 (SPC′)
Rose skin and white pulp
Red and white pulp
Grey and white pulp
Dark red violet and white pulp
Blue black and white pulp
Coloured pulp
Skin and pulp 2 (SPC)
Rose and red skin and white pulp
Grey, dark red violet and blue black skin and white pulp
Coloured pulp

2.7. Anthocyanins content potential covariates

Traits related with the maturity of berries were measured by the Central Laboratory of the Instituto Nacional de Investigação Agrária (INIA-Dois Portos) during the anthocyanin extraction. Brix degree (% m/m), sugar content (g/l), volumic mass (g/cm3) and probable alcohol (%

v/v) were determined by refractometry. Total acidity (g/l tartaric acid) was measured by colorimetric titration. All these measurements were performed according to OIV method (OIV, 2009b).

Data on plants viral infection was obtained by ELISA tests for grapevine fanleaf virus (GFLV), arabic mosaic virus (ArMV), grapevine fleck virus (GFKL), grapevine leafroll-associated viruses (GLRaV1, GLRaV2, GLRaV3, GLRaV7) and grapevine virus B (GVB). These tests results were kindly provided by Instituto Nacional de Investigação Agrária (Oeiras).

2.8. Statistical analysis

Data on anthocyanins concentration and peak areas (relative abundance in percentage of HPLC total peaks area) were considered. Ratios between di and trihydroxylated anthocyanins and between coumarate and acetate derivatives were also analysed since these have been suggested to reflect different enzymatic activities (Benin *et al.*, 1988).

Table III-3 shows the list of variables considered, which were 17 anthocyanin compounds and one isomer. This includes five anthocyanidins (delphinidin, cyanidin, petunidin, peonidin and malvidin) in the form of 3-monoglucosides, acetate and coumarate derivatives. Peonidin and malvidin also appear as caffeoate derivatives. An isomer of malvidin-3-monoglucoside-*p*-coumarate was considered as well. All anthocyanins concentrations were expressed in milligram per litre of extract, milligram per berry and milligram per kilogram of berries. Relative abundance of each anthocyanin expressed as percentage peak area was also measured for each anthocyanin compound. Table III-3 shows total anthocyanins, which were measured as concentration only. Finally, ratios between di and trihydroxylated anthocyanins and between

coumarate and acetate derivatives measured only as relative abundance are shown.

Table III-3 List of variables considered on phenotypic characterisation of cultivars.

Anthocyanin	Concentration			Relative abundance (%)
	Mg per litre of extract	Mg per berry	Mg per kg of berries	
Delphinidin-3-monoglucoside	✔	✔	✔	✔
Cyanidin-3-monoglucoside	✔	✔	✔	✔
Petunidin-3-monoglucoside	✔	✔	✔	✔
Peonidin-3-monoglucoside	✔	✔	✔	✔
Malvidin-3-monoglucoside	✔	✔	✔	✔
Delphinidin-3-monoglucoside-acetate	✔	✔	✔	✔
Cyanidin-3-monoglucoside-acetate	✔	✔	✔	✔
Petunidin-3-monoglucoside-acetate	✔	✔	✔	✔
Peonidin-3-monoglucoside-acetate	✔	✔	✔	✔
Delphinidin-3-monoglucoside-*p*-coumarate	✔	✔	✔	✔
Malvidin-3-monoglucoside-acetate	✔	✔	✔	✔
Peonidin-3-monoglucoside-caffeoate	✔	✔	✔	✔
Cyanidin-3-monoglucoside-*p*-coumarate	✔	✔	✔	✔
Malvidin-3-monoglucoside-caffeoate	✔	✔	✔	✔
Petunidin-3-monoglucoside-*p*-coumarate	✔	✔	✔	✔
Cis-malvidin-3-monoglucoside-*p*-coumarate	✔	✔	✔	✔
Peonidin-3-monoglucoside-*p*-coumarate	✔	✔	✔	✔
Malvidin-3-monoglucoside-*p*-coumarate	✔	✔	✔	✔
Total anthocyanins	✔	✔	✔	
Ratios				
Sum of coumarate/Sum acetate				✔
Sum trihydroxylated/Sum dihydroxylated				✔

2.8.1. Correlation analysis

Correlation between the three different measures of anthocyanin concentration (mg per kg, mg per berry and mg per litre) was evaluated

in order to assess the need to use more than one as a phenotype on genetic association analysis (Chapter IV). For the same reason, correlations between concentration (mg/kg), relative abundance (%) of anthocyanins and visual colour classifications were also assessed. All correlation analyses were performed with SAS v9.1 (SAS Institute Inc., Carry, NC, USA).

2.8.2. Principal component analysis

Principal component analysis (PCA) was performed using data on relative abundance (%) and concentration (mg/kg) of anthocyanins.

To study the variation among the sample of 149 cultivars based on anthocyanins concentration, PCA was applied to the correlation matrix between 19 variables (Table III-3) concerning individual and total anthocyanins concentration.

For the study of variation according to relative abundance of anthocyanins, PCA was performed based on the correlation matrix between the 18 variables of specific anthocyanins percentage of HPLC total peaks area (Table III-3).

The number of principal components necessary to sufficiently approach the sample dimensionality was determined considering the cumulative variance explained and the average root criterion, including components with eigenvalues higher than the average value (Jackson, 1991).

Graphical representations were based on standardised scores and variable loadings. Standardised scores were obtained by dividing each principal component score by the square root of the respective eigenvalue. Variable loadings were obtained by multiplying eigenvectors by the square root of the respective eigenvalue. In order to be

commensurate with score values, variable loadings were scaled before plotting, by multiplying each value by a constant.

SAS v9.1 (SAS Institute Inc., Carry, NC, USA) was used to perform these calculations.

2.8.3. Stepwise regression

To assess the impact of other variables, such as virus infection and maturity state on anthocyanin composition of cultivars, regressions between these variables were performed. Maturity state related features like berries brix degree, sugar content, volumic mass, probable alcohol and total acidity were regressed on total anthocyanin concentration. Viral infections, such as grapevine fanleaf virus, arabic mosaic virus, grapevine fleck virus, grapevine leafroll-associated viruses (GLRaV1, GLRaV2, GLRaV3 and GLRaV7) and grapevine virus B were regressed on the same concentration.

In order to evaluate the relationships between different visual classifications of berries colour and anthocyanins content, concentration and relative abundance of anthocyanins were regressed on visual colour characterisation variables (pulp colour, skin colour, SPC and SPC′). This information was mainly valuable for the interpretation of genetic association results on Chapter IV. All calculations were undertaken in SAS v9.1 (SAS Institute Inc., Carry, NC, USA).

2.8.4. Cluster analysis

Separate cluster analyses were performed using phenotypic data on anthocyanin concentration (mg/kg), relative abundance (%) and microsatellite data for 149 cultivars. For concentration in mg/kg, 18 specific anthocyanins and total anthocyanins concentrations were used as

variables. In the case of relative abundance, the variables were 18 specific anthocyanins relative abundances (Table III-3).

For molecular data, cluster analysis was based on 20 microsatellite loci scattered across 18 linkage groups provided by the Istituto Agrario San Michele all'Adige (IASMA). Two pairs of markers were on the same chromosome. However, according to previous linkage studies, the distance between them was at least 5cM (Doligez, 2006; Troggio *et al.*, 2007). Therefore, the genotyped loci were expected to behave independently. These markers were very informative with an average Polymorphic Information Content (PIC) of 0.7.

Euclidean distances between cultivars were obtained using the phenotypic data. This distance was calculated according to the following formula:

$$d_{EU} = \left[\sum_{i=1}^{n} (X_{ij} - X_{ik})^2 \right]^{\frac{1}{2}}$$

Where *j* and *k* are two points between which the distance is measured in the *n*-dimensional space.

Distances between cultivars based on the proportion of shared alleles were obtained using microsatellite data (Chakraborty and Jin, 1993). This distance was calculated according to the following formula:

$$d_{PSA} = 1 - \frac{\sum_{l=1}^{L} S}{2L}$$

Where S is the number of shared alleles and L is the number of total genotyped loci.

Phenotypic data was standardised by average subtraction and division by standard deviation. Unweighted Arithmetic Average Clustering (UPGMA) method (Sneath & Sokal, 1973) was used to cluster cultivars according to phenotypic and genotypic data.

In order to measure the goodness of fit between the original dissimilarity matrix and the tree, the cophenetic correlation was

calculated for each tree. The cophenetic correlation is a Pearson correlation between the elements of the original matrix and the cophenetic values. The cophenetic values are the minimal dissimilarities implied by the dendrogram (Sokal and Rohlf, 1962).

In order to estimate the statistical error associated with each tree, bootstrap analysis was performed. Proposed by Felsenstein (1985), this method is a means to approximate the underlying sampling distribution by resampling with replacements the original data set. Bootstrap values were obtained following 1000 permutations of traits and microsatellite markers. PAST software was used for phenotypic data while PowerMarker and Phylip consense were used for microsatellite data. Bootstrap values above 50 % were added to the original dendrograms.

However, this method assumes independence of the shuffled sets of observations. This assumption is clearly not met by different concentrations and relative abundances of anthocyanins. Also shuffling of cultivars would not overcome this limitation, since many are also partly related to each other (Chapter IV). Therefore, bootstrap values should be considered cautiously for the tree based on phenotype data.

Comparisons between genotypic distances and phenotypic distances were undertaken using Pearson correlation. Since the matrix elements are not independent bivariate observations as assumed by correlation theory, conventional significance levels for correlation coefficients are not applicable. Therefore, the test proposed by Mantel (1967) was used. In order to create an empirical distribution of the data, *n* rows and corresponding columns of one matrix were randomly rearrangement by 1000 permutations. Correlation coefficients were calculated for these permuted datasets. The coefficient observed with the original dataset was

ranked to calculate the empirical significance level by dividing the rank position by the total number of permuted coefficients.

$$Empirical\ P\text{-}value = \frac{rank\ of\ original\ correlation\ coefficient\ in\ the\ total\ permuted}{total\ number\ of\ permutations}$$

NTSYS software package (Rohlf, 2000) was used to perform all cluster analysis and matrix comparisons.

3. Results

3.1. Anthocyanin content measures

Pairwise correlations were calculated between mg of anthocyanins per litre of extract (mg/l), mg per berry (mg/berry) and mg per kilogram of berries (mg/kg). This was performed in order to assess the importance of considering different anthocyanin concentration measures for cultivar characterisation and genetic association analysis (Chapter IV). The correlation matrix is shown in Appendix 9.

The correlations between different anthocyanin concentration units were significant ($P < 0.0001$) and very high, with r values ranging from 0.90 to 0.99 (red circles and green triangles in Figure III-2). Due to this high correlation, concentration in mg/kg has been selected for cultivar characterisation and genetic association analysis (Chapter IV). Concentration measure in terms of mg/kg was selected rather than mg/l or mg/berry because its interpretation is clearer.

Relative abundance and concentration (mg/kg) of anthocyanins were also correlated for most cases. However, r values were not as high as between different concentration measures (blue diamonds in Figure III-2, where the dark are significant and the light are non significant). Significant ($P < 0.01$) correlations between concentration (mg/kg) and relative abundance of anthocyanins showed coefficients between 0.33

and 0.94. Relative abundance and concentration (mg/kg) of cyanidin-3-monoglucoside did not correlate significantly (r = -0.001). This anthocyanin is an extreme example of distinct information obtained by measuring relative abundance and concentration.

Total anthocyanins concentration correlated significantly with each specific anthocyanin concentration, except with the rare ones (peonidin-3-monoglucoside-caffeoate, malvidin-3-monoglucoside-caffeoate and cis-malvidin-3-monoglucoside-*p*-coumarate). However, considering relative abundance, only near 40 % of the anthocyanins correlated ($P > 0.01$) with total concentration.

Total anthocyanins concentration was significantly ($P < 0.01$) negatively correlated with the ratio between coumarate and acetate derivatives, even though the r values were low (-0.3). On the other hand, trihydroxylated/dihydroxylated anthocyanins ratio was not significantly correlated with total anthocyanins concentration (r = 0.018).

Considering the low correlation coefficients between relative abundance and concentration of anthocyanins, it was decided to consider both for cultivar characterisation and genetic association analysis (Chapter IV). The correlation matrix is shown on Appendix 10.

3.2. Diversity of anthocyanin composition

Table III-4 shows the moments for the identified anthocyanins, total anthocyanins, sums of acylation types, ratio of coumarate/acetate derivatives and tri/dihydroxylated anthocyanins. The moments are shown for data expressed in concentration (mg/kg) and relative abundance. Also frequencies are shown to highlight the rarest and most common pigments.

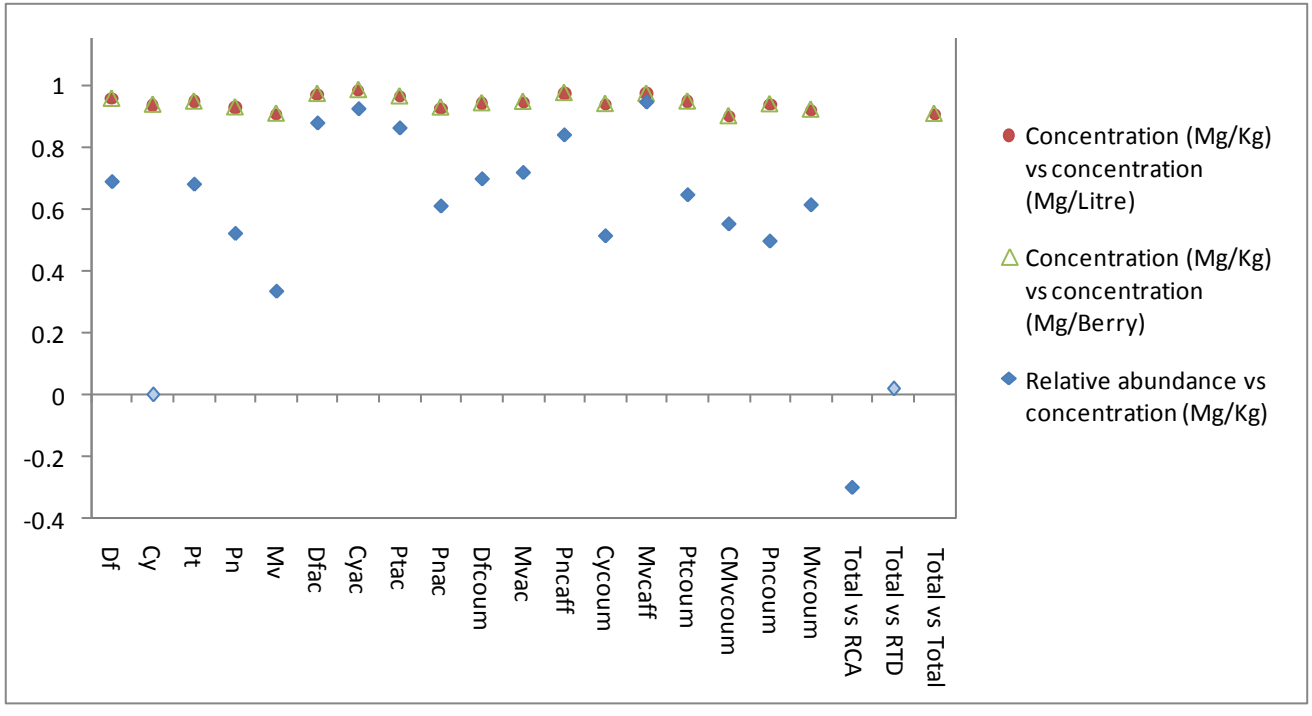


Figure III-2 Correlation coefficients plot between different anthocyanins concentration units and peak area. The x axis shows the anthocyanins, ratios or total concentrations with plotted values. RTD means ratio between sum of trihydroxylated and dihydroxylated anthocyanins; RCA means ratio between sum of coumarate and acetate derivative anthocyanins.

Total anthocyanins concentration in mg/kg had a minimum of 1.3 mg/kg (cultivar Chasselas Roxo), a maximum of 2643.3 mg/kg (cultivar Tinta Ferreira), with average 555 mg/kg and standard deviation of 471.5 (Table III-4).

Concerning relative abundance of anthocyanins, the predominant pigments were the glucosides ranging from 38.07 % of the total peak area to 100 %, with an average of 77.5 % (SD = 13.59). Exceptions to this were cultivars Tinta Pomar and Tinto Cão which have a higher percentage of coumarate derivatives (Figure III-3).

In general, coumarate derivative anthocyanins were the second most abundant with an average of 18.46 %. Exceptions were cultivars Espadeiro Mole and Trollinger where the percentage of acetate derivatives exceeded coumarate (Figure III-3). Caffeoate derivatives were the rarest ones, absent in 81.21 % of the cultivars. When present, these pigments were always the less abundant showing small areas with a maximum of 0.17 % (Table III-4; Figure III-3).

In most cases (81.8 %), malvidin-3-monoglucoside was the predominant pigment. This anthocyanin was in average 39.4 % of the total anthocyanins measured in a cultivar. Peonidin-3-monoglucoside was the second most abundant with an average of 14.5 %. It was the main pigment in 14 cultivars (9.4 %). Some exceptions occurred where the most abundant pigment was delphinidin-3-monoglucoside (cultivars Tinto Velasco[1], Vinhão and Espadeiro Mole), cyanidin-3-monoglucoside (cultivars Uva Moranga[1], Verdelho Roxo, Ahmeur bou Ahmeur, Imperial Rojo, Malvasia Fina Roxa, Folgasão Roxo, Chasselas Roxo and Gewürztraminer) and malvidin-3-monoglucoside-*p*-coumarate (cultivars Tinto Cão and Tinta Pomar) (Figure III-4; Table III-4).

[1] Cultivar identification unconfirmed.

Table III-4 Summary of moments and frequency of anthocyanins, sums and ratios.

Anthocyanin, sums and ratios	Concentration in mg per kg of berries				HPLC peak area				Freq.
	Mean	St. Dev.	Min.	Max.	Mean	St. Dev.	Min.	Max.	
Delphinidin-3-monoglucoside	58.830	103.102	0.000	913.168	8.399	6.231	0.000	37.340	97.32
Cyanidin-3-monoglucoside	19.289	30.628	0.300	227.928	7.941	19.786	0.164	100.000	100.00
Petunidin-3-monoglucoside	49.809	66.785	0.000	431.829	7.304	3.854	0.000	21.347	95.97
Peonidin-3-monoglucoside	79.161	87.978	0.000	443.542	14.463	12.521	0.000	65.074	96.64
Malvidin-3-monoglucoside	226.131	181.875	0.000	877.471	39.393	13.661	0.000	78.236	95.97
Delphinidin-3-monoglucoside-acetate	2.510	6.201	0.000	49.489	0.295	0.547	0.000	3.634	59.06
Cyanidin-3-monoglucoside-acetate	0.626	2.548	0.000	29.384	0.067	0.192	0.000	1.983	35.57
Petunidin-3-monoglucoside-acetate	2.605	5.389	0.000	37.824	0.331	0.517	0.000	3.451	63.76
Peonidin-3-monoglucoside-acetate	3.006	4.020	0.000	18.797	0.525	0.705	0.000	5.853	83.22
Delphinidin-3-monoglucoside-*p*-coumarate	7.149	10.137	0.000	66.831	1.100	0.960	0.000	5.202	85.91
Malvidin-3-monoglucoside-acetate	15.338	20.958	0.000	130.103	2.723	2.904	0.000	15.236	87.92
Peonidin-3-monoglucoside-caffeoate	0.024	0.093	0.000	0.574	0.004	0.016	0.000	0.106	8.05
Cyanidin-3-monoglucoside-*p*-coumarate	2.591	4.029	0.000	27.514	0.474	0.473	0.000	2.988	89.26
Malvidin-3-monoglucoside-caffeoate	0.060	0.176	0.000	0.894	0.010	0.028	0.000	0.140	13.42
Petunidin-3-monoglucoside-*p*-coumarate	8.030	9.655	0.000	59.233	1.370	1.019	0.000	4.634	90.60
Cis-Malvidin-3-monoglucoside-*p*-coumarate	1.626	1.457	0.000	7.877	0.390	0.389	0.000	2.181	84.56
Peonidin-3-monoglucoside-*p*-coumarate	16.834	17.356	0.000	131.832	3.223	2.142	0.000	10.456	93.96
Malvidin-3-monoglucoside-*p*-coumarate	60.471	56.204	0.000	296.701	11.907	9.062	0.000	47.409	93.96
Total anthocyanins	554.992	471.497	1.306	2643.29	_	_	_	_	_
Sum of monoglucosides	433.220	403.682	1.306	2248.90	77.501	13.585	38.068	100.000	100.00
Sum of acetate derivatives	24.085	35.251	0.000	195.239	3.940	4.266	0.000	22.868	90.60
Sum of coumarate derivatives	96.702	87.017	0.000	467.010	18.463	11.734	0.000	57.493	94.63
Sum of caffeoate derivatives	0.084	0.208	0.000	1.078	0.014	0.035	0.000	0.174	18.79
Ratio									
Sum of coumarate/Sum acetate	_	_	_	_	8.002	6.948	0.415	44.141	_
Sum trihydroxylated/Sum dihydroxylated	_	_	_	_	4.977	3.886	0.000	19.732	_

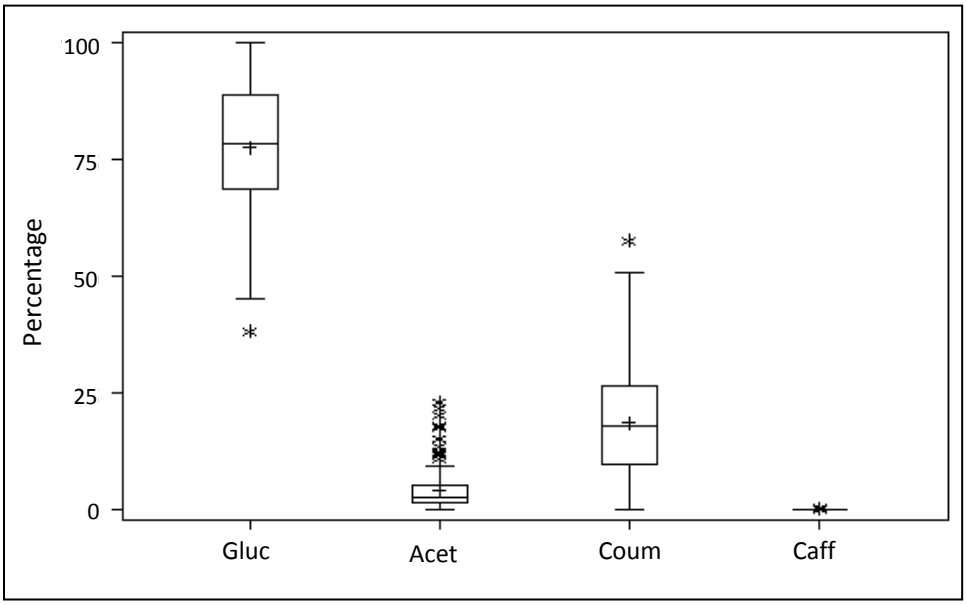


Figure III-3 Box and Whiskers plot showing the distribution of the percentage of the different anthocyanin groups according to acylation types. The acylation types are identified on the x axis. *Gluc*: no acylation; *Acet*: acetate derivatives; *Coum*: coumarate derivatives; *Caff*: caffeoate derivatives.

The predominant acetate derivative pigment was peonidin with an average of 0.53 % (SD = 0.71). Among the coumarate derivative anthocyanins, malvidin was the most abundant with an average of 11.9 % (SD = 9.06). Concerning caffeoate derivative anthocyanins, despite the low frequency and abundance, malvidin-caffeoyl was the most abundant (average 0.01 %, SD = 0.028).

The ratio between tri and dihydroxylated anthocyanins was quite diverse, ranging from zero to 19.73, with average 5.0 (SD = 3.89). The coumarate/acetate derivatives ratio range was even wider, between 0.42 and 44.14, with an average of eight (SD = 6.95) (Table III-4).

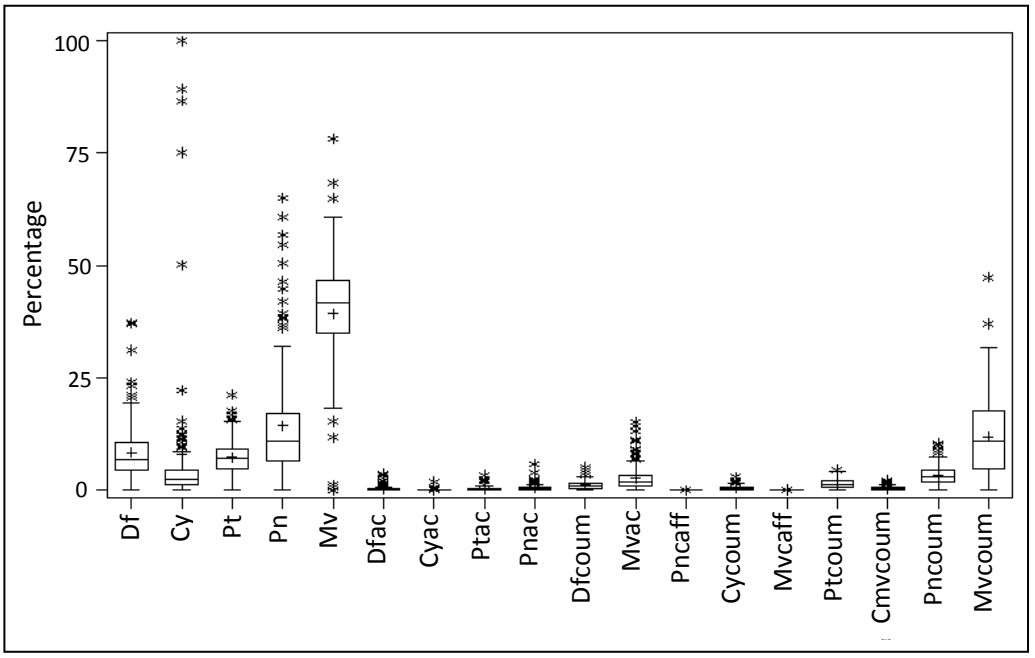


Figure III-4 Box and Whiskers plot showing the distribution of the percentage of the different anthocyanins.

3.3. Principal component analysis based on anthocyanins concentration

The principal component analysis based on the correlation matrix between 18 specific anthocyanins and total anthocyanins concentration in mg/kg showed the first five principal components to explain 86.1 % of the total variation (Table III-5) and to have eigenvalues greater than the average (0.99) .

Figure III-5 shows a bidimensional plot of principal components 1 and 2 of anthocyanin concentration data in mg/kg. This plot shows cultivar sample scores (black) and variable loadings (colour). The first principal component (PC1) explained 46.1 % of total variance. The majority of the variables had high loadings on this component. This means they

correlated highly with this component and therefore contributed strongly to this component. Along PC1, cultivars were separated based on their anthocyanins concentration, both specific and total anthocyanins (Figure III-5), although the latter had the highest loading (0.915). Therefore, cultivars with high concentrations of total and specific anthocyanins tended to have high scores in PC1. This shows that the main distinguishing trait among the cultivars is anthocyanins concentration, especially total anthocyanins.

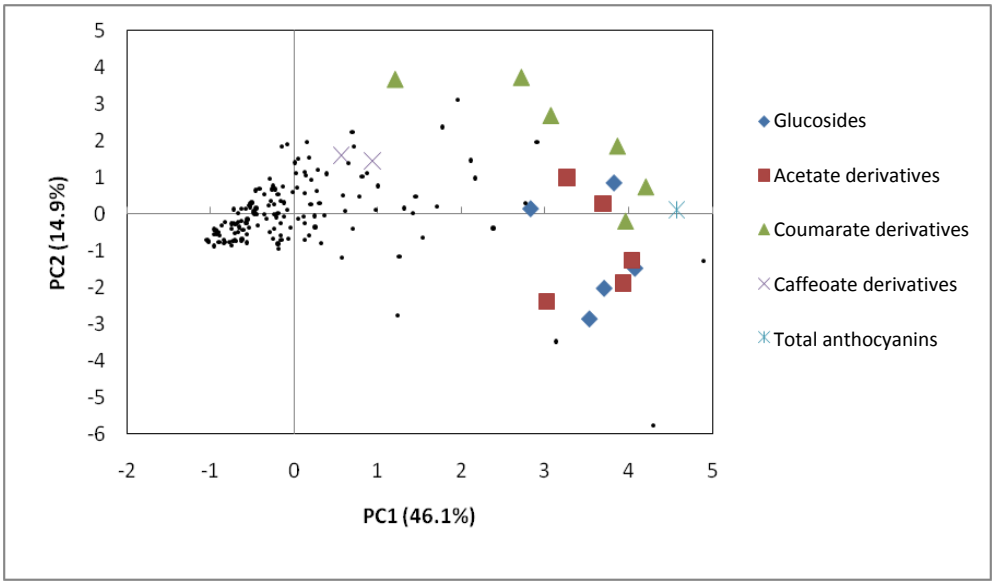


Figure III-5 Bidimensional plot of principal components 1 and 2 of anthocyanin concentration data in mg/kg. The variable loadings are scaled by a factor of 5 on the plane defined by principal components 1 and 2 in order to be commensurate with score values. The variable loadings are represented by coloured symbols and the samples scores by black dots.

Since the plane defined by PC1 and PC2 showed only cultivar differences based on concentration, it was decided to assess how other components would separate cultivars. Figure III-6 shows two identical plots of PC2 and PC3 showing sample scores (black) and variable

loadings (colour). These plots differ in the highlighted properties of the plotted variables. In plot A, variables are coloured according to the anthocyanidin type. In plot B, variables are coloured according to acylation type.

Table III-5 Variable loadings on first three principal components of anthocyanin concentration (mg/kg) data.

Variables	Principal Components		
	PC1	PC2	PC3
Delphinidin-3-monoglucoside	0.741	-0.404	0.268
Cyanidin-3-monoglucoside	0.706	-0.571	0.244
Petunidin-3-monoglucoside	0.815	-0.295	0.313
Peonidin-3-monoglucoside	0.565	0.029	0.640
Malvidin-3-monoglucoside	0.765	0.169	0.398
Delphinidin-3-monoglucoside-acetate	0.786	-0.378	-0.428
Cyanidin-3-monoglucoside-acetate	0.604	-0.477	-0.337
Petunidin-3-monoglucoside-acetate	0.808	-0.254	-0.471
Peonidin-3-monoglucoside-acetate	0.738	0.055	-0.229
Delphinidin-3-monoglucoside-*p*-coumarate	0.841	0.151	-0.134
Malvidin-3-monoglucoside-acetate	0.652	0.199	-0.503
Peonidin-3-monoglucoside-caffeoate	0.188	0.319	0.327
Cyanidin-3-monoglucoside-*p*-coumarate	0.792	-0.036	-0.066
Malvidin-3-monoglucoside-caffeoate	0.113	0.291	-0.002
Petunidin-3-monoglucoside-*p*-coumarate	0.773	0.373	-0.164
Cis-Malvidin-3-monoglucoside-*p*-coumarate	0.241	0.738	-0.201
Peonidin-3-monoglucoside-*p*-coumarate	0.613	0.541	0.279
Malvidin-3-monoglucoside-*p*-coumarate	0.543	0.749	-0.147
Total anthocyanins	0.915	0.023	0.342
Eigenvalues	8.767	2.832	2.035
% of variance	46.1%	14.9 %	10.7 %

PC2 and PC3 explained respectively 14.9 % and 10.7 % of total variation. PC2 separated cultivars based on anthocyanidin type, while on PC3 this was based on acylation pattern (Figure III-6). Figure III-6 shows the variables ordered by cyanidin, delphinidin, petunidin, peonidin and malvidin derivatives along PC2. This order corresponded to an increase in methylation ranging from non-methylated anthocyanidins (cyanidin and delphinidin) to anthocyanidins with one (petunidin and peonidin) and two methyl groups (malvidin). With PC3, variables with acetate and coumarate derivatives correlated negatively and variables with non acylated anthocyanins correlated positively. The only exception was peonidin-3-monoglucoside-*p*-coumarate. Total anthocyanins also correlated positively with PC3, since glucosides are often the most abundant anthocyanins (Figure III-6).

Very similar results were obtained from a principal component analysis based on the log transformed data. PCA based on the covariance matrix was also investigated. Although the different variables units were the same, variances were very distinct. Since no biological interpretation could be derived from these variance differences, it was concluded that the use of the correlation matrix was more adequate.

3.4. Principal component analysis based on relative abundance of anthocyanins

The principal component analysis based on the correlation matrix between relative abundance of 18 anthocyanins, showed that the first 6 principal components explained 83.7 % of the total variation and had eigenvalues greater than the average (1.00) (Table III-6).

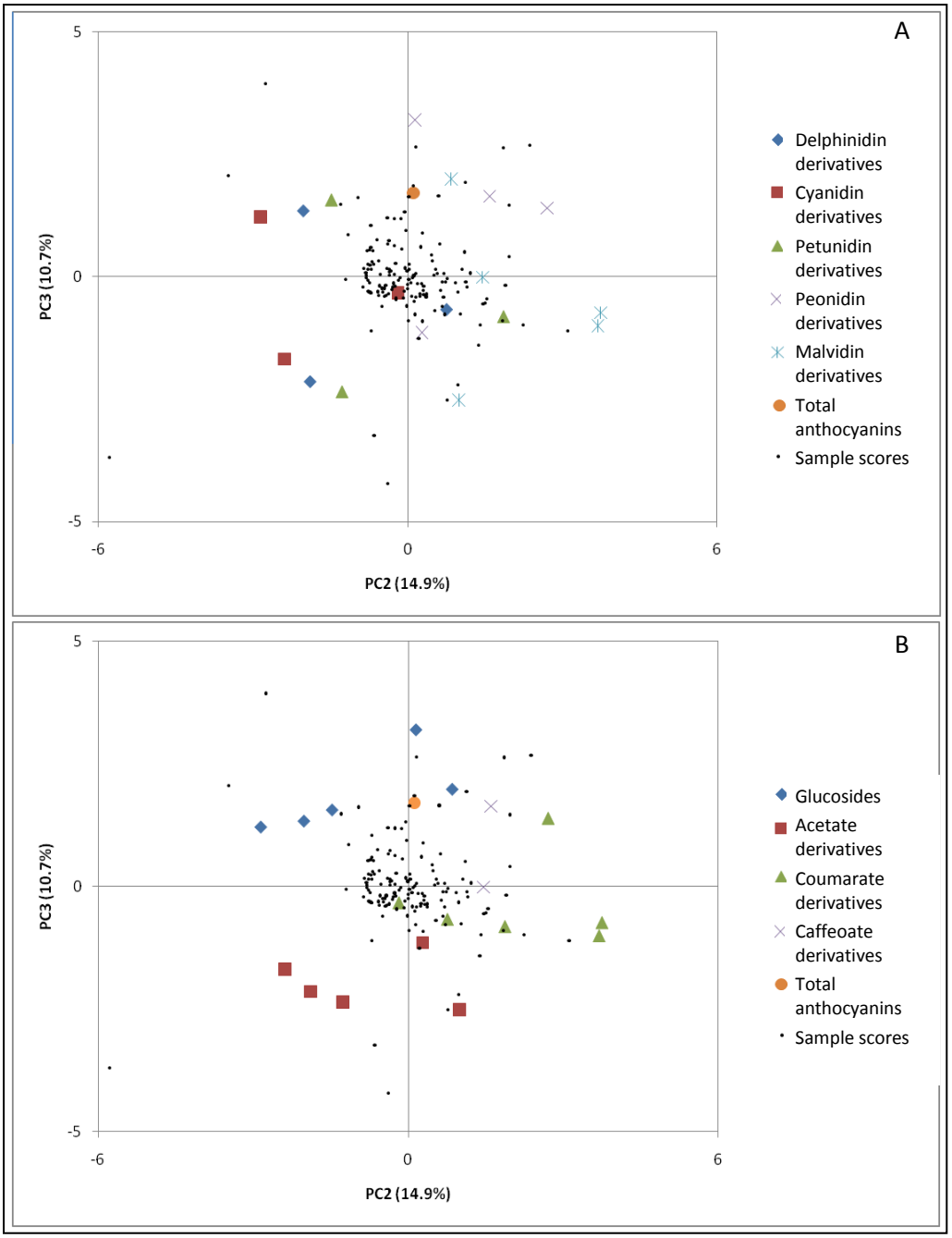


Figure III-6 Bidimensional plots of principal components 2 and 3 of anthocyanin concentration data in mg/kg. In plot A, variable loadings are identified according to anthocyanidin type. In plot B, variable loadings are identified according to acylation pattern. Variable loadings are represented by coloured symbols and samples scores by black dots. Cultivar sample scores and variable loadings are scaled by a factor of 5.

Table III-6 Variable loadings on first three principal components of relative abundance of anthocyanins.

Variables	Principal Components		
	PC1	PC2	PC3
Delphinidin-3-monoglucoside	0.354	-0.670	-0.449
Cyanidin-3-monoglucoside	-0.448	-0.389	0.158
Petunidin-3-monoglucoside	0.467	-0.463	-0.547
Peonidin-3-monoglucoside	-0.501	-0.128	0.400
Malvidin-3-monoglucoside	0.174	0.403	-0.316
Delphinidin-3-monoglucoside-acetate	0.774	-0.454	0.325
Cyanidin-3-monoglucoside-acetate	0.533	-0.488	0.397
Petunidin-3-monoglucoside-acetate	0.812	-0.342	0.346
Peonidin-3-monoglucoside-acetate	0.458	0.040	0.720
Delphinidin-3-monoglucoside-*p*-coumarate	0.741	0.052	-0.411
Malvidin-3-monoglucoside-acetate	0.703	0.252	0.431
Peonidin-3-monoglucoside-caffeoate	-0.114	0.088	0.351
Cyanidin-3-monoglucoside-*p*-coumarate	0.269	-0.233	-0.101
Malvidin-3-monoglucoside-caffeoate	0.115	0.204	0.162
Petunidin-3-monoglucoside-*p*-coumarate	0.683	0.389	-0.362
Cis-Malvidin-3-monoglucoside-*p*-coumarate	0.253	0.801	-0.071
Peonidin-3-monoglucoside-*p*-coumarate	0.078	0.627	0.385
Malvidin-3-monoglucoside-*p*-coumarate	0.399	0.836	-0.123
Eigenvalues	4.415	3.611	2.502
% of variance	24.5 %	20.1 %	13.9 %

Figure III-7 shows two identical plots of PC1 and PC3 showing sample scores and variable loadings. These plots differ in the highlighted properties of the plotted variables. In plot A, variables are coloured according to the anthocyanidin type. In plot B, variables are coloured according to acylation type.

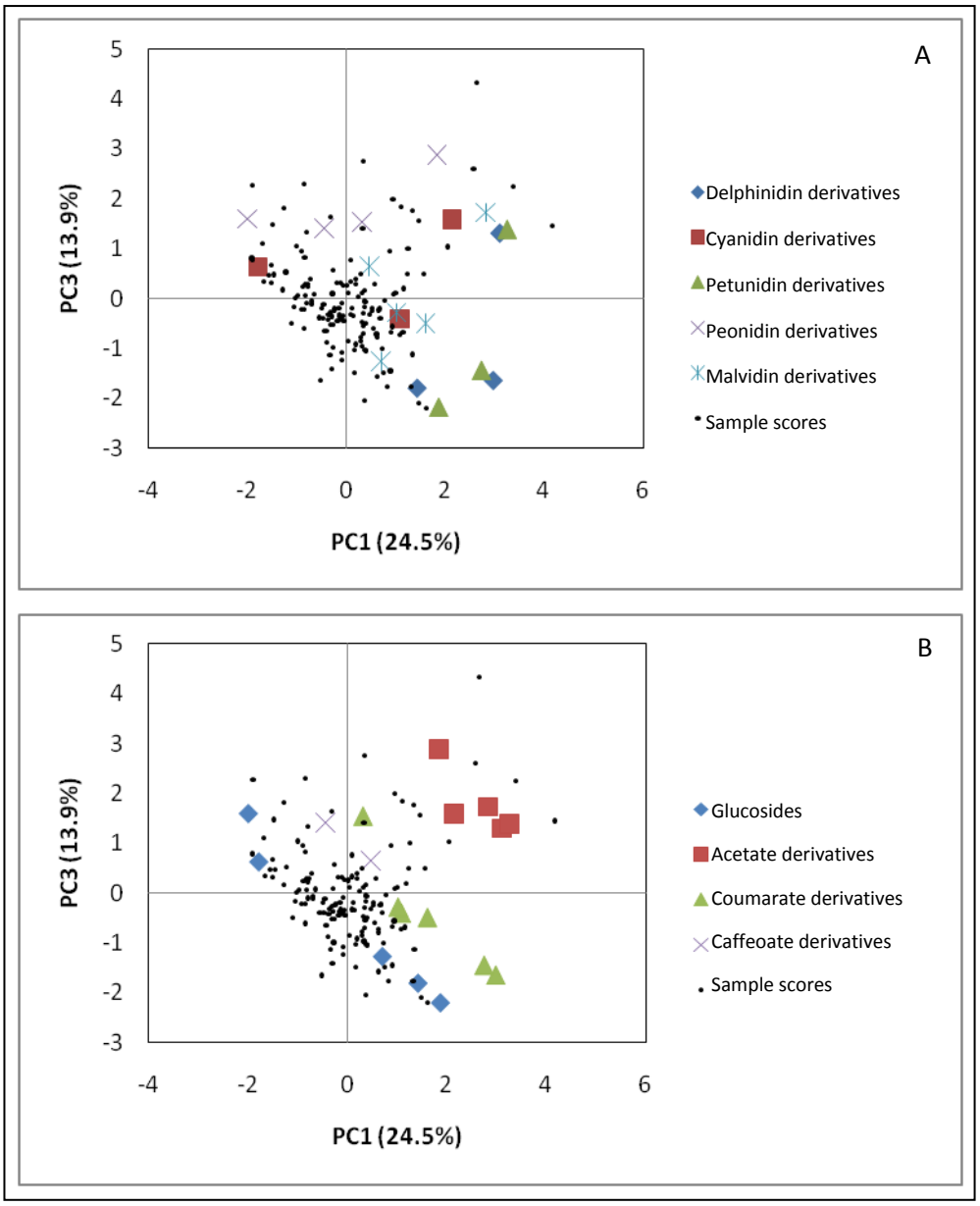


Figure III-7 Bidimensional plot of principal components 1 and 3 of anthocyanin relative abundance. In plot A, variable loadings are identified according to anthocyanidin type. In plot B, variable loadings are identified according to acylation pattern. Variable loadings are represented by coloured symbols and sample scores by black dots. Sample scores and the variable loadings are scaled by a factor of 4.

PC1 and PC3 explained respectively 24.5 % and 13.9 % of the total variance. Both components separated cultivars according to anthocyanidin type and acylation pattern. Along both axis variables were ordered by peonidin, cyanidin, malvidin, delphinidin and petunidin in opposite directions. This order corresponded to an increase in hydroxylation ranging from dihydroxylated anthocyanins (peonidin and cyanidin) to trihydroxylated anthocyanins (malvidin, delphinidin and petunidin). Variables were also ordered along both axes by non acylated, acetate derivatives and coumarate derivatives (Figure III-7).

The principal component analysis based on the log transformed data showed very similar results. Similarly to the PCA of anthocyanin concentration data, the use of a covariance matrix was also investigated. Once again, due to very different variances with no associated biological interpretation it was concluded that the use of the correlation matrix was more adequate.

3.5. Cluster analysis based on anthocyanins concentration

The dendrogram obtained using mg/kg had a very good-fit with a cophenetic coefficient of 0.95. The tree did not show very clear clusters (Figure III-8). Nevertheless, it was possible to identify two subdivisions I and II where dissimilaritites between cultivars were smaller. These two clusters were identified in the principal component analysis by plotting sample scores on PC2 and PC3 (Figure III-9). Blue and yellow were the colours used to identify clusters I and II respectively in Figures III-8 and III-9. These clusters were separated along PC2. Cluster I included cultivars with higher concentrations of peonidin and malvidin-3-monoglucoside-acetate and slightly higher concentrations of coumarate derivative anthocyanins. Cluster II included mainly cultivars with higher

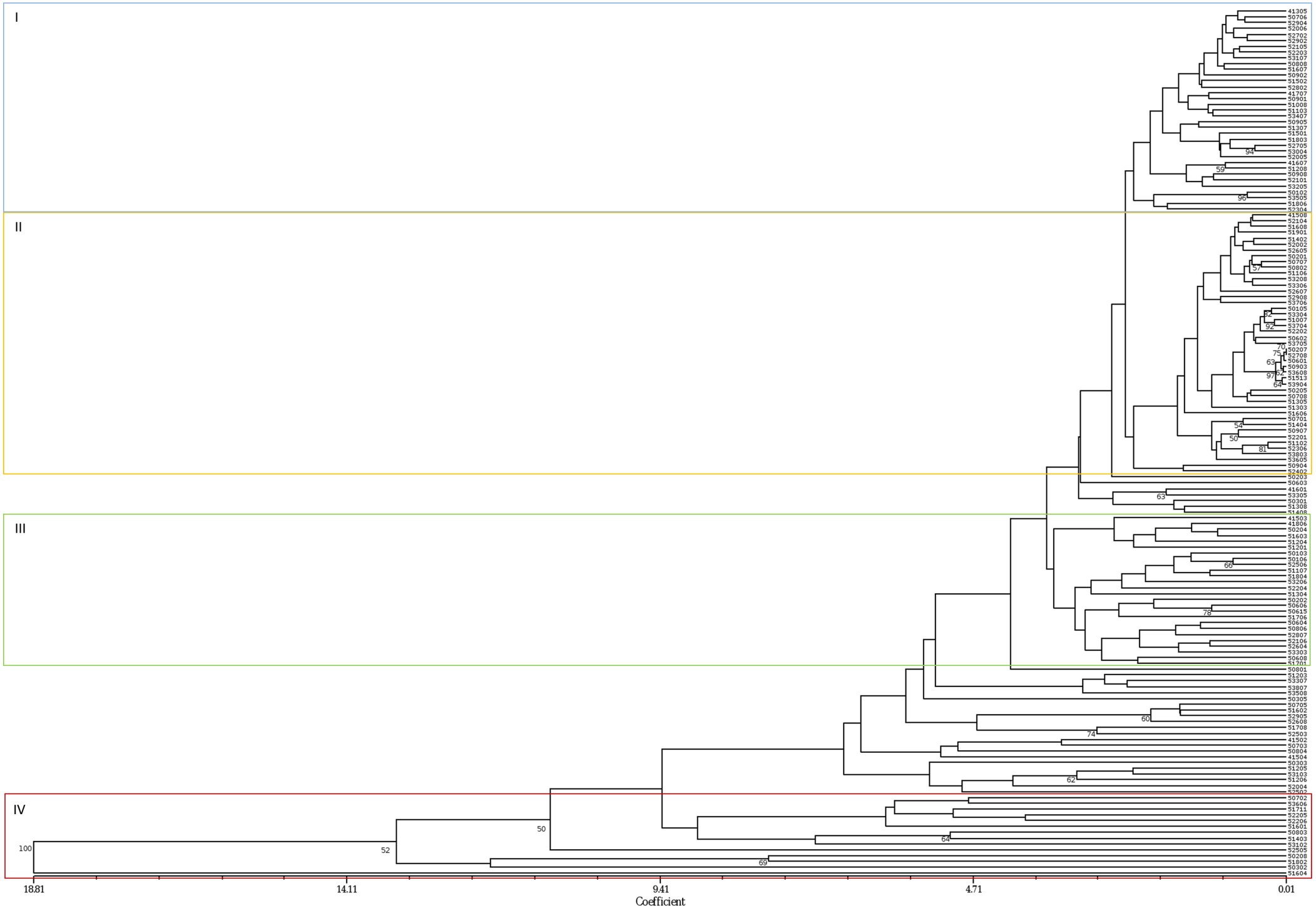


Figure III-8 Cluster analysis dendrogram of 149 cultivars based on anthocyanin concentration in mg/kg. Bootstrap values above 50 % are shown.

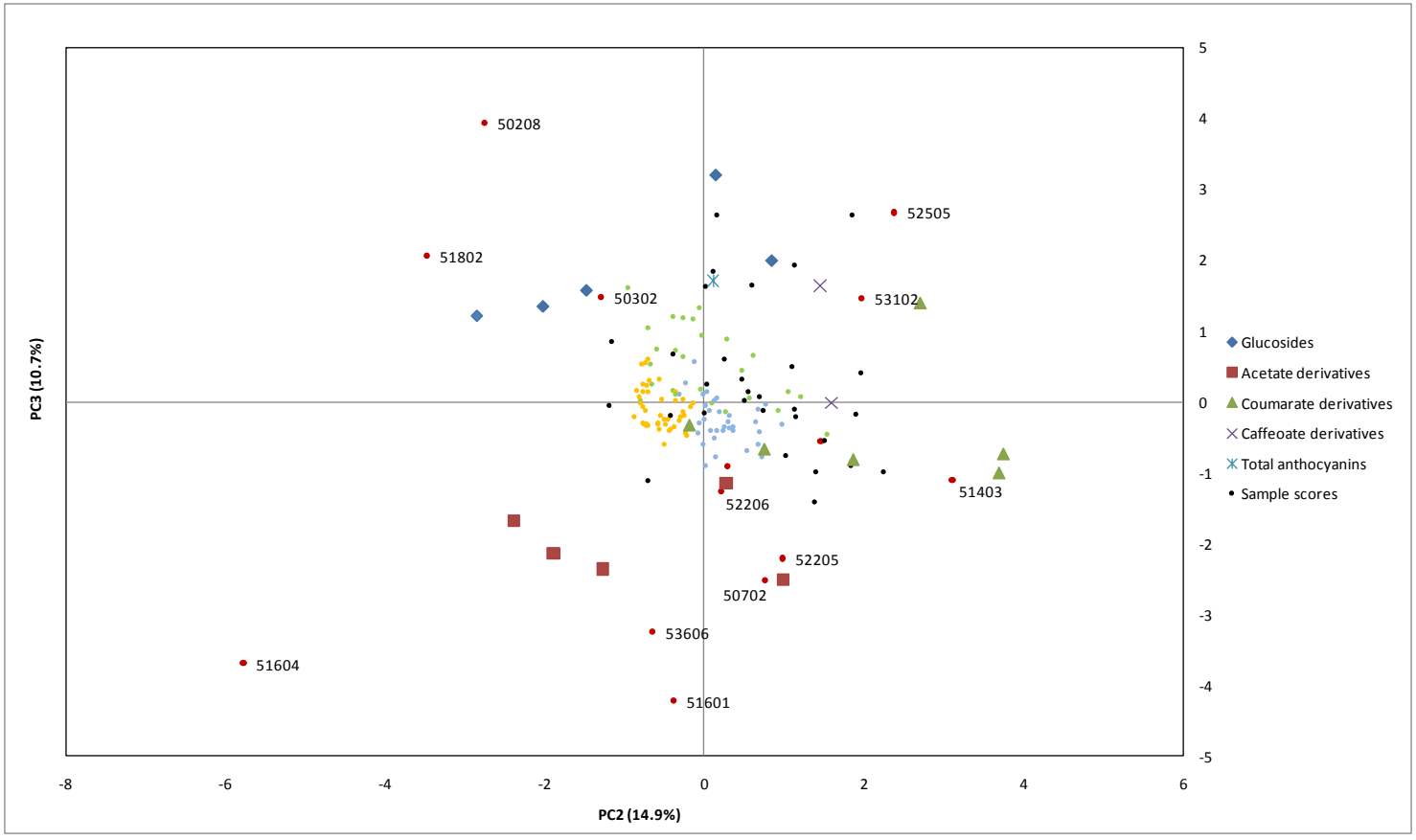


Figure III-9 Bidimensional plot of principal components 2 and 3 of anthocyanin concentration data in mg/kg with sample scores identified according to UPGMA clusters. The variable loadings are scaled by a factor of 5 on the plane defined by principal components 2 and 3 in order to be commensurate with score values. The sample scores are represented by dots coloured according to UPGMA clusters. Blue: cluster I; yellow: cluster II; green: cluster III. The variable loadings are represented by the remaining shapes.

concentrations of glucosides and acetate derivatives of cyanidin, delphinidin and petunidin.

The remaining branches of the tree showed several minor clusters and outliers. Among these, a larger cluster (III) was identified despite having higher dissimilarities between cultivars than clusters I and II. The cultivars in cluster III (marked in green in Figure III-8) were more dispersed along the axes of PC2 and PC3. However, it was PC3 that separated the cultivars in this cluster from the ones in clusters I and II. The cultivars in cluster III were mainly characterised by smaller concentrations of acetate derivatives.

The most dissimilar cultivars grouped together (cluster IV) showing the longer branches in the tree. The cultivars included in this cluster were identified in the PCA plot by the red scores. The most isolated cultivars had their accession number shown in the graph. The cultivars in this cluster located mostly in the outskirts of the scores cloud. The remaining cultivars included mainly small clusters and outliers.

3.6. Cluster analysis based on relative abundance of anthocyanins

The clustering based on relative abundance of anthocyanins had a slightly lower goodness-of-fit, with a cophenetic correlation of 0.88. The tree did not show very clear clusters (Figure III-10). Nevertheless, it was possible to identify five clusters. Figure III-11 shows clusters I to V and identifies cultivars on the PCA plot. Cluster I (blue) included mainly cultivars with higher abundance of malvidin-3-monoglucoside and peonidin and malvidin-3-monoglucoside-*p*-coumarate. Cluster II (yellow) was characterised by higher abundance of dihydroxylated glucosides. Cluster III (green) was mostly formed by cultivars with higher abundance of acetate derivative anthocyanins, coumarate derivatives of delphinidin, petunidin and cyanidin and monoglucosides of petunidin and delphinidin.

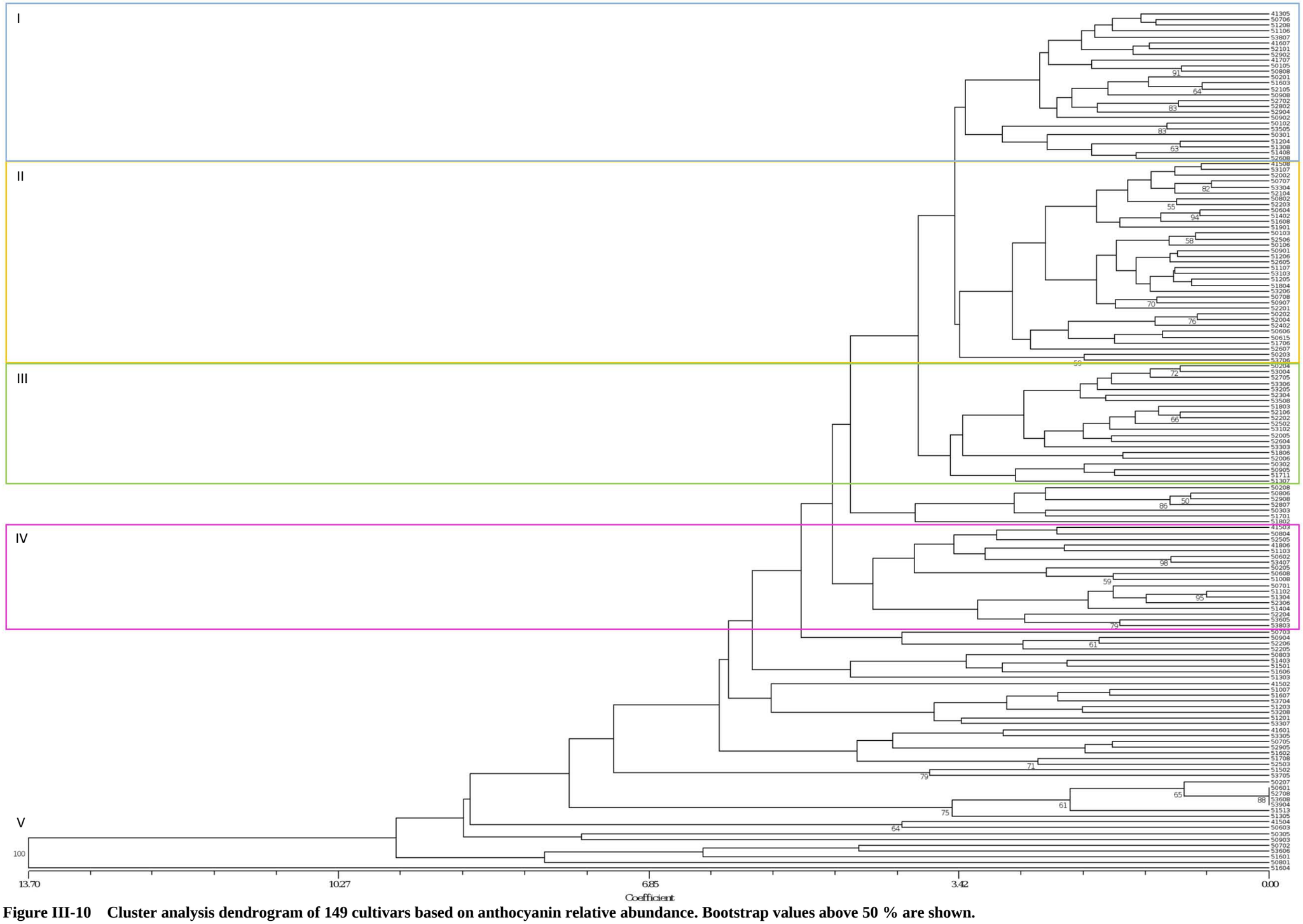


Figure III-10 Cluster analysis dendrogram of 149 cultivars based on anthocyanin relative abundance. Bootstrap values above 50 % are shown.

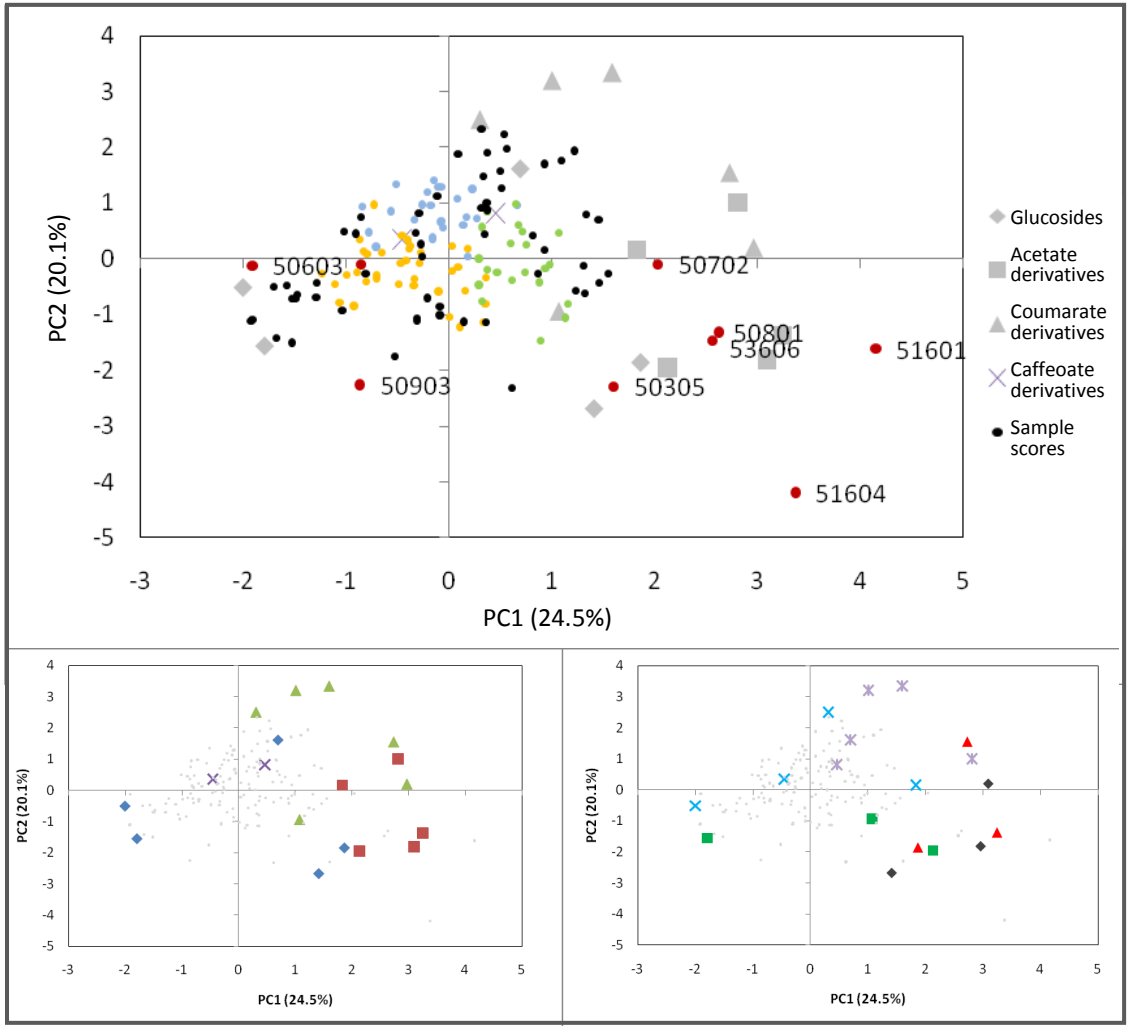


Figure III-11 Bidimensional plot of principal components 1 and 2 of anthocyanin relative abundance data with sample scores identified according to UPGMA clusters. On top, the plot highlights sample scores according to UPGMA cluster. Blue: cluster I, yellow: cluster II, green: cluster III, red: cluster V. Variable loadings are in grey.
In the bottom the two bidimensional plots of the same principal components, only highlighting the variable loadings (in colour) instead of sample scores (in grey). On the left, variable loadings have different colours and shapes according to acylation pattern and on the right, according to anthocyanidin type.

Legend for bottom left plot
- Glucosides
- Acetate derivatives
- Coumarate derivatives
- Caffeoate derivatives
- Sample scores

Legend for bottom right plot
- Delphinidin derivatives
- Cyanidin derivatives
- Petunidin derivatives
- Peonidin derivatives
- Malvidin derivatives
- Sample scores

A forth cluster was also identified. To distinguish the cultivars in this cluster from the remaining ones the sample scores of PC1 and PC3 were plotted (Figure III-12). This cluster was characterised by higher abundance of dihydroxylated glucosides as happened with cluster II. However, in this case the difference in abundance of these compounds was stronger than in cluster II.

3.7. Comparison of genotypic and phenotypic distances

The distance matrix based on the proportion of shared alleles did not correlate with Euclidean distances matrix based on the anthocyanins concentration. Comparison between the Euclidean distances matrix based on relative abundance of anthocyanins and the distance matrix based on proportion of shared alleles yielded a significant correlation (P = 0.001) despite a low coefficient (r = 0.22). Figure III-13 shows a scatter plot of the Euclidean distance values calculated from relative abundance of anthocyanins and distance values based on the proportion of shared alleles. Low distances based on proportion of shared alleles were associated with low phenotypic distances only. However, large distances based on proportion of shared alleles were associated with low and high phenotypic distances. Therefore, in some cases low phenotypic distances did not correspond to low distances based on allelic information.

Figure III-14 shows graphical representations of some examples of relative abundance of anthocyanins on cultivars. Charts in Box A show some cultivars of the different clusters that were identified in the cluster analysis. Pie charts in Box B show cultivars with the maximum area of each specific non acylated anthocyanin. Finally, charts in Box C show cultivars with extreme relative abundances of anthocyanins according to acylation patterns. Pie charts for all the studied cultivars are on Appendix 11.

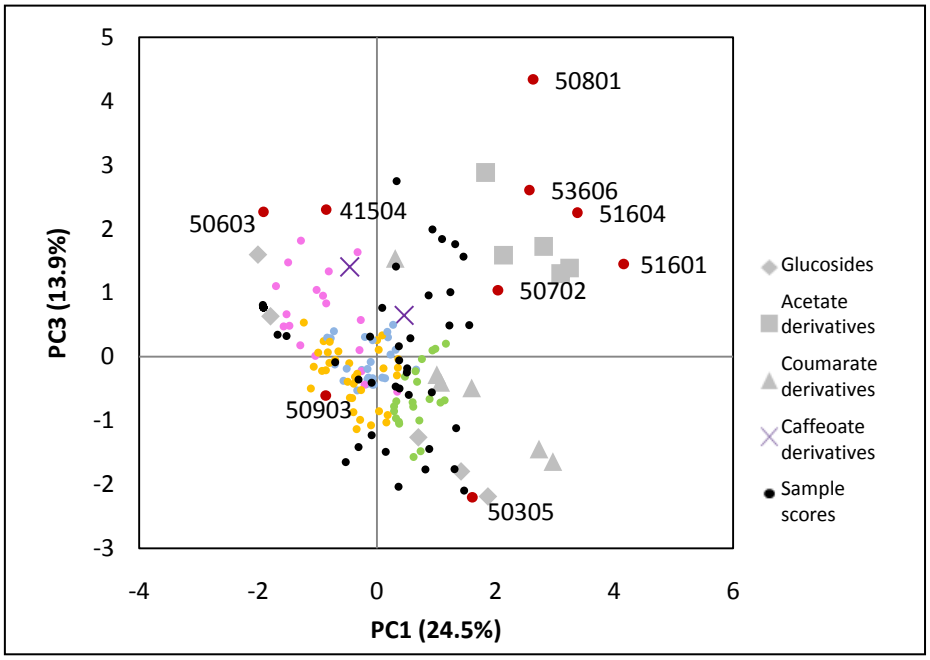
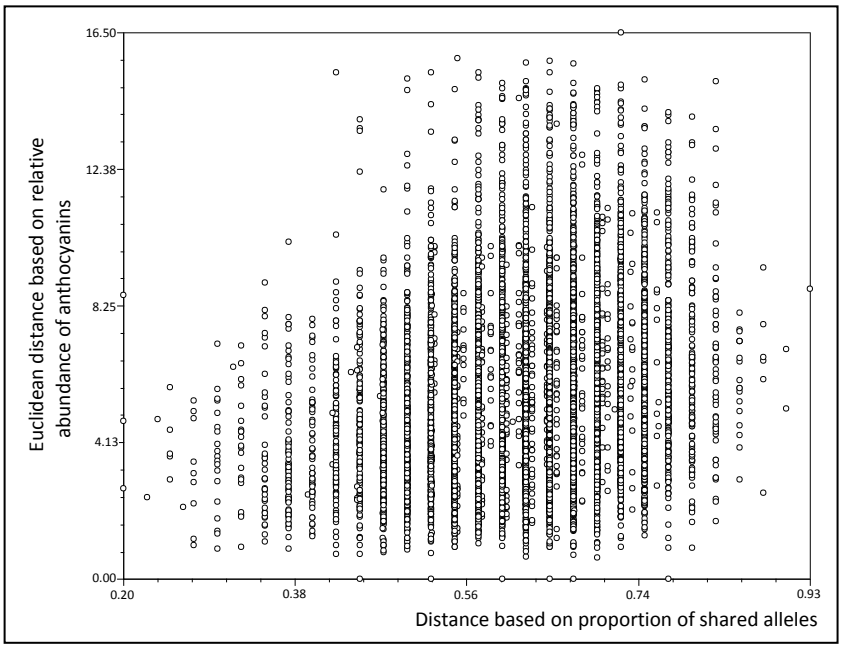


Figure III-12 Bidimensional plot of principal components 1 and 3 of relative abundance of anthocyanins with sample scores identified according to UPGMA cluster. Variables are in grey and sample scores are Blue: cluster I, yellow: cluster II, green: cluster III, pink: cluster IV, red: cluster V.



FigureIII-13 Plot of distances based on proportion of shared alleles and relative abundance of anthocyanins for 149 cultivars.

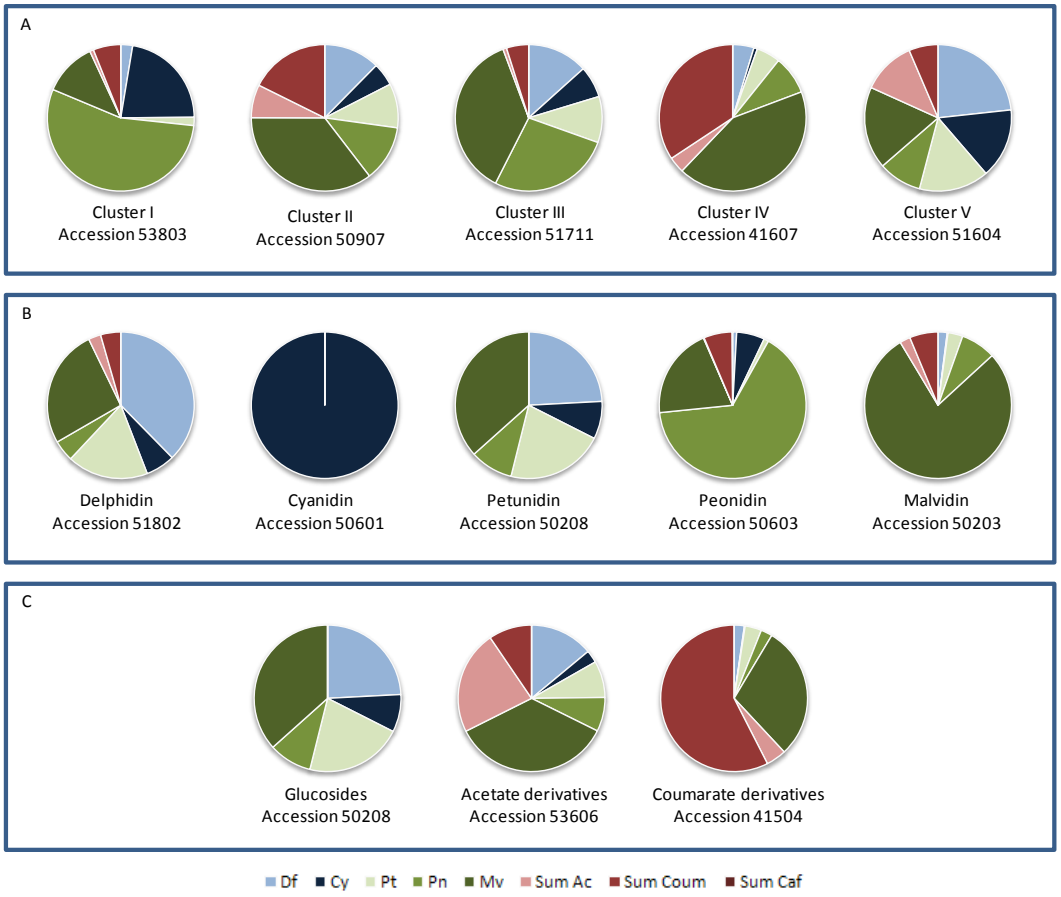


Figure III-14 Graphical representation of cultivars relative abundance of anthocyanins (%). Each pie chart represents one cultivar and the percentages of anthocyanin identified in it. Charts group A shows five cultivars, each belonging to a different cluster identified in the cluster analysis. Charts group B shows five cultivars, each with the maximum percentage of each glucoside anthocyanin. Charts group C shows 3 cultivars, each with the higher percentage of one acylation type.

3.8. Anthocyanins content potential covariates

3.8.1. Virus infection

Stepwise regression for total anthocyanin concentration (mg/kg) on virus infection state by GFLV, ArMV, GFKL, GLRaV1, GLRaV2, GLRaV3, GLRaV7, GVB did not show any significant association at 1 % level for the sample studied. However, GFLV, GFKL and GLRaV2 were

significant at 5 % level. The *P*-values for these tests are shown in Appendix 12.

3.8.2. Berry maturation parameters

Stepwise regression for total anthocyanin concentration (mg/kg) on parameters related with maturity, brix, sugar, volumic mass, probable alcohol and total acidity did not show significant associations for the sample studied (Appendix 12).

3.9. Berry colour visual characterisation

Correlation analyses between different visual characterisations of berries skin and pulp showed that pulp colour (PC) was more strongly correlated with concentration of anthocyanins while skin colour (SC) was essentially correlated with relative abundance of anthocyanins. SPC and SPC′ were correlated with a mixture of concentration and relative abundance variables.

4. Discussion

Different measures of anthocyanin concentration either mg/kg, mg/berry and mg/l are equally useful and provide overlapping information. Concentration in mg/kg was selected for cultivar characterisation and genetic association analysis since its interpretation is clearer.

Data on relative abundance and concentration of anthocyanins do not always overlap. For some anthocyanins the information on concentration and on relative abundance is identical but for others is different. Cyanidin-3-monoglucoside is an extreme example of different information provided by relative abundance and concentration. In some

cultivars this anthocyanin is present in moderate concentrations but represents 100 % of total pigments. This kind of difference is most likely the reason for the non significant correlation in cyanidin-3-monoglucoside case and the small correlation coefficients for other anthocyanins. These observations suggest that different results may be obtained using these two types of phenotypes in multivariate and association analysis and therefore excluding one of them could hide valuable information.

Total anthocyanins concentration varied widely across cultivars. Relative abundance of each anthocyanin also showed some variation but glucosides were the predominant pigments and malvidin-3-monoglucoside was the most abundant. Great diversity has also been observed at the ratios of coumarate/acetate derivative anthocyanins and tri/dihydroxylated anthocyanins, showing a rich gradient of enzymatic activities, respectively acyl transferase (Gonzalez-Sanjosé and Diez, 1990) and hydroxylase (Roggero *et al.*, 1988; Bogs *et al.*, 2007).

Principal component analysis based on both concentration and relative abundance of anthocyanins allowed cultivar distinction by acylation pattern and anthocyanidin type. However, the two approaches did not yield completely similar results. PCA based on concentration data showed that the strongest discriminating feature was total concentration for each anthocyanin and total anthocyanins, hiding more subtle differences. Also these two approaches were different concerning the anthocyanidin variable groupings. For concentration based PCA the different anthocyanidins were separated according to methylation level and for relative abundance discrimination was based on di/trihydroxylation of the B ring, reflecting respectively methyl transferase and hydroxylase enzymatic activities. Cluster analysis did not

reveal major clusters of cultivars. The largest clusters formed were in agreement with PCA analysis.

No significant correlation was found between anthocyanins concentration and genotypic variation based on DNA co-dominant markers. A low but significant correlation was found between the relative abundance phenotype and molecular variance. This observation is in agreement with previous works in maize (Rebourg *et al.*, 2003; Hartings *et al.*, 2008). The distribution of the calculated distances supports the higher accuracy of molecular markers data compared with phenotypic data. This difference is most certainly due to the effects of the environment.

Overall, virus infections and maturation parameters have not shown high association with anthocyanin concentration or relative abundance in the studied sample. Therefore, these are not considered important covariates to include on a genetic association study using this sample. It must be considered that different results may be obtained in a different sample of cultivars or using several clones of the same cultivar as described by previous publications (Guidoni *et al.*, 2000; Lider *et al.*, 1975; Goheen, 1958; Cabaleiro *et al.*, 1999).

The study of association of visual characterisation of pulp and skin colour with concentration and relative abundance of anthocyanins has shown that pulp colour reflects mainly concentration differences in anthocyanin content while skin colour is strongly related to relative abundance of anthocyanins. The variables resulting from the combination of these two are related to both concentration and relative abundance variation of anthocyanins.

It must be considered that the results here presented were based on phenotypic measurement of only one year. Therefore, care must be taken

when generalizing these conclusions for other harvest years, especially concerning total skin anthocyanin concentration. Data on anthocyanins relative abundance has been shown to be less sensitive to environmental variations (Mazza, 1999). Nevertheless, the phenotypic information here collected is the first comprehensive characterisation of anthocyanin content of the majority of the cultivars that compose the Portuguese Grapevine Collection. This data now available will be ideally collated with future characterisations in other years and provide basis for further works.

5. References

Aradhya, M.K., Dangl, G.S., Prins, B.H., Boursiquot, J.M., Walker, M.A., *et al.* (2003). Genetic structure and differentiation in cultivated grape, *Vitis vinifera* L. *Genetic Research* **81**, 179-192.

Arozarena, I., Ayestaran, B., Cantalejo, M.J., Navarro, M., Vera, M., *et al.* (2002). Anthocyanin composition of Tempranillo, Garnacha and Cabernet Sauvignon grapes from high and low quality vineyards over two years. *European Food Research and Technology* **214**, 303-309.

Benin, M., Gasquez, J., Mahfoudi, A., Bessis, R. (1988). Biochemical characterization of *Vitis vinifera* L. cultivars by electrophoresis of leaf isoenzymes – an attempt to classify grapevine varieties. *Vitis* **27**, 157-172.

Bogs, J., Jaffé, F.W., Takos, A.M., Walker, A.R., Robinson, S.P. (2007). The grapevine transcription factor *VvMYBPA1* regulates proanthocyanidin synthesis during fruit development. *Plant Physiology* **143**, 1347-1361.

Bowers, J.E. and Meredith, C.P. (1996). Genetic similaritites among wine grape cultivars revealed by restriction fragment-length polymorphism

(RFLP) analysis. *Journal of the American Society for Horticultural Science* **121**, 620-624.

Cabaleiro, C., Segura, A., Garcia-Berrios, J.J. (1999). Effects of grapevine leafroll-associated virus 3 in the physiology and must of *Vitis vinifera* L. cv. Albarino following contamination in the field. *American Journal of Enology and Viticulture* **50**, 40-44.

Cacho, J., Fernandez, P., Ferreira, V., Castells, J.E. (1992). Evolution of five anthocyanin-3-glucosides in the skin of the Tempranillo, Moristel, and Garnacha grape varieties and influence of climatological variables. *American Journal of Enology and Viticulture* **43**, 244-248.

Caló, A., Tomasi, D., Cravero, M.C., Di Stefano, R. (1994). Varietal analysis and classification of the species (*Vitis* sp.) by determination of anthocyanins and of hydroxycinnamoyl tartaric acids in the skin of the red-berry cultivars. *Rivista di Viticoltura e di Enologia* **3**, 13–25.

Carreño, J., Almeida, L., Martínez, A., Fernandéz-López. (1997). Chemotaxonomical classification of red table grapes based on anthocyanin profile and external colour. *Lebensmittel-Wissenschaft und-Technologie* **30**, 259-265.

Cervera, M.T., Cabezas, J.A., Sancha, J.C., Martínez de Toda, F., Martínez-Zapater, J.M. (1998). Application of AFLPs to the characterization of grapevine *Vitis vinifera* L. Genetic resources. A case study with accessions from Rioja (Spain). *Theoretical and Applied Genetics* **97**, 51-59.

Chakraborty, R. and Jin, L. (1993). Determination of relatedness between individuals using DNA-fingerprinting. *Human Biology* **65**, 875-895.

Doligez, A., Adam-Blondon, A.F., Cipriani, G., Di Gaspero, G., Laucou, V., *et al.* (2006). An integrated SSR map of grapevine based on five mapping populations. *Theoretical and Applied Genetics* **113**, 369-382.

Falush, D., Stephens, M., Pritchard, J.K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, 7, 574-578.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using boostrap. *Evolution* **39**, 783-791.

Franks, T., Botta, R., Thomas, M.R. (2002). Chimerism in grapevines: implications for cultivar identity, ancestry and genetics improvement. *Theoretical and Applied Genetics* **104**, 192-199.

Goheen, A.C., Harmon, F.N., Weinberger, J.H. (1958). Leafroll (white emperor disease) of grapes in California. *Phytopathology* **48**, 51-54.

Gonzalez-San José, M.L., Santa-Maria, G., Diez, C. (1990). Anthocyanins as parameters for differentiating wines by grape variety, wine-growing region, and wine-making methods. *Journal of Food Composition and Analysis* **3**, 54-66.

Grotewold, E., Chamberlin, M., Snook, M., Siame, B., Butter, L., *et al.* (1998). Engineering secondary metabolism in maize cells by ectopic expression of transcription factors. *The Plant Cell* **10**, 721-740.

Guidoni, S., Mannini, F., Ferrandino, A., Argamante, N., Di Stefano, R. (1997). The effect of grapevine leafroll and rugose wood sanitation on agronomic performance and berry and leaf phenolic content of Nebbiolo clone (*Vitis vinifera* L.). *American Journal of Enology and Viticulture* **48**, 438-442.

Hartings, H., Berardo, N., Mazzinelli, G.F., Valoti, P., Verderio, A., *et al.* (2008). Assessment of genetic diversity and relationships among maize (*Zea mays* L.) Italian landraces by morphological traits and AFLP profiling. *Theoretical and Applied Genetics* **117**, 831-842.

Jackson, D.A. and Somers, K.M. (1991). Putting things in order – the ups and down of detrended correspondence analysis. *American Naturalist* **137**, 704-712.

Lider, L.A., Goheen, A.C., Ferrari, N.L. (1975). Comparison between healthy and leafroll-affected grapevine planting stocks. *American Journal of Enology and Viticulture* **26**, 144-147.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209-220.

Mazza, G. and Miniati, E. (1993). Grapes. In *Anthocyanins in Fruits, Vegetables and Grains.* Mazza, G. and Miniati, E., Eds. CRC Press: Boca Raton, pp 149-199.

Mazza, G., Fukumoto, L., Delaquis, P., Girard, B., Ewert, B. (1999). Anthocyanins, phenolics, and color of Cabernet Franc, Merlot, and Pinot Noir wines from British Columbia. *Journal of Agriculture and Food Chemistry* **47**, 4009–4017.

Mori, K., Goto-Yamamoto, N., Kitayama, M., Hashizume, K. (2007). Effect of high temperature on anthocyanin composition and transcription of flavonoids hydroxylase genes in 'Pinot noir' grapes (*Vitis vinifera*). *Journal of Horticultural Science & Biotechnology* **82**, 199-206.

Nuñez, V., Monagas, M., Gomez-Cordovés, M.C., Bartolomé, B. (2003). *Vitis vinifera* L. cv. Graciano grapes characterized by its anthocyanin profile. *Postharvest Biology and Technology* **31**, 69-79.

Organisation Internationale de Vigne et du Vin (OIV). (1983). *1st Edition of the OIV Descriptor list for grape varieties and Vitis species.* OIV: Paris.

Organisation Internationale de Vigne et du Vin (OIV). (2009a). *2nd Edition of the OIV Descriptor list for grape varieties and Vitis species.* OIV: Paris.

Organisation Internationale de Vigne et du Vin (OIV). (2009b). *Recueil des méthodes internationales d'analyse des vins et des moûts.* OIV: Paris.

Pritchard, J. K., Stephens, M., Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Rebourg, C., Chastanet, M., Gouesnard, B., Welcker, C., Dubreil, P., *et al.* (2003). Maize introduction into Europe: the history reviewed in the light of molecular data. *Theoretical and Applied Genetics* **106**, 895-903.

Riaz, S., Garrison, K.E., Dangl, G.S., Boursiquot, J.M., Meredith, C.P. (2002). Genetic divergence and chimerism within ancient asesually propagated wine grape cultivars. *Journal of the American Society for Horticultural Science* **127**, 508-514.

Ribéreau-Gayon, P. (1959). Recherches sur les anthocyanes des végétaux. Application au genre *Vitis.* Ph.D. Thesis, University of Bourdeux.

Ribéreau-Gayon, P. (1964). Les composés phénoliques du raisin et du vin. II. Les flavonosides et les anthocyanosides. *Annales de Physiologie Végétale* **6**, 211-242.

Ribéreau-Gayon, P. (1978). In *Plant phenolics.* Heywood, V.H., Ed. Hafner Publishing Co.: New York, pp 54.

Rohlf, F.J. (2000). *NTSYSpc. Numerical taxonomy and multivariate analysis system, version 2.1.* Applied Biostatistics Inc.: New York.

Roggero, J.P., Larice, J.L., Rocheville-Divorne, C., Archier, P., Coen, S. (1988). Composition anthocyanique des cépages. I-Essai de classification par analyse en composantes principales et par analyse factorielle discriminante. *Revue Française d'Œnologie* **112**, 41–48.

Ryan, J.M. and Revilla, E. (2003). Anthocyanin composition of Cabernet Sauvignon and Tempranillo grapes at different stages of ripening. *Journal of Agriculture and Food Chemistry* **51**, 3372-3378.

Sefc, K.M., Lefort, F., Grando, M.S., Scott, K.D., Steinkellner, H., *et al.* (2001). Microsatellite markers for grapevine: a state of the art. *Molecular Biology & Biotechnology of the Grapevine* **463**, 433 – 463.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical taxonomy – the principles and practice of numerical classification.* Freeman: San Francisco, pp 573.

Sokal, R.R. and Rohlf, F.J. (1995). *Biometry*. 3rd edition. Freeman: New York, pp 887.

Tessier, C., David, J., This, P., Boursiquot, J.M., Charrier, A. (1999). Optimization of the choice of molecular markers for varietal identification in *Vitis vinifera* L. *Theoretical and Applied Genetics* **98**, 171-177.

Troggio, M., Malacarne, G., Coppola, G., Segala, C., Cartwright, D., *et al.* (2007). A dense sigle-nucleotide polymorphism-based genetic linkage map of grapevine (*Vitis vinifera* L.) anchoring pinot noir bacterial artificial chromosome contigs. *Genetics* **176**, 2637-2650.

Vignani, R., Bowers, J.E., Meredith, C.P. (1996). Microsatellite DNA polymorphism analysis of clones of *Vitis vinifera* 'Sangiovese'. *Scientia Horticulturae* **65**, 163-169.

Wenzel, K., Dittrich, H.H., Heimfarth, M. (1987). Anthocyanin composition in berries of different grape varieties. *Vitis* **26**, 65-78.

Wulf, L.W. and Nagel, C.W. (1978). High-pressure liquid chromatographic separation of anthocyanins of *Vitis vinifera*. *American Journal of Enology and Viticulture* **29**, 42-49.

6. Acknowledgements

I would like to acknowledge Flávia Moreira (IASMA) for performing SSR genotyping; Margarida Santos for ELISA tests (INIA - Oeiras) and Isabel Spranger, Conceição Leandro and Baoshan Sun (INIA - Dois Portos) for assistance on anthocyanin extraction and measurement. I would also like to thank Nikolas Maniatis for his assistance with data analysis.

CHAPTER IV

A CANDIDATE GENE ASSOCIATION STUDY FOR BERRY COLOUR AND ANTHOCYANIN CONTENT IN *VITIS VINIFERA* L.

This chapter is an extended version of a manuscript accepted for review in *PLoS One*:

Cardoso, S., Lau, W., Eiras Dias, J., Fevereiro, P., Maniatis, N. (Submitted). A Candidate Gene Association Study for Berry Colour and Anthocyanin Content in *Vitis vinifera* L. *PLoS One*.

Contributions to this chapter:

- Designed the study: Cardoso, S., Eiras Dias, J.E., Fevereiro, P., Maniatis, N.
- Performed experiments: Cardoso, S.
- Analysed data: Cardoso, S., Lau, W., Maniatis, N.

Summary

Anthocyanin content is a trait of major interest in *Vitis vinifera* L.. These compounds affect grape and wine quality, and have beneficial effects on human health. A candidate gene approach was used to identify genetic variants associated with anthocyanin content in grape berries. A total of 445 polymorphisms were identified in five genes encoding transcription factors and 10 genes involved in either the biosynthetic pathway or transport of anthocyanins. A total of 124 SNPs were selected to examine association with a wide range of phenotypes based on RP-HPLC analysis and visual characterisation. The phenotypes were total skin anthocyanin (TSA) concentration but also specific types of anthocyanins and relative abundance. The visual assessment was based on OIV descriptors for berry and skin colour. Association tests were performed both with and without accounting for population structure and relatedness. The genes encoding the transcription factors *MYB11*, *MYBCC* and MYC_B were significantly associated with TSA concentration. *UFGT* and *MRP* were associated with several different types of anthocyanins. Skin and pulp colour were associated with nine genes (*MYB11*, *MYBCC*, MYC_B, *UFGT*, *MRP*, *DFR*, *LDOX*, *CHI* and *GST*). Pulp colour was associated with a similar group of 11 genes (*MYB11*, *MYBCC*, MYC_B, MYC_A, *UFGT*, *MRP*, *GST*, *DFR*, *LDOX*, *CHI* and CHS_A). Statistical interactions were observed between SNPs within the transcription factors *MYB11*, *MYBCC* and MYC_B. SNPs within *LDOX* interacted with *MYB11* and MYC_B, while SNPs within *CHI* interacted only with *MYB11*. Together, these findings suggest the involvement of these genes in anthocyanin content and on the regulation of anthocyanin biosynthesis. This work forms a benchmark for replication and functional studies.

1. Introduction

1.1. Anthocyanins

Anthocyanins are natural pigments which accumulate especially in fruits and flowers (Brouillard, 1982). The interest in these compounds has increased due to their potential as natural innocuous food colourants and as antioxidants with several benefits for human health (Giusti and Wrolstad, 2003). Anthocyanins play an important role in wine and grape industry since their accumulation gives colour to grapes and influences organoleptic characteristics of wines (Ribéreau-Gayon, 1982).

Anthocyanins are part of the larger group of flavonoids. There are several different kinds of anthocyanins found in nature, differing at B-ring position, sugar residue, and organic acid. Their structural differences and the amount accumulated determine the colour observed (Eder, 2000; Harborne, 1998).

The anthocyanins biosynthetic pathway is well characterised since it has been thoroughly studied in petunia, snapdragon and maize (Martin and Gerats, 1993). This pathway may be divided in two main stages. First, phenylalanine is converted to 4-coumaroyl Co-A in the general phenylpropanoid pathway. Subsequently the 4-coumaroyl Co-A is converted into anthocyanins in the flavonoid pathway (Verpoort, 2000).

In many plant species, anthocyanin biosynthesis has been shown to be regulated by regulatory genes belonging to three major families, *Myb*, *β helix-loop-helix* (*bHLH*) (also known as *Myc*) and *tryptophan-aspartic acid repeat* (*WDR* or *WD40* repeats) families (Baudry *et al.*, 2004; Borovsky *et al.*, 2004; Holton and Cornish, 1995; Matus *et* al., 2010; Payne *et al.*, 2000; Ramsay *et al.* 2003; Robbins *et al.* 2003; Sainz *et al.* 1997; Schwinn *et al.* 2006; Spelt *et al.* 2000). In grapevine, *Myb* and *Myc* family genes have been shown to affect expression of structural genes in

the biosynthetic pathway and to interact between themselves (Bogs *et al.*, 2007; Cutanda-Perez *et al.*, 2009; Deluc *et al.*, 2006, 2008; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; Matus *et al.*, 2009, 2010; Terrier *et al.*, 2009; This *et al.*, 2007; Ageorges, 2006; Walker, 2007). Recently, Matus *et al*. (2010) observed also a correlation between *WDR1* expression and anthocyanin accumulation in grapevine.

Flavonoid composition in grapes has been shown to be affected by environmental effects such as temperature and light exposure (Cortell and Kennedy, 2006; Downey *et al.*, 2004, 2006; Jeong *et al.*, 2004; Matus *et al.*, 2009). *Myb* family genes have an important role in response to these factors (Matus *et al.*, 2009).

Anthocyanin biosynthetic pathway genes have been mapped to five different linkage groups and *Myb* transcription factors to two linkage groups (LG) (Salmaso *et al*., 2008). Berry skin colour considered simply as a dichotomous trait, with berries with non-coloured skin versus berries with coloured skin, was observed to have Mendelian segregation (Fischer, 2004; Salmaso, 2008). This trait was mapped to LG2 and Salmaso (2008) have mapped one transcription factor (*MybA1*) to the same locus (Doligez, 2002, 2006; Fischer, 2004). Fournier-Level *et al.* (2009) mapped colour as a quantitative trait to LG2.

Expression, functional and association studies have contributed to a better understanding of the regulation of anthocyanins by *Myb* family genes. The absence of anthocyanins has been shown to be determined by the homozygous presence of a *MybA1* allele with a retrotransposon insertion (*Gret1*) in the gene promoter region (Fournier-Level *et al.*, 2009; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; This *et al.*, 2007). Genes *MybA1* and *MybA2* multiallelic mutations control the biosynthetic step mediated by *UFGT* (Kobayashi *et al.*, 2002; Walker *et*

al., 2007). However, this phenotype seems to be influenced by other genes, since some exceptions occur where white cultivars do not have *Gret1* insertion (This *et al.*, 2007). Other transcription factors, *Myb5a*, *Myb5b*, *MybPA1* and *MybPA2* have been found to affect expression of genes coding enzymes involved in earlier steps of the pathway by promoter activation (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Matus *et al.*, 2009; Terrier *et al.*, 2009).

Also, variation within coloured cultivars has not been completely understood yet. Four polymorphisms in *MybA1* have shown to be associated with pink/red cultivars (This *et al.*, 2007). Anthocyanin amount as a quantitative trait is expected to be determined by small contributions of many genes. Recently, Fournier-Level *et al.* (2009) identified four polymorphisms in *MybA1*, *MybA2* and *MybA3* accounting for 23 % of colour variance. However, this study used an overrepresented sample of white-berried cultivars.

1.2. Association mapping

The aim of association mapping is to find a correspondence between genotypes and phenotypes on a population scale.

Linkage mapping has been successfully used to identify single-gene "Mendelian traits" (Corder *et al.*, 1993; Zielenski and Tsui, 1995) and QTL regions (Alpert and Tanksley, 1996; Stuber et al., 1999). However, the success of this method to identify genes involved in quantitative trait loci (QTLs) is limited. This limitation has been attributed to the low power and resolution of linkage studies to detect variants of small effects (Altmüller *et al.*, 2001; Risch and Merikangas, 1996; Risch, 2000).

Linkage disequilibrium (LD) is the statistical association of alleles at two loci in a population (Balding, 2006). Patterns of LD are mainly

shaped by recombination. However, other factors such as mutation, selection, genetic drift, population demography and breeding systems affect LD. LD is the genetic basis for association mapping. The location of a trait-influencing allele can be inferred by measuring the association between the trait phenotype and a marker allele in LD with the trait-influencing allele.

Both linkage mapping and association mapping rely on the co-inheritance of DNA variants to infer trait genes. However, association mapping explores recombination events over many generations, while linkage mapping relies on recombination events taking place over the few generations included on a pedigree. As a consequence, trait associated regions identified by association mapping are much smaller than the regions identified by linkage mapping.

Continuous advances in molecular genetics techniques, bioinformatics and statistical analysis make association mapping an increasingly appealing approach to unveil the genetic basis of traits.

Association mapping studies design may vary depending on the trait and species of interest, and the available resources. Case-control studies are used when the trait of interest is dichotomous. In this design, evidence of association is obtained if marker allele frequencies differ significantly between affected and unaffected individuals. In the case of continuous traits, evidence of genetic association is obtained by statistical association between allelic variants and the trait of interest.

The genome area under study in association mapping studies may include the whole genome or focus on candidate genes or candidate regions. Genome-wide association mapping does not rely on previous hypothesis about genes associated with the trait of interest. Genome-wide association mapping explores variation across the whole genome to find

association with the trait of interest. The candidate-gene approach tests association between the trait of interest and variation on genes hypothesised to be associated with this trait based on previous research data. The selection of candidate genes depends mainly on regions associated with the trait of interest on previous linkage or association mapping studies. However, often candidate genes selection is also based on genes involved in pathways or regulatory processes likely to affect the trait of interest. Evidence of differential gene expression associated with the trait of interest has also been used to select candidate genes.

Association studies are often based on single marker tests. However, haplotype tests may also play an important role. Different studies have reached contradictory conclusions about power comparisons between these two methodologies. Akey *et al.* (2001) concluded that haplotype tests have higher power to detect associations while Long and Langley (1999) and Kaplan and Morris (2001) obtained higher power with single SNP tests.

Association mapping is a useful tool for the identification of genetic variation that contributes to diseases and traits. However, many studies have shown problems, raising doubts on the reliability of the findings (Terwilliger et al., 1998; Gambaro et al., 2000; Weiss and Terwilliger, 2000). Power to detect genetic associations is influenced by sample size, linkage disequilibrium (LD) between the genotyped marker and the causal variant, effect size, and marker and causal variant frequencies. Several factors can lead to spurious association. For example, population structure, relatedness, poor study design and inaccurate phenotypic data (Cardon and Bell, 2001). Nevertheless, population stratification and more recently cryptic relatedness have received a great deal of attention.

Population stratification is the case where a population includes

subgroups of individuals characterised by different allele frequencies (Cardon and Bell, 2001). This may give rise to spurious associations when the trait of interest is most prevalent in one subpopulation and therefore associates with any allele with higher frequency in this subpopulation (Pritchard and Rosenberg, 1999).

Many methods have been developed to deal with this problem. Genomic Control (GC) is one of these methodologies. It was developed by Devlin and Roeder (1999) to deal with population structure in population based designs, by using random markers to calculate an inflation factor to adjust significance tests bias. Pritchard *et al.* (2000) proposed another statistical correction. This method is based on a Bayesian clustering approach that estimates the proportion of each individual's variation that came from each subpopulation. This proportion is then included in association tests. Also a method based on Principal Component Analysis is widely used in genome-wide association studies for dealing with structure problems (Price *et al.*, 2006). The axes of genetic variation based on the analysis of molecular markers are used to adjust phenotypes and genotypes for association tests.

In grapevine and many other agricultural species, a certain degree of relatedness is expected due to selection and breeding history (Zhu *et al.*, 2008). Uneven familial relationships between groups of individuals may also cause spurious associations. Recent studies suggest that correcting for pairwise relatedness besides structure decreases false positives and increases power (Malosetti *et al.*, 2007; Yu *et al.*, 2006; Zhao *et al.*, 2007). It has been argued that this is due to structure and relatedness capturing different levels of variation (Yu *et al.*, 2006). Especially in association mapping of plant species, where germplasm collections tend

to gather related and admixed accessions that have high interest to breeders, this is an issue of major concern (Zhu et al., 2008).

Yu *et al.* (2006) developed a mixed model approach to account for population structure and cryptic relatedness detected by molecular markers while testing for genetic association. The mixed model has traditionally been used in animal breeding studies with well described pedigrees for genetic evaluation of livestock. In association genetics, pedigree records are often incomplete or inaccurate. As an alternative, marker-based relatedness matrices have been suggested. Yu *et al.* (2006) used a model that considers structure using the method presented by Pritchard *et al.* (2000) and a relatedness matrix based on Ritland's kinship coefficient (RKC). This measure of kinship is estimated based on the probability of Identity by State (IBS) between two individuals adjusted to the average probability of IBS between random individuals in the population (Ritland, 1996; Yu *et al.*, 2006). The matrix obtained may not be positive semidefinite generating mathematical problems which might bias further likelihood estimates (Kang *et al.*, 2008).

Alternatively, Kang *et al.* (2008) suggested a kinship matrix based on the proportion of shared alleles (PSA), a similarity measure first proposed by Chakraborty and Jin (1993). Zhao *et al.* (2007) used haplotype data on 95 *Arabidopsis* accessions and showed with simulation data that this matrix is at least equally effective for taking into account relatedness. Kang *et al.* (2008) showed that this evades convergence and mathematical problems compared to the relationship matrix based RKC.

Besides the concerns about false positive results, false negatives are also a problem in association mapping. The presence of interactions between loci is one of the reasons often cited to justify the inability to successfully identify associations (Culverhouse *et al.*, 2002; Moore,

2003). If a genetic factor has an effect on the phenotype through a complex mechanism involving other genes, examining each gene separately may not have enough power to detect this effect (Cordell, 2009). As a result several methods have been developed to analyze statistical interactions between loci that may be informative on biological pathways underlying traits. The most common are regression models; however, other methods have been proposed and are revised by Cordell (2009; Chanda *et al.*, 2007; Dong *et al.*, 2008; Kang *et al.*, 2008; Moore *et al.*, 2006; Yang *et al.*, 2008).

In association mapping studies, on single locus analyzes or on several loci interaction analyzes, multiple testing is an important concern. The genotyping of a large number of markers on the same sample of individuals leads to a large number of tests performed and increased chances of false positives. Bonferroni correction is very conservative and assumes marker independence. Therefore, it is inadequate when the markers used are in LD. Dataset permutations have been argued as the most adequate method to correct for multiple testing (Cardon and Bell, 2001).

Despite the large number of studies on grape colour, there is still no clear understanding on the genetics underlying this phenotype. The studies performed to date, have focused on either presence or absence of colour, categorical variation of colour or total concentration of anthocyanins. This is the first study that examines a wide range of phenotypes including different types of anthocyanin concentration and relative abundance (RA). Obtaining samples for association studies in grapevine is still very challenging. This is because germplasm collections have often limited numbers of cultivars or cultivars with no phenotypic

records. This study uses one of the largest samples for association mapping and sequence data to identify polymorphisms in grapevine.

The aim is to look for associations between 15 candidate genes and grape colour using 124 newly discovered SNPs and a wide range of colour related phenotypes, including visual assessment, TSA concentration and specific types of anthocyanins concentration and RA. We also investigate the importance of population structure and relatedness in grapevine.

2. Material and methods

This candidate gene study was divided in two phases. In a first phase, SNP identification in the candidate genes was performed by sequencing a small sample of 22 cultivars. The second phase involved association analysis by genotyping 124 SNPs in 149 cultivars. Figure IV-1 shows a scheme of the study design stages.

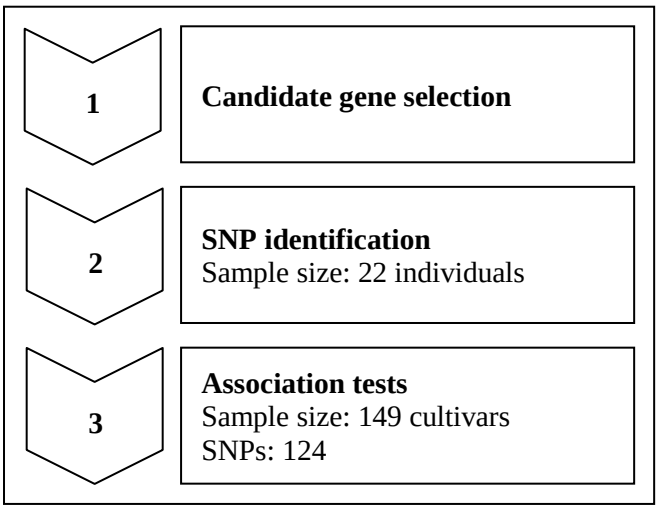


Figure IV-1 Scheme of the study design stages.

2.1. Candidate genes

Selection of candidate genes was based on their biological functions and expression analysis. Information on function was obtained from published results whereas expression analysis was undertaken for the purpose of this thesis (Chapter II).

Table IV-1 shows the list of selected genes for this study and the source of information supporting their selection. Also the name and symbol of the encoded proteins are listed. A total of fifteen candidate genes in total were selected. This included genes encoding enzymes involved in the biosynthetic pathway of anthocyanins, chalcone synthase (*CHS*), chalcone isomerase (*CHI*), flavanone 3-hydroxylase (*F3H*), flavonoid 3′-hydroxylase (*F3'H*), dihydroflavonol reductase (*DFR*), leucoanthocyanidin dioxygenase (*LDOX*) and UDP-glucose: flavonoid 3-*O*-glucosyltransferase (*UFGT*) (Figure I-1). Other genes involved in the pathway were not included due to difficulties on the design of good quality primer pairs.

Two genes related to transport and accumulation of anthocyanins in the vacuole were also selected as candidate genes. Glutathione *S*-transferase (*GST*) has been shown to be involved in vacuolar accumulation of anthocyanins in grapevine by Ageorges *et al.* (2006). It has been suggested that *GST* binds anthocyanins through hydrofobic interactions and transports them to the tonoplast membrane (Mueller *et al.*, 2000). Multidrug resistance–associated protein (*MRP*) has been shown to be involved in anthocyanins transport across the tonoplast in maize (Goodman *et al.*, 2004).

Table IV-1 List of candidate genes.

Chr.	Scaf.[1]	Gene ID[1]	Code	Coded protein name	Function	SNPs
Unk	168	GSVIVT00006341001	*CHS$_A$**	Chalcone synthase	Involved in anthocyanins biosynthetic pathway. Catalyzes the condensation of one molecule of 4-coumaroyl CoA and three molecules of malonyl-CoA into a naringenin chalcone.	5
14	9	GSVIVT00037967001	*CHS$_C$**	Chalcone synthase		7
13	48	GSVIVT00029513001	*CHI**	Chalcone isomerase	Involved in anthocyanins biosynthetic pathway. Catalyzes the isomerisation of the naringenin chalcone into a naringenin flavanone.	3
4	83	GSVIVT00036784001	*F3H**	Flavanone 3-hydroxylase	Involved in anthocyanins biosynthetic pathway. Catalyzes the hydroxylation of naringenin flavanone to dihydrokaempfenol.	5
17	12	GSVIVT00016215001	*F3'H$_B$**	Flavonoid 3′-hydroxylase	Involved in anthocyanins biosynthetic pathway. Catalyzes the hydroxylation of dihydrokaempferol at the 3' position of the B-ring.	3
18	1	GSVIVT00014584001	*DFR**	Dihydroflavonol reductase	Involved in anthocyanins biosynthetic pathway. Catalyzes the reduction of the dihydroflavonols into leucoanthocyanidins.	12
2	112	GSVIVT00001063001	*LDOX**	Leucoanthocyanidin dioxygenase	Involved in anthocyanins biosynthetic pathway. Catalyzes the conversion of leucoanthocyanidins into anthocyanidins.	3
16	10	GSVIVT00014047001	*UFGT**	UDP-glucose: flavonoid 3-*O*-glucosyltransferase	Involved in anthocyanins biosynthetic pathway. Catalyzes the conversion of anthocyanidins into anthocyanins.	19
9	7	[1]XM_002276176	*MRP**	Multidrug resistance–associated protein	Involved on vacuolar accumulation of anthocyanins in maize. ATP-binding transporter which mediates the primary transport of anthocyanins across the tonoplast.	14
Unk.	30	GSVIVT00023496001	*GST**	Glutathione *S*-transferase	Involved in vacuolar accumulation of anthocyanins in grapevine. Thought to bind anthocyanins through hydrofobic interactions and escort them to the tonoplast membrane.	3
Unk.	203	GSVIVT00008627001	*MYC$_A$**	Basic helix-loop-helix transcription factor	Involved in regulation of the flavonoid and anthocyanin metabolism in other plants and in grapevine.	6
2	11	GSVIVT00015763001	*MYC$_B$*[§]	Basic helix-loop-helix transcription factor	Involved in regulation of the flavonoid and anthocyanin metabolism in other plants and in grapevine.	10
9	7	GSVIVT00034097001	*MYB11*[§]	Myb transcription factor	Involved in regulation of the flavonoid and anthocyanin metabolism in other plants and in grapevine.	18
4	83	GSVIVT00036753001	*MYB9*[§]	Myb transcription factor	Involved in regulation of the flavonoid and anthocyanin metabolism in other plants and in grapevine.	4
Unk.	342	[1]XM_002272552.1	*MYBCC*[§]	Myb transcription factor	Involved in regulation of the flavonoid and anthocyanin metabolism in other plants and in grapevine.	12

Chr. stands for chromosome, *Scaf.* for scaffold and *Unk.* for unknown chromosome. *SNPs* column shows the total number of SNPs genotyped for association analysis. [1]Scaffold and gene IDs are according to Genoscope, sequencing version 8x coverage. NCBI locus nomenclature is shown for *MYBCC* and *MRP* because Genoscope annotation was not available in these cases. *Candidate genes selection based on literature review. [§]Candidate genes selection based on previous expression analysis (personal communication). Codes used to designate candidate genes selected based on expression analysis were retrieved from UniProt database description.

Five genes encoding transcription factors were selected as candidate genes (*MYC_A*, *MYC_B*, *MYB9*, *MYB11*, *MYBCC*). These genes showed subtle differential expression in Aragonez cultivar clones with contrasting grape skin colours (Chapter II). From a final gene list of 24 genes showing differential expression, these five genes were selected based on the transcription family, on the ability to replicate results for different *t*-tests, on the *P*-values and on the ability to design good quality primer pairs. *Myb* and *Myc* families were given priority since these constitute major families of transcription factors, shown to regulate anthocyanin biosynthesis in many plants (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; Matus *et al*., 2009, 2010; Terrier *et al*., 2009; This *et al*., 2007).

2.2. Phase 1: SNP identification

2.2.1. PCR and sequencing of 22 cultivars

Cultivars with white and coloured pulp and with a range of skin colours were included. When possible, data from previous works were used to include dark skin cultivars covering a wide range of total skin anthocyanin (TSA) concentration in berry skin (Garcez, 1997). These cultivars and the phenotypic data are shown on Appendix 13.

Approximately 100mg of leaf fresh weight was used to extract genomic DNA. Mortar and pestle grinding with sterile quartz sand were used with Quiagen Mini Kit (Quiagen Inc, Hilden, Germany). Quantification was done spectrophotometricaly.

Sequences available on NCBI were used to design the primers with Primer3 software (Rozen and Skaletsky, 2000). Primers were designed to amplify DNA fragments of approximately 800bp. These fragments, overlapping by approximately 100 bp, covered the candidate genes and

respective predicted promoter regions. Published characterisation of promoters was available for *CHI* (Bogs *et al.*, 2007), *DFR* (Gollop *et al.*, 2002), *UFGT* (Kobayashi *et al.*, 2001) and *LDOX* (Gollop *et al.*, 2001). For the remaining genes, promoters were not described. In these cases, TSSP promoter prediction program for plant genes available on SoftBerry network server (http://www.softberry.com) was used to predict the transcription start site and identify promoter motifs up to 2000 bp upstream the transcription start site.

Ninety seven primer pairs were tested. Amplifications were performed in a 10 µl final volume containing 1X PCR buffer, 0.2 mM of each dNTP, 0.5 µM of each primer, 0.75 ng of genomic DNA as template and 0.025 U of Taq DNA Polymerase (Promega). A touchdown cycling strategy was adopted using a Biometra Thermocycler (Biometra, Göttingen, Germany). The thermocycler was programmed as follows: an initial denaturing step of 3 minutes at 94 ºC, 45 cycles and a final extension of 10 minutes at 72 ºC. Each cycle consisted on a denaturing step of 45 seconds at 94 ºC, an annealing step of 30 seconds, starting at a temperature according to primer pair and decreasing each cycle by 0,5 ºC along 15 cycles, and an extension step of 30 seconds at 72 ºC. Fragments were checked by electrophoresis in a 2 % agarose gel. Automated sequencing was performed by STAB Vida, Lda. (Portugal).

Some gene regions did not succeed on PCR amplification or sequencing. These difficulties occurred more commonly on genes which are part of gene families, on especially repetitive regions and also on regions with successive heterozygous INDELs. The remaining regions were amplified successfully on 22 cultivars. Appendix 14 shows the sequences of the primer pairs that amplified successfully. SNPs were identified using CodonCode Aligner software (Codon Code Corp.).

2.2.2. SNP selection

A total of 445 DNA polymorphisms, including 407 SNPs and 38 INDELs, were identified in the studied sequences. These polymorphisms are listed in Appendix 15. Table IV-2 shows the number and percentage of polymorphisms in coding and non-coding regions and Table IV-3 shows their frequency per base pair distance.

Table IV-2 Polymorphisms identified in the sequenced regions.

	Number	**Percentage**
Non-coding regions (SNPs and INDELs)	311	69.89
Coding regions (SNPs)		
Synonymous SNPs	70	15.73
Non-synonymous SNPs	64	14.38
Total	445	

As is commonly observed, polymorphisms in non-coding regions were overall more common than in coding regions. Also, synonymous SNPs were overall more common than non-synonymous. INDELs were only identified in non-coding regions. The highest overall frequency of polymorphisms was observed in the gene coding *UFGT* and the lowest on *MYB9*. *UFGT* showed also the highest frequency of non-synonymous SNPs. The gene coding *DFR* had the lowest frequency of non-synonymous SNPs.

The selection of SNPs for further genotyping within the candidate genes was based on various quality control criteria. The selection criteria were missing values, minor allele frequency (MAF), amino acid changes caused by the SNP and base pair distance between different SNPs. Even though the sample size was small (22 cultivars), additional information on Hardy Weinberg (HW) and LD was obtained. SNPs with missingness

Table IV-3 Frequency of polymorphisms in the studied genomic regions.

	1 polymorphism per bp				
	Overall (SNPs or INDELs)	**Coding regions** (SNPs)	**Non-coding regions** (SNPs or INDELs)	**Synonymous** (SNPs)	**Non-synonymous** (SNPs)
Overall	82.50	273.97	118.05	524.46	573.63
CHI	292.00	584.00	584.00	0.00	584.00
CHS2A	202.50	202.50	0.00	202.50	0.00
CHS2C	50.27	94.25	107.71	125.67	377.00
DFR	44.03	869.50	46.37	1159.33	3478.00
F3H	143.00	1430.00	158.89	1430.00	0.00
F3'H$_B$	184.88	1479.00	211.29	0.00	1479.00
LDOX	157.00	0.00	157.00	0.00	0.00
UFGT	30.51	47.38	85.65	117.21	79.54
MRP	100.59	242.79	171.73	502.93	469.40
GST	76.46	0.00	76.46	0.00	0.00
MYC$_A$	137.08	253.08	299.09	470.00	548.33
MYB11	49.92	312.00	59.43	624.00	624.00
MYB9	260.40	1302.00	325.50	0.00	1302.00
MYC$_B$	162.55	357.60	298.00	447.00	1788.00
MYBCC	83.69	609.71	97.00	1067.00	1422.67

> 20 % were avoided. Markers with MAF < 2 % were excluded, since this would represent a unique allele among the 22 sampled individuals and therefore could be a genotyping error. Agreement with HW proportions was also considered for SNP selection. SNPs with strong deviations from HW ($\chi^2 > 10$) were excluded since these could be due to genotyping errors. Pairwise LD was estimated. Genotype frequencies were used to estimate haplotype frequencies by employing the iterative Expectation-Maximisation algorithm of Excoffier and Slatkin (1995). LD values were high across the studied genomic regions. Care was taken not to select more than one representative of pairs of SNPs in complete LD,

since these would provide overlapping information.

It is important to mention that analysis of LD, HWE and MAF were based on a small sample size (22 cultivars). However, these assessments provided good guidance for SNP selection. When genotyped on a larger sample (149 cultivars) most of these SNPs were polymorphic, did not deviate from HWE and had MAF < 0.02. Only 11.4 % of the SNPs were excluded by these criteria. SNPs causing amino acid substitutions were preferred as these are more likely to have a functional effect.

Finally, the length of the genes was considered. The selected SNPs were spaced on the physical map so that the whole gene would be covered. In cases where the selected SNPs were separated by a long distance, information from the SNP database hosted by The Institute for Genomic Research (TIGR) was used. Four SNPs retrieved from this database were genotyped to guarantee gene coverage. Also more than three SNPs within 20 base pairs were avoided due to further genotyping technology restrictions.

Following these selection criteria 140 SNPs in total were selected across the 15 genes for genotyping in a larger sample (149 cultivars). The supplementary table on Appendix 16 shows the list of these SNPs.

2.3. Phase 2: Association study

2.3.1. Association study sample

A sample of 149 cultivars with coloured berries was collected on the same vineyard in Dois Portos, Portugal, where the national ampelographic collection is established (Appendix 7). Young leaves were collected and stored at -80ºC.

2.3.2. Phenotypic characterisation

For phenotypic characterisation, fifty berries were collected from each cultivar and stored at -20ºC. Probable alcohol percentage was used as an indicator of berries maturity state at harvest. The collected berries had approximately 9 % probable alcohol.

Anthocyanin extraction was conducted with acidified methanol (0.1 % HCl) and identification was performed by reverse-phase high performance liquid chromatography (HPLC). A calibration curve was used to calculate anthocyanin concentration. This curve was obtained by regression through the origin of HPLC peak areas on concentration (in mg/l) of an external pattern of malvidin-3-O-glucoside chloride (Hoffman-La Roche, Switzerland) (See Chapter II for more details on anthocyanins extraction, identification and quantification).

The phenotypes included a wide range of traits based on either RP-HPLC analysis or visual characterisation of berry colour. Phenotypes based on RP-HPLC analysis included TSA concentration (milligrams of anthocyanins per kilogram of berries) but also specific types of anthocyanins and relative abundance (RA). Anthocyanins were grouped based on the type they belonged to, for example anthocyanidin and acylation types. Additionally, ratios between di/trihydroxylated anthocyanins and coumarate/acetate derivatives were used as phenotypes. The entire list of phenotypes is presented in Table IV-4. This table also shows which of the phenotypes were quantitative or qualitative. Different visual characterisations of berries colour were used as phenotypes. These were categorical variables and the different categories considered are shown in Table IV-4. Seventeen anthocyanins and one isomer were considered in both concentration and RA. These anthocyanins included monoglucosides, acetate derivatives and coumarate derivatives of

delphinidin, cyanidin, petunidin, peonidin and malvidin. Caffeoate derivatives of peonidin and malvidin were also considered. An isomer of coumarate derivative malvidin was included in the phenotypes group as well. Total concentration of anthocyanins was also considered. Sums of each anthocyanidin type and acylation type, expressed both as concentration and RA were also part of the phenotype list.

Pulp colour (PC) is a dichotomous trait (coloured versus white pulp). Skin colour (SC) was classified according to descriptor number 225 by the International Organisation of Vine and Wine (OIV). This descriptor establishes the following five categories: rose, red, grey, dark red violet and blue black. OIV descriptor 225 has a certain degree of subjectivity and PC is likely to influence to some extent the classification of SC. Since these visual phenotypes were aimed to be used for genetic association analysis, two new classifications were tested targeting higher accuracy. These classifications were named SPC and SPC′ and were obtained by joining pulp and skin colour classifications (Table IV-4). Only three categories for SPC were established. The first included cultivars with rose and red skin berries with white pulp. The second included grey, dark red violet and blue black skin cultivars with white pulp and the third included only coloured pulp cultivars. For SPC′ a sixth category for coloured pulp cultivars was added to the five OIV225 categories (Table IV-4).

TSA concentration varied widely across cultivars and it was treated as the main phenotype. RA of anthocyanins also varied although most often glucosides and malvidin-3-monoglucoside were the most abundant. Ratios of coumarate/acetate derivative anthocyanins and tri/dihydroxylated anthocyanins showed great diversity among cultivars, revealing a range of acyl transferase and hydroxylase enzymes,

Table IV-4 List of the phenotypes used for association analysis.

Phenotypes		Concentration (mg/kg)	Relative abundance (%)	Variable type
Anthocyanins	Delphinidin-3-monoglucoside	✔	✔	Q
	Cyanidin-3-monoglucoside	✔	✔	Q
	Petunidin-3-monoglucoside	✔	✔	Q
	Peonidin-3-monoglucoside	✔	✔	Q
	Malvidin-3-monoglucoside	✔	✔	Q
	Delphinidin-3-monoglucoside-acetate	✔	✔	Q
	Cyanidin-3-monoglucoside-acetate	✔	✔	Q
	Petunidin-3-monoglucoside-acetate	✔	✔	Q
	Peonidin-3-monoglucoside-acetate	✔	✔	Q
	Delphinidin-3-monoglucoside-*p*-coumarate	✔	✔	Q
	Malvidin-3-monoglucoside-acetate	✔	✔	Q
	Peonidin-3-monoglucoside-caffeoate	✔	✔	Q
	Cyanidin-3-monoglucoside-*p*-coumarate	✔	✔	Q
	Malvidin-3-monoglucoside-caffeoate	✔	✔	Q
	Petunidin-3-monoglucoside-*p*-coumarate	✔	✔	Q
	Cis-Malvidin-3-monoglucoside-*p*-coumarate	✔	✔	Q
	Peonidin-3-monoglucoside-*p*-coumarate	✔	✔	Q
	Malvidin-3-monoglucoside-*p*-coumarate	✔	✔	Q
Total skin anthocyanin (TSA)		✔		Q
Sum of anthocyanin groups	Sum of delphinidin derivatives	✔	✔	Q
	Sum of cyanidin derivatives	✔	✔	Q
	Sum of petunidin derivatives	✔	✔	Q
	Sum of peonidin derivatives	✔	✔	Q
	Sum of malvidin derivatives	✔	✔	Q
	Sum of monoglucosides	✔	✔	Q
	Sum of acetate derivatives	✔	✔	Q
	Sum of coumarate derivatives	✔	✔	Q
	Sum of caffeoate derivatives	✔	✔	Q
Ratio	Sum of coumarate/Sum acetate		✔	Q
	Sum trihydroxylated/Sum dihydroxylated		✔	Q
Visual colour characterisations	Pulp colour (PC) (categories: white pulp/coloured pulp)			D
	Skin colour OIV 225 (SC) (categories: rose skin/red skin/grey skin/ dark red violet/blue black skin)			P
	Skin and pulp colour (SPC) (categories: rose and red skin with white pulp/ grey, dark red violet and blue black skin and white pulp/ coloured pulp)			P
	Skin and pulp (SPC′) (categories: rose skin and white pulp/red skin and white pulp/grey skin and white pulp/ dark red violet and white pulp/ blue black skin and white pulp/ coloured pulp)			P

The last column shows variable types, where Q, D and P mean quantitative, dichotomous and polychotomous, respectively.

respectively (Chapter III).

Principal component analysis (PCA) using anthocyanins concentration and RA of anthocyanins showed that cultivars differed by anthocyanidin type and acylation pattern. PCA using anthocyanins concentration showed that the different anthocyanidins were separated according to methylation level. The analysis using RA of anthocyanins separated anthocyanidins according to hydroxylation level. These patterns of anthocyanidin discrimination show variation at enzymatic activities of methyl transferase and hydroxylase (Chapter III).

Other variables which could interfere with anthocyanin content of the berries were measured. Traits related with the maturity of the berries were measured by the Central Laboratory of the Instituto Nacional de Investigação Agrária (INIA-Dois Portos) during the anthocyanin extraction. Berries brix degree (% m/m), sugars content (g/l), volumic mass (g/cm3) and probable alcohol (% v/v) were measured by refractometry. Total acidity (g/l tartaric acid) was measured by colorimetric titration (Curvelo-Garcia, 1988). All these measurements were performed using the OIV method (2009). Several viral infections were assessed with the ELISA test. These were grapevine virus B (GVB), grapevine fanleaf (GFLV), arabic mosaic (ArMV), grapevine fleck (GFKL) and grapevine leafroll-associated viruses (GLRaV1, GLRaV2, GLRaV3, GLRaV7). These tests were performed by the National Institute of Biologic Resources, Portugal. None of these covariates were significant using stepwise regression for the studied sample ($P < 0.01$). Therefore, these covariates were excluded from the association analyses. All statistical analyses were performed using SAS v9.1 (SAS Institute Inc., Carry, NC, USA).

2.3.4. Genotyping

Data on 20 SSR loci scattered across 18 different chromosomes for 149 cultivars were provided by the Istituto Agrario San Michele all'Adige (IASMA). These markers were independent. Although two pairs of markers were located on the same chromosome, the distance between them according to previous linkage studies was 5cM (Doligez *et al.*, 2006; Troggio *et al.*, 2007). The Polymorphic Information Content (PIC) was high, with an average of 0.7 (Appendix 8).

SNP genotyping was performed using KasPar technology at KBiosciences (Hertfordshire, UK). Quality control analysis was performed. SNPs with MAF < 0.02, HW deviation > 10 and with missingness > 20 % were removed from the data. A sample of 124 SNPs was obtained after filtering for these criteria (Appendix 17). Table IV-5 shows the number of SNPs on each candidate gene after filtering for these quality criteria. Appendix 18 shows schematic drawing of the candidate genes and the SNPs genotyped.

Pairwise LD values for each gene are shown in Figure IV-2 and Appendix 19. Black symbols represent significant D' values while white symbols represent non-significant values ($P < 0.01$). Overall, the D' values were very high. The *CHI*, *F3'H*$_B$, *LDOX* and *GST* genes are not presented because there were only three SNPs on each of them. In these four genes, as in the remaining genes, significant LD values of one were observed even at the larger base pair distances.

Table IV-5 List of SNPs selected for genotyping for association analysis.

Gene code	Number of SNPs tested
CHS_A	5
CHS_C	7
CHI	3
F3H	5
$F3'H_B$	3
DFR	12
LDOX	3
UFGT	19
MRP	14
GST	3
MYC_A	6
MYC_B	10
MYBCC	12
MYB9	4
MYB11	18
Total	124

2.3.5. Structure

Data on 20 SSR loci scattered across 18 chromosomes for 149 cultivars were provided by the Istituto Agrario San Michele all'Adige (IASMA). These 20 SSR were used to assess background structure in the population sample. The method developed by Pritchard *et al.* (2000) implemented in STRUCTURE software was used to estimate the number of subpopulations (*K)*. It was assumed that each individual drew some fraction of its genome from each of the *K* populations and that allele frequencies in these populations were correlated. The parameter α which was used to model the degree of admixture was inferred from the data (Pritchard *et al.,* 2000). Lambda, the parameter of the distribution of

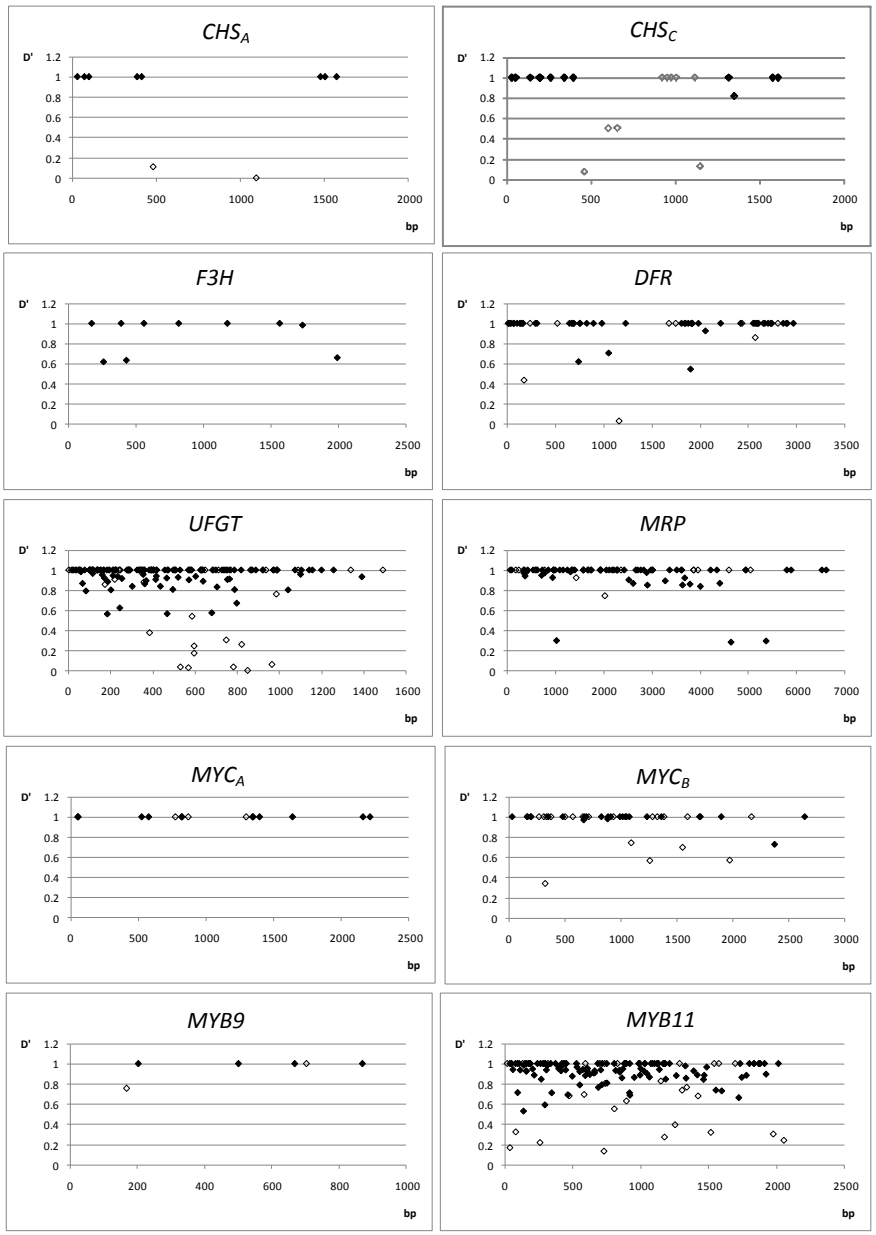


Figure IV-2 Plots of pairwise D' on the studied genes regions. Black significant at 1%, white non-significant.

allele frequencies, was set to the default value of one. The Markov chain Monte Carlo was performed with a burn-in period of 500 000 iterations followed by 500 000 iterations. According to pilot runs, summary statistics (alpha, divergence distances among populations and joint probability) were stable for this length and did not increase for a length of 1 000 000 iterations. Moreover, estimates of joint probabilities, allele frequencies and proportion of admixture were consistent between different runs for the same number of subpopulations (*K*).

In order to estimate the number of subpopulations, a total of 10 runs for each value of *K* were performed. Values of *K* ranged from *K* = 1 to *K* = 10. For *K* = 1 a total of one population was assumed by the model while for *K* = 2 a total of two populations were assumed, continuing in the same way up to 10 subpopulations.

Estimated posterior probabilities should be regarded only as rough guides to the number of subpopulations of the model that best fit the data. Therefore, several informal criteria are advised to select the best number of *K*. Firstly, the posterior probabilities for different *K* values should be considered jointly. It is commonly observed that the value of *K* lower than the true value has a very small posterior probability. Also, several values of *K* usually have similar estimates of posterior probability. In this case, the smallest value of *K* within these must be the true *K* value, since it is the smallest value of *K* that captures the major structure of the data. However, often the value of the posterior probability continues to increase with increasing values of *K*, making this criterion difficult to apply. A second criterion is variation of the parameter α. In the case of true structure, after Markov chain convergence, α will settle to relatively constant values with a range of 0.2 or less. Finally, when there is no population structure, the proportion of individuals assigned to each

subpopulation is near symmetric and most individuals are severely admixed (Pritchard *et al.*, 2009).

Figure IV-3 shows the estimated log probabilities of the current data for given values of *K*. The smaller value of log probability was obtained for $K = 1$. From $K = 2$ to $K = 7$ the log probability values were very similar, but increased with higher values of K instead of being stable. The highest log probability value was for $K = 10$ but a lot of variation was observed between runs. The largest log probability difference was between $K = 1$ and $K = 2$. Therefore, according to the informal criterion based on the smallest *K* within similar values of log probability for several *K*, $K = 2$ would be the most adequate number of subpopulations.

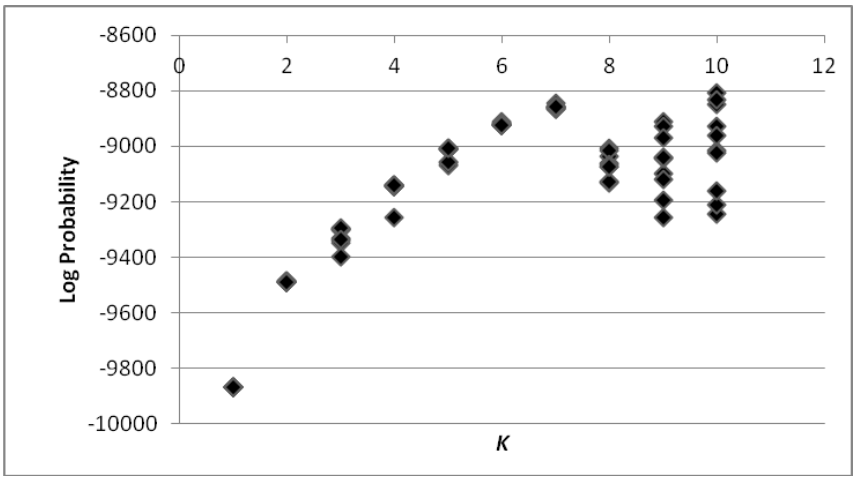


Figure IV-3 Plot of the log probability of data as a function of *K*.

However, alpha showed a high range towards the end of iterations (≈0.2) and the estimates of the proportion of each individual's variation that came from each subpopulation was near symmetric, suggesting the absence of structure (Pritchard, 2009). The matrix with these values is shown on Appendix 20. Figure IV-4 shows that the proportion of individuals assigned to each subpopulation was near symmetric (0.433;

0.567). Data on the same runs were also used to assess the number of subpopulations according to Evanno *et al.* (2005). This method also gave $K = 2$; however, it is not appropriate to detect the absence of structure since it cannot find the optimal K if $K = 1$.

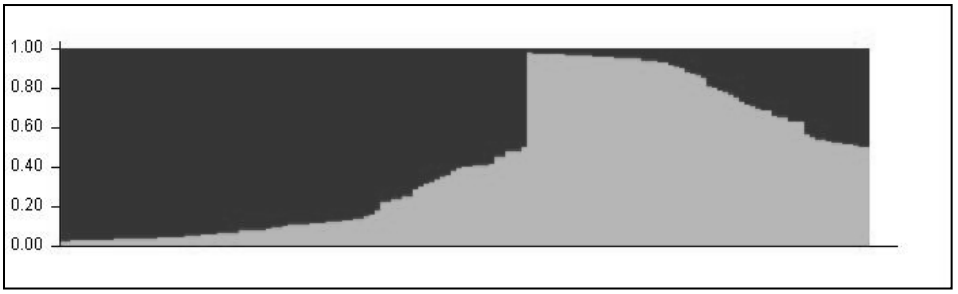


Figure IV-4 Plot of estimates of each individual estimated membership in each subpopulation. Each individual is represented by a single vertical line broken in as many coloured segments as subpopulations estimated. The y axis shows in each colour the estimated membership to each subpopulation.

Evanno's method (2005) uses ΔK, an ad hoc quantity based on the second order rate of change of the likelihood function with respect to K, to assess the best value of K. This method was calculated as:

$$\Delta K = mean\frac{\left[\left|\Pr(X|K+1) - 2\Pr(X|K) + \Pr(X|K-1)\right|\right]}{SD\left[\Pr(X|K)\right]}$$

Where *mean* represents the mean and SD represents standard deviation across runs and $\Pr = (X|K)$ means the posterior probability of the data for a given value of K (number of subpopulations).

The modal value of ΔK distribution was observed to be at the true K for most the situations investigated by Evanno *et al.* (2005). Although it has been shown that the number of subpopulations is better detected by this method than by the estimated log probability of data, it was also emphasised that this is an *ad hoc* approach. This method is not useful to detect the absence of structure since it cannot find the best K if $K = 1$.

Figure IV-5 shows the plot of delta *K* as a function of *K* calculated according to Evanno *et al.* (2005). According to this method the number of subpopulations (*K*) was two, since the mode of *ΔK* is observed at *K* = 2.

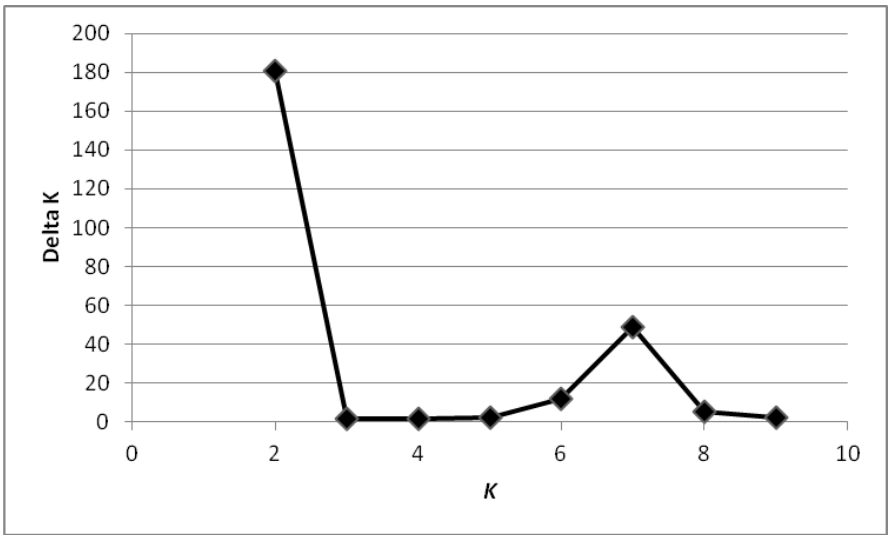


Figure IV-5 Plot of delta *K* as a function of k calculated according to Evanno's method (Evanno *et al.*, 2005)**.**

2.3.6. Relatedness

The same data (20 SSR loci) were used to measure pairwise relationships using two different methodologies. Firstly, a pairwise relationship based on the proportion of shared alleles (PSAR) between pairs of individuals was calculated as proposed by Chakraborty and Jin (1993):

$$P = \frac{\sum_{l=1}^{L} S}{2L}$$

Where *S* is the number of shared alleles and *L* is the number of total genotyped loci.

A PERL script was written for this purpose. The second measure of pairwise relationship was based on Ritland's (1996) kinship coefficient (RKCR). The software SPAGeDi (Hardy and Vekemans, 2002) was used to perform the calculations according to the following formula:

$$F_{ij} = \sum_l ((\sum_a \sum_{ci} \sum_{cj} (x_{lcia} x_{lcja} / p_{la}) / \sum_{ci} \sum_{cj} 1) - 1) / \sum_l (m_l - 1)$$

Where *Fij* is the kinship coefficient between individuals *i* and *j*; p_{la} is the frequency of allele *a* at locus *l* in the reference sample; x_{lcia} is an indicator variable (x_{lcia}= 1 if the allele on chromosome *c* at locus *l* for individual *i* is *a*, otherwise x_{lcia}= 0); m_l is the number of different alleles found in the sample at locus *l; Σci* stands for the sum over the homologous chromosomes of individual *i.*

The relatedness matrix based on RKC was transformed prior to any analysis. Negative values were set to zero, as this means that they are less related than a random pair of individuals (Hardy and Vekemans, 2002). Diagonals were equal to 1 + F , where F was the inbreeding coefficient obtained by SPAGeDi (Falconer and Mackay, 1996; Wright, 1922; Zhang *et al.,* 2009). All the off-diagonals were transformed to pairwise relationships between cultivars by multiplying by two the kinship coefficient (Falconer and Mackay, 1996; Wright, 1922). The matrix of PSA was used in the mixed model without any alterations. The matrices based on these two methods are shown on Appendices 21 and 22. Both methods, revealed some degree of relatedness among the individuals in the sample. Correlation between the two pairwise estimates was high (0.71697, $P < 0.0001$) and was confirmed by permutation analysis (Mantel, 1967). Relatedness based on RKC ranged from -0.19751 to 0.81851, while relatedness based on the PSA ranged from 0.075 to 0.8.

Table IV-6 shows the difference between the two methods for different percentiles of the distributions. Table IV-7 shows the highest pairwise relationship values between individuals for both methods. Pairwise relationships between individuals were higher when measured as the proportion of shared alleles (Table IV-7). RKCR ranged from -0.19751 to 0.81851, while PSAR ranged from 0.075 to 0.8 (Table IV-6; Appendices 5 and 6).

Table IV-6 Percentile distribution of relatedness values.

	Relatedness values based on RKC multiplied by 2	Relatedness values based on PSA
Percentile 90	0.094	0.500
Percentile 75	0.031	0.425
Percentile 50	-0.022	0.375
Percentile 25	-0.064	0.325
Percentile 10	-0.095	0.275

Table IV-7 Ten highest pairwise relatedness values obtained.

Pairwise individuals ID		Relatedness values based on RKC multiplied by 2	Pairwise individuals ID		Relatedness values based on PSA
50801	53606	0.819	41502	50804	0.800
41607	50201	0.760	41504	50204	0.800
50702	51208	0.710	50301	52506	0.800
51708	53608	0.670	50303	53103	0.775
50603	50701	0.647	50105	53704	0.763
53305	53505	0.644	50301	53102	0.750
53305	53306	0.640	50303	53107	0.750
50102	51106	0.608	51107	51711	0.750
50105	53704	0.586	52105	53208	0.750
53306	53505	0.585	52205	52905	0.750

2.3.7. Association models

Several models were applied and compared. A list of the different models used is shown on Table IV-8 and a list of the comparisons performed on Table IV-9. Model comparisons were performed using the phenotype of TSA concentration. The selected models were used to test association for the remaining phenotypes.

Only single SNP tests were performed. It was decided not to perform haplotype tests. On one hand side, this decision was based on the fact that haplotype estimation is associated to a considerable amount of error. On the other hand, it is not consensual if higher power is or not obtained with this approach. Moreover, very often haplotype analysis helps select a genomic region to be further explored. Testing haplotype association in this case would be quite uninformative since very fine mapping was performed. The same genes showing association with phenotypes on single SNP tests would most likely show associated haplotypes.

The simplest model tested, Model A, was a linear regression with phenotype as the response variable and SNP as the predictor variable. The genotypes AA, Aa, aa were coded as 0, 1 and 2, respectively. Model B was as Model A but it included the relationship matrix based on PSA as a random effect and structure as a covariate. Model C was identical to Model B, but RKC matrix substituted PSA matrix. Model D was similar to Model A but with structure added. Finally, models E and F were similar to models B and C but excluding structure (Table IV-8). The SAS PROC MIXED procedure was used for all the mixed model analyses.

Table IV-8 List of the statistical models tested.

Model name	Model in matrix notation	Description	
A	$y = X\beta + e$	*PHE = SNP*	Regression model where phenotype is the response variable and genotype (0, 1, 2) is the independent variable.
B	$y = X\beta + Q\gamma + Z\delta + e$	*PHE = SNP + Q + PSAR*	Same as Model A, but with structure and relatedness based on PSA included.
C	$y = X\beta + Q\gamma + Z'\delta' + e$	*PHE = SNP + Q + RKCR*	Same as Model A, but with structure and relatedness based on RKC included.
D	$y = X\beta + Q\gamma + e$	*PHE = SNP + Q*	Same as Model A, but with structure included.
E	$y = X\beta + Z\delta + e$	*PHE = SNP + PSAR*	Same as Model A, but with relatedness based on PSA included.
F	$y = X\beta + Z'\delta' + e$	*PHE = SNP + RKCR*	Same as Model A, but with relatedness based on RKC included.

In the formula y is a vector of phenotypic observations (TSA concentration), β is a vector of SNP allele effects to be estimated, X contains the genotypes, γ is a vector of population structure effects, Q is a matrix with the proportion of individuals genome inherited from ancestors in each subpopulation inferred by STRUCTURE, δ is a vector of random effects due to relatedness based on PSA, Z is an incidence matrix relative to δ, δ' is a vector of random effects due to relatedness based on RKC, Z' is an incidence matrix relative to δ', and e is a vector of residual effects. In the schematic representation *PHE* means phenotypic data (TSA concentration), *SNP* means genotype data on SNPs, *Q* means population structure measured as the proportion of individuals genome inherited from ancestors in each subpopulation, *PSAR* represents pairwise relationships between individuals based on the PSA and *RKCR* represents the pairwise relationships between individuals based on RKC.

To assess the importance of structure in association analyses, two models were compared by F-test (comparison 1; Table IV-9), according to the following formula:

$$F = \frac{\frac{ESSr - ESSf}{Edfr - Edff}}{\frac{ESSf}{Edff}}$$

Where *ESS* represents the error sum of squares, *df* represents the degrees of freedom, *r* represents the reduced model and *f* the full model.

Table IV-9 List of model comparisons performed.

		Test	Reduced model	Full model	Objective
Model comparison	1	F-test	A *(PHE = SNP)*	D (*PHE = SNP + Q)*	Assess importance of structure in the association model.
	2a	Likelihood Ratio Test	D (*PHE = SNP + Q)*	B (*PHE = SNP + Q + PSAR)*	Assess importance of relatedness using PSA in the association model.
	2b	Likelihood Ratio Test	D (*PHE = SNP + Q)*	C *(PHE = SNP + Q + RKCR)*	Assess importance of relatedness using RKC in the association model.
	3	Likelihood Ratio Test	E (*PHE = SNP + PSAR)*	E' (*PHE = SNP + PSAR)*	Assess differences in importance of relatedness in the association model using RKC or PSA.

E' stands for model E but with covariance parameters set using the covariance parameters estimated for model F.

In this comparison, Model A is the reduced model (*r*), where the phenotype was regressed on genotype and the full model (*f*) is model B where the phenotype was regressed on genotype and structure. The F-test showed significant differences ($P < 0.05$) for 94.35 % of the markers. The percentage of markers for which model B was significantly different from Model A was very high and therefore it was decided to include structure in further analysis.

To assess the importance of relatedness (comparisons 2a and 2b; Table IV-9) in the association model, a likelihood ratio test was performed

according to the following formula:

$$\chi^2 = -2\ln likelihood\ nm - (-2\ln likelihood\ \ fm)$$

$$df = number\ of\ \ parameters\ of\ \ the\ nm - number\ of\ \ parameters\ of\ \ the\ \ fm$$

Where *nm* means null model and *fm* means full model.

In these comparisons (2a and 2b; Table IV-9), the reduced model was model D where phenotype was the dependent variable, genotype the independent variable and structure effects were included. The full models were models D1 and D2. These were mixed models that contained genotype and structure but also the relationship matrix as random effects. For comparison 2a, the Model B included the relatedness matrix based on PSA. For comparison 2b, the model C was the same as B but the random effects were the relatedness matrix was based on RKC. These comparisons revealed significant differences ($P < 0.05$) for both relatedness measures. For comparison 2b, model D was significantly different from the model C for 99.19 % of the markers. For comparison 2a, the models were significantly different for 97.58 % of the markers. As the percentage of markers for which the models B and C were significantly different from model D was very high, relatedness was included in further analysis.

To assess the significance of the two different measures of relatedness, a likelihood ratio test was used (comparison 3; Table IV-9). Comparing the covariance parameters for Models E and F we found that the two matrices were not significantly different for all markers ($P < 0.01$). Only one SNP was significant for $P < 0.05$. Therefore, relatedness based on PSA was used for further analysis since it raises fewer problems with convergence and non-positive definiteness.

Tests of association were performed with two different models, Model

A and Model B. All statistical analyses were performed with the original phenotypes and the log transformed values but the results were essentially the same. Throughout the study we only present the analyses performed with the original phenotypes. The results for all the tests are shown on Appendices 23 - 27.

2.3.8. Statistical interactions

Interactions between SNPs at different genes have recently received a great deal of attention (Cordel, 2009). Examining a gene in isolation when it functions through a mechanism involving other genes might hinder the effect it has on the phenotype. When a locus variation affects a phenotype by interacting with another locus, using a statistical model that allows for this interaction will increase the power to detect the effect of this locus.

Gene-gene interactions were tested using SNP data on a model for two-locus interactions. Statistically, these tests were performed by using a multiple regression model where the phenotype was regressed on genotype on locus 1, genotype on locus 2 and on interaction between locus 1 and 2, according to the following formula:

$$y = b_1 x_1 + b_2 x_2 + b_{12} x_{12} + e$$

y represents phenotypic observations; x_1 is SNP1 genotypes; b_1 is the regression coefficient for x_1; x_2 is SNP2 genotypes; b_2 is the regression coefficient for x_2; x_{12} is interaction between the two loci; b_{12} is the regression coefficient for x_{12}.

Interaction tests were performed for TSA concentration between each of three genes coding transcription factors (*MYB11*, *MYBCC*, MYC_B) and the remaining genes. These three transcription factors were selected for interactions due to the significant associations shown with TSA

concentration on single SNP tests. Transcription factors were also especially interesting since previous works have shown interactions among different transcription factors, between these and genes involved in the biosynthetic pathway of anthocyanins and also between transcription factors and genes related to anthocyanin transport (Bogs *et al.*, 2007; Cutanda-Perez *et al.*, 2009; Fournier-level *et al.*, 2009; Matus *et al.*, 2009; Terrier *et al.*, 2009).

2.3.9. Permutations

A total of 10 000 permutations were performed for TSA concentration, PC and SPC under Model A. For the SNP-SNP interaction analyses, 1000 permutations were performed only for the highest significant interaction of each gene pair with high proportion of SNPs involved in significant interactions (> 25 %). Multiple test corrections are necessary due to the high number of SNPs and phenotypes considered. Bonferroni correction assumes marker independence. This assumption was not met by the genotyped SNPs since strong LD was observed. Therefore, Bonferroni correction would be too conservative.

Permutations were used to estimate the significance empirically and thus avoiding assumptions about normality and problems with Type I errors due to multiple testing. The phenotype was randomly shuffled one thousand times. Permutations were performed for TSA concentration and for SPC. The nominal *P*-value was ranked in order to calculate the empirical *P*-value according to the following formula:

$$Empirical\ P\text{-}value = \frac{rank\ of\ nominal\ P\text{-}value\ in\ the\ total\ observed\ P\text{-}values}{total\ number\ of\ observed\ P\text{-}values}$$

3. Results

3.1. Association results for single SNP tests

This section gives a detailed account of the results from TSA concentration, PC and SPC. However, detailed results for all the association tests performed (i.e. all phenotypes) may be found on Appendices 23 and 26.

TSA concentration was considered the main phenotype since it shows colour variation in a rather accurate way and includes concentration of different types of anthocyanins. Visual characterisation of SPC and PC were considered important phenotypes since they showed association with both anthocyanins concentration and RA. Also, these phenotypes are important for future replication studies, since they are based on visual characterisation of colour and may therefore be easily obtained. In addition, these phenotype showed significant associations with a large number of SNPs, while SPC' was associated with only two SNPs and SC did not show any significant associations.

Figure IV-6 shows the results of the associations using models A and B. The –log10 of the *P*-values were plotted on all genes for TSA concentration, SPC and PC. Table IV-10 shows the *P*-values for SNPs associated with TSA concentration, SPC and PC for models A and B.

Five SNPs (**s36**, s65, **s68**, s89, **s90**) in three different genes coding transcription factors (*MYB11*, *MYBCC* and MYC_B) yielded significant associations with TSA concentration ($P < 0.01$) using Model A (Table IV-10). Similar levels of significance for these genes were also observed for **s36**, **s68** and **s90** with the mixed model (Model B, Table IV-10). These three SNPs were also significantly associated with SPC and PC under both models (Table IV-10).

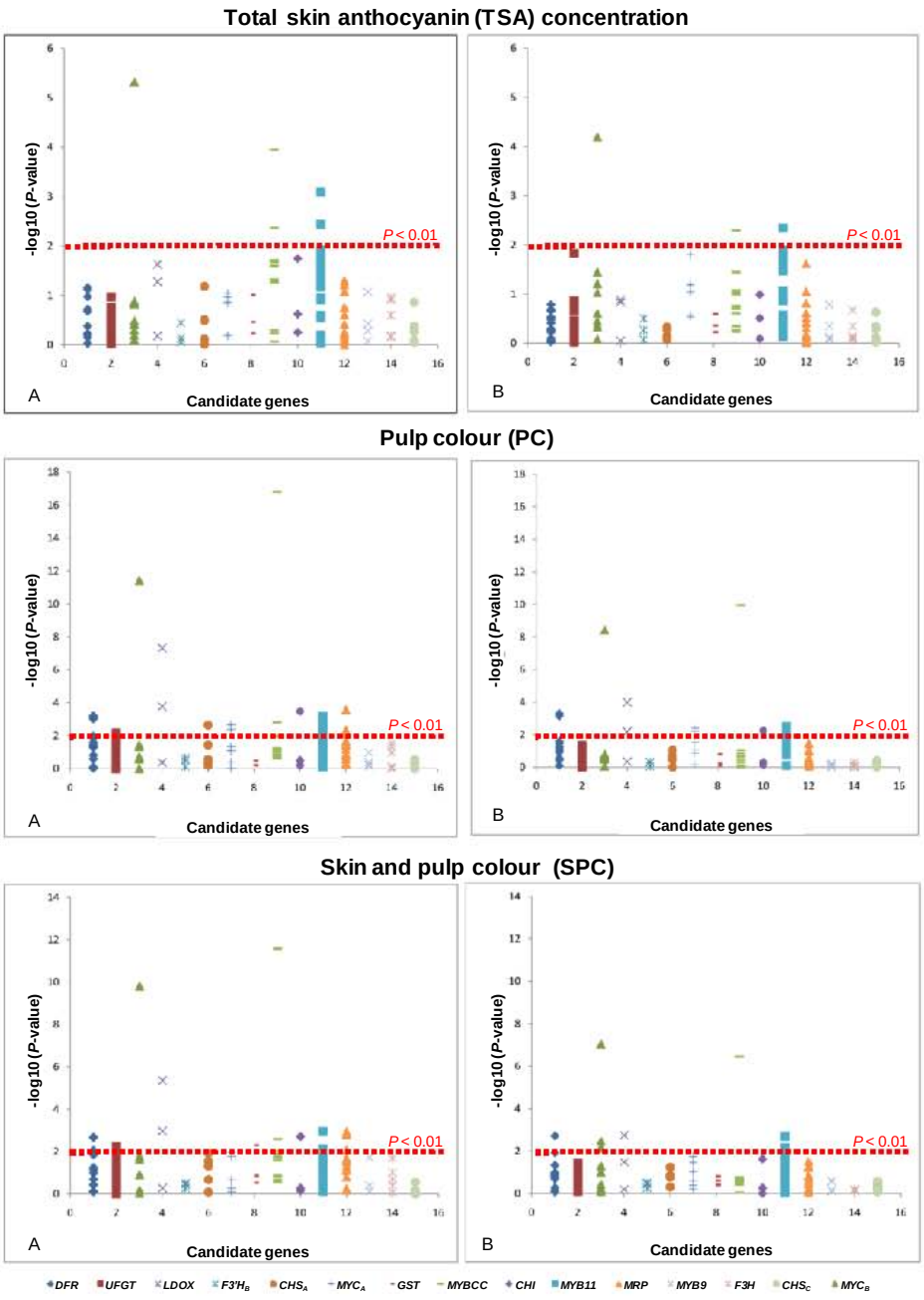


Figure IV-6 Results of association test of TSA concentration, PC and SPC. Graphs A and B show results for tests of association based on models A and B, respectively. The *y* axis shows -log10 (*P*-values). The different genes studied are shown along *x* axis and identified according to the legend colour code.

Table IV-10 List of SNPs showing significant associations with total skin anthocyanin (TSA) concentration, pulp colour (PC) and skin and pulp colour together (SPC).

			TSA concentration		PC		SPC	
Gene	SNP ID	MAF	Model A	Model B	Model A	Model B	Model A	Model B
MYB11	s80	0.40			$9.68x10^{-03}$			
	s83	0.38						$9.34x10^{-03}$
	s84	0.49				$3.11x10^{-03}$		$2.04x10^{-03}$
	s86	0.40				$5.14x10^{-03}$		$7.11x10^{-03}$
	s87	0.07			$2.95x10^{-03}$			
	s89	0.33	$3.64x10^{-03}$		$1.84x10^{-03}$		$7.40x10^{-03}$	
	s90	**0.06**	**$8.29x10^{-04}$**	**$4.64x10^{-03}$**	**$6.79x10^{-04}$**		**$1.14x10^{-03}$**	**$9.73x10^{-03}$**
	s93	0.07			$6.67x10^{-03}$			
	s94	0.35			$5.43x10^{-03}$			
MYBCC	s65	0.21	$4.31x10^{-03}$		$1.47x10^{-03}$		$9.61x10^{-03}$	
	s68	**0.06**	**$1.12x10^{-04}$**	**$5.14x10^{-03}$**	**$1.64x10^{-17}$**	**$1.10x10^{-10}$**	**$2.62x10^{-12}$**	**$3.39x10^{-07}$**
	s71	0.48					$2.55x10^{-03}$	
MYC_B	s33	0.32						$6.60x10^{-03}$
	s34	0.32						$3.29x10^{-03}$
	s36	**0.15**	**$4.77x10^{-06}$**	**$6.29x10^{-05}$**	**$3.66x10^{-12}$**	**$3.66x10^{-09}$**	**$1.51x10^{-10}$**	**$8.84x10^{-08}$**
	s37	0.31						$7.21x10^{-03}$
	s40	0.31						$7.17x10^{-03}$
MYC_A	s55	0.26			$2.16x10^{-03}$	$3.90x10^{-03}$		
	s58	0.27			$4.00x10^{-03}$	$5.88x10^{-03}$		
CHS_A	s49	0.22			$2.21x10^{-03}$			
CHI	s75	0.44			$3.24x10^{-04}$	$5.42x10^{-03}$	$1.95x10^{-03}$	
DFR	s1	0.33			$9.10x10^{-04}$	$5.22x10^{-04}$	$8.72x10^{-03}$	
	s11	0.03			$6.18x10^{-04}$	$7.41x10^{-04}$	$2.14x10^{-03}$	$1.86x10^{-03}$
LDOX	s42	0.19			$4.63x10^{-08}$	$1.49x10^{-04}$	$4.41x10^{-06}$	$1.80x10^{-03}$
	s44	0.28			$1.65x10^{-04}$	$6.33x10^{-03}$	$1.07x10^{-03}$	
UFGT	s20	0.31			$9.79x10^{-03}$			
	s22	0.32					$9.94x10^{-03}$	
	s29	0.34					$8.02x10^{-03}$	
	s30	0.37			$6.78x10^{-03}$		$5.77x10^{-03}$	
MRP	s95	0.26			$2.64x10^{-04}$		$1.42x10^{-03}$	
	s98	0.42			$7.59x10^{-03}$		$8.28x10^{-03}$	
	s100	0.30			$6.85x10^{-03}$		$1.62x10^{-03}$	
	s102	0.31			$4.33x10^{-03}$		$1.13x10^{-03}$	
GST	s59	0.15			$8.02x10^{-03}$		$4.99x10^{-03}$	

This table shows significant nominal *P*-values obtained under Models A and B ($P < 0.01$). Significance was confirmed by 10 000 permutations for Model A. Some SNPs that were not significant for TSA concentration, PC and SPC but were associated with other phenotypes are shown on Appendices 23 and 25.

SPC showed significant associations with 19 SNPs under Model A ($P < 0.01$). These 19 SNPs were distributed across nine genes (*DFR*, *UFGT*, *LDOX*, *CHI*, *GST*, *MRP*, MYC_B, *MYBCC*, *MYB11*), three genes encoding transcription factors, four involved in the biosynthetic pathway and two involved in the transport of anthocyanins (Table IV-10). Only five of these genes were significantly associated with this phenotype for Model B. However, for *MYB11* and MYC_B the overall number of significant SNPs increased (Table IV-10). Under Model A, PC was associated with 24 SNPs ($P < 0.01$) on a similar group of 11 genes (*MYB11*, *MYBCC*, MYC_A, MYC_B, *GST*, *MRP*, *UFGT*, *DFR*, *LDOX*, *CHI* and CHS_A). Only seven genes showed association with this phenotype for Model B. Nine of these genes were associated with both PC and SPC (Table IV-10).

3.1.1. Genes coding transcription factors (*MYB11*, *MYBCC*, MYC_B, *MYB9*, MYC_A)

Table IV-11 presents the percentage of SNPs and phenotypes that showed at least one significant association ($P < 0.01$) on each of the genes.

Table IV-12 shows the percentage of variable groups associated with each gene under Model A. Each variable group represents only those phenotypes that are relative to concentration, RA, anthocyanidin type, acylation type and ratios. The percentages shown concern the proportion of each variable group among the variables associated with SNPs in each gene.

At the *MYB11* locus two SNPs (s89 and **s90**) were significantly associated with TSA concentration, PC and SPC under Model A ($P < 0.01$; Table IV-10). SNP s89 is synonymous and **s90** is in the predicted promoter region. For TSA and SPC, **s90** was significant for both models

showing the strongest association for TSA concentration under Model A (P = 8.3x10^{-04}; Table IV-10). The SNP s89 was not significantly associated with any of these three phenotypes using Model B (P < 0.01; Table IV-10). Three additional SNPs (s83, s84 and s86) were associated with SPC for Model B only (P < 0.01). Also s80, s87, s93 and s94 were associated with PC under Model A, and s84 and s86 for Model B (P < 0.01; Table IV-10).

Two intronic SNPs within *MYBCC* (s65 and **s68**) were associated with TSA concentration, PC and SPC (P < 0.01; Table IV-10) under Model A. The SNP **s68** was significant for these three phenotypes with both models. The strongest associations were found between **s68** and PC using Model A (P = 1.64x10^{-17}) and Model B (1.10x10^{-10}; Table IV-10). A third SNP (s71), also located on an intron region, was found to be associated with SPC for Model A (P = 2.55x10^{-03}; Table IV-10). In *MYC*$_B$, one synonymous SNP (**s36**) showed association with the three phenotypes under all the models (P < 0.01). The highest significance was observed for PC with Model A (3.66x10^{-12}; Table IV-10). This SNP was also associated with a large percentage of phenotypes (33 %; Appendix 28). Four additional SNPs (s33, s34, s37 and s40) showed association with SPC (P < 0.01; Table IV-10). All the associations using Model A were verified empirically through permutations.

For these three genes, a high percentage of SNPs (approximately 80 %) were found to be associated with at least one of the phenotypes (P < 0.01; Table IV-11). In *MYB11*, over half of the phenotypes (53 %) were associated with at least one of the SNPs (P < 0.01). Association tests with Model B yielded similar results (Table IV-11). These phenotypes were mainly acetate and coumarate derivative anthocyanins (Table IV-12). In *MYBCC* many phenotypes (27.9 %) were significantly associated with at

least one SNP under Model A (Table IV-11). Ten SNPs out of 12 (83.3 %) associated with at least one of these phenotypes for this model (P < 0.01; Table IV-11). However, **s68** yielded the strongest signal (P < 0.001; Appendix 23). The phenotypes were mainly concentrations of different types of anthocyanins (Table IV-12). Peonidin-3-monoglucoside concentration was significantly associated with seven SNPs in this gene (58.3 %) (P < 0.01; Appendix 23). Association tests using the mixed model (Model B) showed a smaller number of significant associations (Table IV-11). However, **s68** was significant for several phenotypes (20 %) for Model B (P < 0.01; Appendix 28). In MYC_B, 44.3 % of the phenotypes were associated with at least one of the SNPs (Table IV-11).

Pairwise D′ was generally very high between SNPs within *MYB11*, *MYBCC* and MYC_B. LD between SNPs s89 and **s90** on *MYB11* was 0.89 (P < 0.01; Appendix 19). Minor allele frequency (MAF) varied for these two SNPs (0.33 and 0.06 for s89 and **s90**, respectively) (Table IV-10). Complete LD was found between the pairs of SNPs **s68**-s65 and **s68**-s71. Between s65 and s71, and s34 and **s36** LD was low but not significant (Appendix 19). In *MYBCC*, MAF varied for the significant markers with **s68** being the rarest (0.06) and s71 being the most common (0.48) (Table IV-10). SNPs within MYC_B showed MAF between 0.06 and 0.32 (Appendix 28). MAF of **s36** was 0.15 while the remaining SNPs (significant for SPC) showed MAF near 0.3 (Table IV-10), which may explain the different results for association.

Overall, *MYB9 and* MYC_A genes did not reveal any associations with TSA concentration and SPC. Two SNPs (s55 and s58), however, within MYC_A showed association with PC under both models (Table IV-10).

Table IV-11 Percentage of SNPs and phenotypes showing significant association ($P < 0.01$) for each gene.

		Model A		Model B	
Genes		**Percentage of SNPs associated with at least one phenotype**	**Percentage of phenotypes associated with at least one SNP**	**Percentage of SNPs associated with at least one phenotype**	**Percentage of phenotypes associated with at least one SNP**
Transcription factors	*MYC_A*	50.00	8.20	66.67	9.84
	MYC_B	80.00	44.26	50.00	36.07
	MYB9	25.00	1.64	0.00	0.00
	MYB11	83.33	52.46	77.78	44.26
	MYBCC	83.33	27.87	41.67	24.59
Biosynthetic pathway of anthocyanins	*CHS_A*	40.00	9.84	20.00	1.64
	CHS_C	14.29	8.20	28.57	6.56
	CHI	100.00	14.75	100.00	4.92
	F3H	40.00	6.56	60.00	11.48
	$F3'H_B$	0.00	0.00	0.00	0.00
	DFR	50.00	13.11	41.67	14.75
	LDOX	66.67	9.84	66.67	4.92
	UFGT	57.89	34.43	10.53	27.87
Vacuole accumulation	*MRP*	57.14	39.34	42.86	18.03
	GST	100.00	4.92	66.67	1.64

Table IV-12 Percentage of variable groups associated with each gene under Model A.

Variable groups*	Candidate genes														
	MYC$_A$	***MYC$_B$***	***MYB9***	***MYB11***	***MYBCC***	***CHS$_A$***	**CHS$_C$**	***CHI***	***F3H***	***F3'H$_B$***	***DFR***	***LDOX***	***UFGT***	***MRP***	***GST***
Relative abundance	67	62	100	31	14	50	100	30	75		46		65	49	50
Concentration		32		60	70	33		50	25		18	50	21	36	
Delphinidin derivatives	33			8					25				15	21	
Cyanidin derivatives		21		15	3		20								
Petunidin derivatives	33	9		10			20						15	6	
Peonidin derivatives		19		13	38	67		20	25		27	25	33	38	
Malvidin derivatives		15		23	16	17	20	40	25		18	13	6	11	50
Glucoside derivatives	33	23		7	43	33	20	20	25		9	25	21	21	
Acetate derivatives		13		39		17							27	8	
Coumarate derivatives		34		39	14		80	30	25		36		9	28	50
Caffeoate derivatives		4			3			10						2	
Visual characterisation	33	6		8	16	17		20			36	50	15	15	50
Ratios			100						25		9		12	6	

*Each variable group represents only those phenotypes that are relative to concentration, RA, anthocyanidin type, acylation type and ratios. Values were rounded to the unit.

3.1.2. Genes coding enzymes involved on the biosynthetic pathway of anthocyanins (*CHS_A*, *CHS_C*, *CHI*, *F3H*, *$F3'H_B$*, *DFR*, *LDOX*, *UFGT*)

The SNP s30 on *UFGT* was associated with both PC and SPC under Model A ($P < 0.01$; Table IV-10). Under this model two other SNPs (s22 and s29) were significantly associated with SPC and one (s20) with PC ($P = 9.79x10^{-03}$; Table IV-10). SNP s11 within *DFR* and in the 3′UTR region was associated with PC and SPC for both models (Table IV-10). Another SNP (s1) in the predicted promoter region was associated with PC and SPC using Model A (Table IV-10). Both SNPs were associated with PC using Model B ($P < 0.01$) (Table IV-10). On *LDOX* the SNP s42 in the 3′UTR region associated under both models with PC and SPC showing the strongest significance for PC ($P = 4.63x10^{-08}$). The SNP s44 in the promoter region was associated with PC and SPC for Model A and with PC under Model B ($P < 0.01$; Table IV-10). The non-synonymous SNP s75 in *CHI* was associated with PC and SPC using Model A, and with PC for Model B ($P = 5.42x10^{-03}$; Table IV-10).

Overall, a high percentage of the SNPs (near or above 50 %) within these four genes that are involved on the biosynthetic pathway of anthocyanins, showed association with at least one of the phenotypes under Model A. *DFR*, *LDOX* and *CHI* showed similar percentages for Model B as well (Table IV-11). A large proportion of phenotypes (34.4 %) within *UFGT* were significantly associated with at least one of the SNPs using Model A (Table IV-11). Interestingly, SNP s25, which causes an amino acid substitution, was the most important as it was associated with 25 % of total phenotypes (Appendix 28). Most phenotypes were included in the RA variable group, especially involving peonidin derivatives (Table IV-12). Only two SNPs (s22 and s25) showed associations under Model B ($P < 0.01$; Appendix 26).

In general, these four genes showed great variation on MAF but LD was very high (Appendices 19 and 28). MAF ranged between 0.06 and 0.37 for SNPs within *UFGT*. SNPs s25 was the rarest (MAF = 0.06) among the SNPs that were found significant for PC and SPC and with frequencies near 0.3 (Table IV-10). For the *DFR* gene, s1 was quite common (MAF = 0.33) in contrast to s11 (MAF = 0.03). For *LDOX*, s42 showed a MAF of 0.19 while s44 was more common (MAF = 0.28; Table IV-10). SNP s75 was the most frequent SNP genotyped within *CHI* (MAF = 0.44; Appendix 28). Across these genes, values of D′ between the significant markers for TSA concentration, PC and SPC, were between 0.81 and 1 (Appendix 19).

No other genes coding enzymes involved on the biosynthetic pathway of anthocyanins showed SNPs significantly associated with TSA concentration or with SPC. CHS_A showed one SNP (s49) to be associated with PC ($P = 2.21 \times 10^{-03}$) (Table IV-10).

3.1.3. Genes coding enzymes involved on the transport of anthocyanins to the vacuole (*MRP, GST*)

Four SNPs within *MRP* gene (s95, s98, s100 and s102) were associated with PC and SPC under both models (Table IV-10). SNP s98 is synonymous, while the other three cause amino acid changes. For the *GST* gene, the intronic SNP (s59) was found to be associated with SPC and PC for Model A (Table IV-10). All the SNPs within *GST* showed significant associations with at least one of the 61 phenotypes using Model A and 66.7 % under Model B (Table IV-11).

Using Model A, more than half of the SNPs (57.1 %) within *MRP* genes were associated with the remaining phenotypes. Overall, these associations included 24 phenotypes (Table IV-11). Six of these

polymorphisms were significantly associated with a relatively high number of phenotypes, between six and 12 among the total of 61 phenotypes (Appendix 23). Concentration and RA of Peonidin derivatives were a large proportion of the significantly associated phenotypes (Table IV-12). The association test using Model B showed a smaller number of SNPs (42.9 %) and phenotypes (18 %) to be significantly associated (Table IV-11) within *MRP* gene.

LD was high across the *MRP* and *GST* genes (Appendix 19). Pairwise LD estimates between the SNPs within the *MRP* gene that were significant for the TSA concentration, PC and SPC ranged between 0.87 and 1 (Appendix 19). The MAF for these SNPs varied between 0.26 and 0.42 (Table IV-10). SNP s59 showed the lowest MAF (0.15) for the three genotyped SNPs within the gene *GST* (Appendix 27).

Significant associations under Model A were confirmed empirically using 10 000 permutations. The association tests performed using the log transformed values of TSA concentration yielded similar results to models A and B (Appendices 25 and 27).

3.1.4. Statistical interactions

Tests for statistical interactions were performed between SNPs within each of three transcription factors (*MYB11*, *MYBCC*, MYC_B) and the remaining genes for TSA concentration. A number of significant interactions were observed between SNPs within several genes ($P < 0.001$). *P*-values for these interactions are shown on Appendix 29. Only pairs of genes showing more than 25 % of SNPs involved in significant interactions were further explored.

Figure IV-7 shows a schematic representation of the biosynthetic pathway of anthocyanins and of the pairs of genes within which SNP x

SNP interactions were observed. Scheme A shows a simplified biosynthetic pathway of anthocyanins. Scheme B shows the genes coding transcription factors with SNPs (> 25 %) involved in SNP x SNP interactions (P < 0.001). The interactions are represented by dashed arrows and the numbers beside the arrows indicate the number of SNPs involved in significant SNP x SNP interactions.

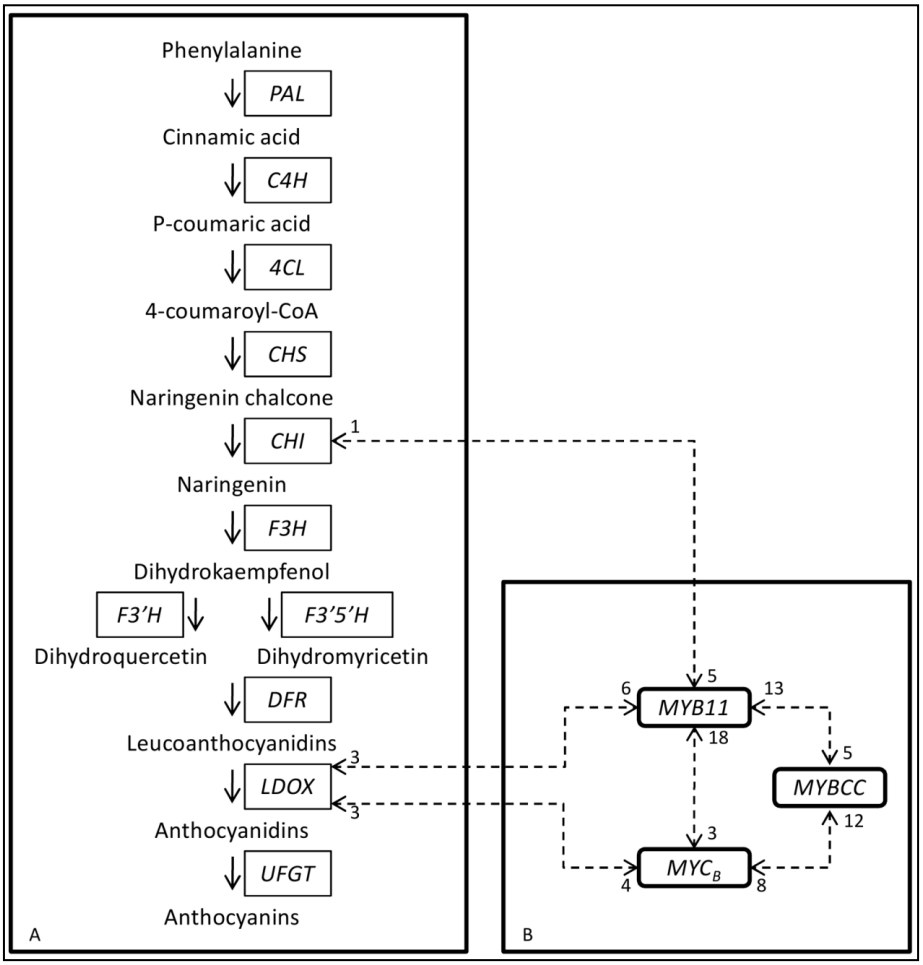


Figure IV-7 Schematic representation of the genes showing SNP x SNP interactions. Scheme A shows a simplified biosynthetic pathway of anthocyanins. Scheme B shows the genes coding transcription factors with SNPs (> 25%) involved in SNP x SNP interactions (P < 0.001). The interactions are represented by dashed arrows and the numbers beside the arrows indicate the number of SNPs involved in significant interactions.

Significant interactions between more than 25 % of the SNPs were identified between *MYB11* and the following four genes: *LDOX*, *CHI*, MYC_B and *MYBCC* ($P < 0.001$; Appendix 30). This was also observed between the gene coding MYC_B, and *MYBCC* and *LDOX*. Table IV-13 presents the top four SNP x SNP interactions for which the model significance and the interaction effects were significant. These SNP x SNP interactions were within the *MYB11*, *MYBCC*, MYC_B, *LDOX* and *CHI* genes. The results show that the significance of the model was stronger than any single SNP tests. The significance values on Table IV-13 were confirmed by 1000 permutation tests.

Table IV-13 Interactions between SNPs in different genes.

Interactions	Model *P*-value	Interaction *P*-value	Single SNP tests *P*-values			
$MYB11_{s93}$ x $LDOX_{s42}$	$4.3X10^{-04}$	$1.2X10^{-03}$	s93	$4.4X10^{-02}$	s42	$2.4X10^{-02}$
$MYB11_{s89}$ x CHI_{s75}	$2.4X10^{-05}$	$2.6X10^{-03}$	s89	$3.6X10^{-03}$	s75	$1.8X10^{-02}$
$MYB11_{s90}$ x MYC_{Bs36}	$1.2X10^{-06}$	$1.6X10^{-02}$	s90	$8.3X10^{-04}$	s36	$4.8X10^{-06}$
$MYB11_{s91}$ x $MYBCC_{s65}$	$5.0X10^{-05}$	$1.0X10^{-03}$	s91	$3.4X10^{-02}$	s65	$4.3X10^{-03}$
MYC_{Bs36} x $MYBCC_{s63}$	$4.9X10^{-07}$	$4.3X10^{-03}$	s36	$4.8X10^{-06}$	s63	$2.5X10^{-02}$
MYC_{Bs36} x $LDOX_{s42}$	$4.5X10^{-06}$	$1.6X10^{-02}$	s36	$4.8X10^{-06}$	s42	$2.4X10^{-02}$

The first column shows the name of the two genes and in subscript the name of the SNPs in the model. The last column shows the *P*-values of the singles SNP tests under Model A. Both tests were performed with the phenotype TSA concentration. The model *P*-values were confirmed by 1000 permutation tests.

4. Discussion

Population structure and more recently cryptic relatedness have been suggested as potential causes for spurious results in association studies. Relatedness in plant populations can be a problem since germplasm collections often select cultivars of high interest to breeders but which are related (Zhu *et al.*, 2008). Several statistical methods have been

developed to address these problems (Yu *et al.*, 2006; Pritchard *et al.*, 2000; Devlin and Roeder, 2009; Price *et al.*, 2006). However, the extent of the impact of cryptic relatedness in association analyses has not been studied extensively.

Here we examined structure and relatedness using two different matrices (PSA and RKC). We found that for more than 90 % of the SNPs, the simplest model was significantly different to the full model that considered both effects. However, the analyses of the main phenotype (TSA concentration) showed that the simple and full model yielded similar results. These associations with the simple model were examined both nominally and empirically. So the question that arises is whether the use of a less parsimonious model could lead to type II errors.

In this study relatedness was considered by two different matrices. One based on the PSA and the other on RKC. Our analyses showed that the two matrices were not statistically different which is in agreement with the simulation study performed by Zhao (Zhao *et al.*, 2007). The choice of a method to infer a relatedness matrix is debatable. Zhao (2007) showed that a matrix based on the PSA (Chakraborty and Jin, 1993) effectively corrects for cryptic relatedness. Kang (2008) showed that this matrix guarantees positive semidefiniteness and convergence when there is no missing data. However, this matrix does not take into account allele frequencies. The relatedness matrix based on RKC (Ritland, 1996) proposed by Yu (2006), takes into account the allele frequencies but it is very sensitive to non-positive definiteness and convergence problems. It is also a relative measure of relatedness as it considers the average relatedness in the population sample.

Three genes coding transcription factors, *MYB11*, *MYBCC* and MYC_B were found to be associated with TSA concentration. Three SNPs (s36,

s68, s90) were significantly associated with TSA concentration after correcting for relatedness and structure. Empirical *P*-values for these SNPs confirmed their association with TSA concentration. None of the changes caused by these SNPs are non-synonymous. SNP s36 leads to a G/C base replacement causing no amino acid substitution. Nevertheless, this sequence region is a CG and CHG context (where H = A, T or C), possibly important for epigenetic regulation by cytosine methylation (Henderson and Jacobsen, 2007). Also there is a 12 bp INDEL 239.5 bp upstream **s36** in the 5'UTR region. UTRs have been shown to play an important role on gene expression regulation. This influence may rely on different mechanisms such as the presence of upstream ORFs, secondary structures and protein or short RNAs binding sites (Morello and Breviario, 2008). SNP s68 causes an A/C base substitution in an intron region. As this is a CHH sequence region it may also play a role in methylation (Henderson and Jacobsen, 2007). It is now evident that intronic regions play important functional roles such as expression enhancement, alternative splicing and generation of intronic microRNA (Morello and Breviario, 2008). SNP **s90** causes an A/G base replacement in a sequence predicted to be part of the promoter region. According to the TSSP promoter prediction program for plant genes available on SoftBerry network server (http://www.softberry.com), **s90** is located only 4 bp away from the transcription start site (TSS).

Two additional SNPs, s65 and s89, in *MYBCC* and *MYB11* respectively, were significantly associated with TSA concentration under Model A. SNP s65 is located on an intron region and while s89 causes a mutation on an exon, it does not actually lead to an amino acid change. SNP s65 may play a regulatory role or change methylation on this region since this is a CG or CHH sequence depending on the SNP allele. SNP

s89 does not affect methylation (Henderson and Jacobsen, 2007) however 197.5 bp upstream there is an 8 bp INDEL in the 5'UTR region, only 9 bp away from the TSS. Empirical *P*-values confirmed the significance of these SNPs under the simple model.

Genes coding *UFGT* and *MRP* were not associated with TSA concentration but were associated with specific types of anthocyanins, especially Peonidin derivatives, and with phenotypes that were visually classified (PC and SPC). This indicates that these genes may be important for relative abundance of anthocyanins and less relevant for anthocyanin concentration. This is supported by the highest proportion of associated phenotypes involving relative abundance, especially in the case of *UFGT*. Visual characterisations of skin and pulp colour (PC and SPC) showed associations with a large number of genes, including genes coding transcription factors and related to the biosynthetic pathway and transport of anthocyanins.

The results showed variation on association between different phenotypes and the same SNP. Although all phenotypes are related to the colour of berries, correlation estimates between phenotypes varied substantially. For example, the correlation between TSA concentration and the two visual phenotypes (PC and SPC) were found to be around 0.5 for both ($P < 0.0001$). These two visual phenotypes were highly correlated ($r^2 = 0.83$, $P < 0.0001$) and so the association results were very similar. However, other factors could also be important. Different degrees of penetrance, phenotypic heterogeneity and environmental factors could all lead to differences in association. Also, there was variation between different SNPs for the same phenotype. The different levels of association could be due to differences in MAF for the genetic variant and the set of markers.

Statistical interactions between a high proportion of SNPs were observed for genes coding transcription factors and genes coding enzymes involved in the biosynthetic pathway of anthocyanins. SNPs within *MYB11* showed significant interactions with SNPs on genes coding *LDOX*, *CHI*, *MYBCC* and MYC_B. Also SNPs on MYC_B showed significant interactions with SNPs on *MYBCC* and *LDOX*. Biological interpretation of the statistical interactions must be performed carefully and ideally must be supported by further investigation (Cordell, 2009). The observed statistical interactions suggest that the transcription factor *MYB11* has a regulatory role over the genes involved in the metabolic pathway, *CHI* and *LDOX*. Also MYC_B is suggested to regulate *LDOX*. These results indicate that the three transcription factors *MYB11*, *MYBCC* and MYC_B functionally interact to regulate anthocyanin synthesis. This agrees with previous findings where transcription factors from Myb and Myc family were shown to regulate genes encoding anthocyanins biosynthetic enzymes in maize, petunia, *Arabidopsis* and grapevine (Baudry *et al.*, 2004; Bogs *et al.*, 2007; Cutanda-Perez *et al.*, 2009; Deluc *et al.*, 2006, 2008; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; Matus *et al.*, 2009; Quattrocchio *et al.*, 1998; Splet *et al.*, 2000; Terrier *et al.* 2009, This *et al.*, 2007). In grapevine, different Myb genes, *MYBPA1* (Bogs *et al.*, 2007; Terrier *et al.*, 2009), *MYBPA2* (Terrier *et al.*, 2009), *MYB5a* (Matus *et al.*, 2009), *MYB5b* (Deluc *et al.*, 2008) and *MYBA1* (Cutanda-Perez *et al.*, 2009), were found to activate *LDOX* promoter. Deluc *et al.* (2006, 2008) also showed the regulation of *CHI* by the transcription factors *MYB5a* and *MYB5b*. Members of Myb and Myc transcription factor families have also been previously shown to interact with each other in grapevine and in other species. Differential expression analysis and transient expression assays showed that the interaction

between the different transcription factors was essential for their ability to activate pathway genes expression (Baudry *et al.,* 2004; Bogs *et al.,* 2007; Spelt *et al.,* 2000; Goff *et al.,* 1992).

This study was performed on a small population sample compared to human genetics studies. However, in the area of genome research in grapevine, this is one of the largest samples studied for association mapping. Power calculations are based on the assumption of strong LD between the variant and the marker. Here we have performed fine mapping with average distance between SNPs of near 300 bp which makes the study powerful. Nevertheless, multiple testing is still an issue. Bonferroni correction is highly conservative as all the markers within the genes are in strong LD. Therefore, replication studies would be valuable for verifying the significance of these results. Despite the rapid increase in genomic resources for *Vitis vinifera* L., these are still limited compared to other species such as human. Association mapping is much more recent in plant studies. The availability of larger collections with genomic and phenotypic data would greatly contribute to future association studies. The International HapMap Project has been extremely useful and successful and such large scale studies will soon be available for other species including *Vitis*. The novel findings from this study and the SNPs that have been identified will be of great interest in genome-wide projects. This study has shown association between berry colour and anthocyanin content with interesting genes which need to be further investigated to better understand the genetics underlying colour. The identification of the functional variants will accelerate grapevine breeding programs.

5. References

Akey, J., Jin, L., Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* **9**, 291-300.

Alpert, K.B. and Tanksley, S.D. (1996). High-resoltuion mapping and isolation of a yeast artificial chromosome contig containing fw2.2: A major fruit weight quantitative trait locus in tomato. *Proceedings of the National Academy of Sciences of the USA* **93**, 15503-15507.

Ageorges, A., Fernandez, L., Vialet, S., Merdinoglu, D., Terrier, N. *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.

Altmüller, J., Palmer, L.J., Fischer, G., Scherb, H., Wjst M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. *American Journal of Human Genetics* **69**, 936-950.

Balding, D.J. (2006). A tutorial on statistical methods for populataion association studies. *Nature Reviews Genetics* **7**, 781-791.

Baudry, A., Heim, M.A., Dubreucq, B., Caboche, M., Weisshaar, B. *et al.* (2004). TT2, TT8 and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant Journal* **39**, 366-380.

Bogs, J., Jaffé, F.W., Takos, A.M., Walker, A.R., Robinson, S.P. (2007). The grapevine transcription factor VvMYBPA1 regulates proanthocyanidin synthesis during fruit development. *Plant Physiology* **143**, 1347-1361.

Borovsky, Y., Oren-Shamir, M., Ovadia, R., De Jong, W., Paran, I. (2004). The A locus that controls anthocyanin accumulation in pepper encodes a MYB transcription factor homologous to *Anthocyanin2* of

Petunia. *Theoretical and Applied Genetics* **109**, 23-29.

Brouillard, R. (1982). Chemical structure of anthocyanins. In *Anthocyanins as food colors.* Markakis, P., Ed. Academic Press: New York, pp 1-38.

Cardon, L.R. and Bell, J.I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics* **2**, 91-99.

Chakraborty, R. and Jin, L. (1993). Determination of relatedness between individuals using DNA-fingerprinting. *Human Biology* **65**, 875-895.

Chanda, P., Zhang, A., Brazeau, D., Sucheston, L., Freudenheim, J.L., *et al.* (2007). Information-theoretic metrics for visualizing gene-environment interactions. *American Journal Human Genetics* **81**, 939-963.

Cordell, H. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**, 392-404.

Corder, E.H., Saunder, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., *et al.* (1993). Gene dose of apolopoprotein-E type-4 allele and the risk of alzheimers-disease in late-onset families. *Science* **261**, 921-923.

Cortell, J.M. and Kennedy, J.A. (2006). Effect of shading on accumulation of flavonoid compounds in (*Vitis vinifera* L.) pinot noir fruit and extraction in a model system. *Journal of Agricultural and Food Chemistry* **54**, 8510-8520.

Culverhouse, R., Suarez, B.K., Lin, J., Eich, T.A perspective on epistasis: limits of models displaying no main effect. *American Journal of Human Genetics* **70**, 461-471.

Cutanda-Perez, M., Ageorges, A., Gomez, C., Vialet, S., Terrier, N., *et al.* (2009). Ectopic expression of *VlmybA1* in grapevine activates a narrow set of genes involved in anthocyanin synthesis and transport.

Plant Molecular Biology **69**, 633-648.

Deluc, L., Barrieu, F., Marchive, C., Lauvergeat, V., Decendit, A., *et al.* (2006). Characterisation of a grapevine R2R3-MYB transcription factor that regulates the Phenylpropanoid pathway. *Plant Physiology* **140**, 499-511.

Deluc, L., Bogs, J., Walker, A.R., Ferrier, T., Decendit, A., *et al.* (2008). The transcription factor VvMYB5b contributes to the regulation of anthocyanin and proanthocyanidin biosynthesis in developing grape berries. *Plant Physiology* **147**, 2041-2053.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997-1004.

Doligez, A., Bouquet, A., Danglot, Y., Lahogue, F., Riaz, S., *et al.* (2002). Genetic mapping of grapevine (Vitis vinifera L.) applied to the detection of QTLs for seedlessness and berry weight. *Theoretical and Applied Genetics* **105,** 780-795.

Doligez, A., Adam-Blondon, A. F., Cipriani, G., Di Gaspero, G., Laucou, V., Merdinoglu, D., Meredith, C. P., Riaz, S., Roux, C. and This, P. (2006). An integrated SSR map of grapevine based on five mapping populations. *Theoretical and Applied Genetics* **113**, 369-382.

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., *et al.* (2008). Exploration of gene-gene interaction effects using entropy-based methods. *European Journal of Human Genetics* **16**, 877-905.

Downey, M.O. (2004). Biosynthesis of flavonoids in grapevines (*Vitis vinifera* L.). Thesis submitted for Doctor of Philosophy. University of Adelaide, Australia.

Downey, M.O., Dokoozlian, N.K., Krstic, M.P. (2006). Cultural practice and environmental impacts on the flavonoid composition of grapes and wine: A review of recent research. *American Journal of Enology*

and Viticulture **57**, 257-268.

Eder, A. (2000). Pigments. In *Food analysis by HPLC.* Noller, M.L.N., Ed. Marcel Dekker: New York, pp 845-880.

Evanno, G., Regnaut, S., Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Falush, D., Stephens, M., Pritchard, J.K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* **7**, 574-578.

Fischer, B.M., Salakhutdinov, I., Akkurt, M., Eibach, R., Edwards, K J., *et al.* (2004). Quantitative trait locus analysis of fungal disease resistance factors on a molecular map of grapevine. *Theoretical and Applied Genetics* **108**, 501-515.

Fournier-Level, A., Le Cunff, L., Gomez, C., Doligez, A., Ageorges, A., *et al.* (2009). Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. ssp. *sativa*) berry: a quantitative trait locus to quantitative trait nucleotide integrated study. *Genetics* **183**, 1127-1139.

Garcez, R.M. (1997). Caracterização de cultivares de videira através dos perfis antociânicos (HPLC). Undergraduate Thesis, Escola Superior Agrária de Santarém.

Giusti, M.M. and Wrolstad, R.E. (2003). Acylated anthocyanins from edible sources and their applications in food systems. *Biochemical Engineering Journal* **14**, 217-225.

Goff, S.A., Cone, K.C., Chandler, V.L. (1992). Functional analysis of the transcriptional activator encoded by the maize B gene: evidence for a direct functional interaction between two classes of regulatory proteins. *Genes & Development* **6**, 864–875.

Gollop, R., Farhi, S., Perl, A. (2001). Regulation of the leucoanthocyanidin dioxygenase gene expression in *Vitis vinifera*. *Plant Science* **161**, 579-588.

Gollop, R., Even, S., Colova-Tsolova, V., Perl, A. (2002). Expression of the grape dihydroflavonol reductase gene and analysis of its promoter region. *Journal of Experimental Botany* **53**, 1397-1409.

Goodman, C.D., Casati, P., Walbot, V. (2004). A multidrug resistance-associated protein involved in anthocyanin transport in *Zea mays. The Plant Cell* **16**, 1812-1826.

Harborne, J.B. and Harborne, A.J. (1998). *Phytochemical methods: a guide to modern techniques of plant analysis.* Chapman & Hall: London, pp 302.

Hardy, O.J. and Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**, 618-620.

Henderson, I.R. and Jacobsen, S.E. (2007). Epigenetic inheritance in plants. *Nature* **447**, 418-424.

Holton, T.A. and Cornish, E.C. (1995). Genetics and Biochemistry of Anthocyanin Biosynthesis. *The Plant Cell* **7**, 1071-1083.

Jeong, S. T., Goto-Yamamoto, N, Kobayashi, S., Esaka, A. (2004). Effects of plant hormones and shading on the accumulation of anthocyanins and the expression of anthocyanin biosynthetic genes in grape berry skins. *Plant Science* **167**, 247-252.

Jeong, S. T., Goto-Yamamoto, N., Hashizume, K., Esaka, M. (2006). Expression of the flavonoid 3'-hydroxylase and flavonoid 3',5'-hydroxylase genes and flavonoid composition in grape (*Vitis vinifera*). *Plant Science* **170**, 61-69.

Kang, H.M., Zautlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., *et al.*

(2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-1723.

Kobayashi, S., Ishimaru, M., Ding, C. K., Yakushiji, H., Goto, N. (2001). Comparison of UDP-glucose:flavonoid 3-*O*-glucosyltransferase (UFGT) gene sequences between white grapes (*Vitis vinifera*) and their sports with red skin. *Plant Science* **160**, 543-550.

Kobayashi, S., Ishimaru, M., Hiraoka, K., Honda, C. (2002). *Myb*-related genes of the Kyoho grape (*Vitis labruscana*) regulate anthocyanin biosyntesis. *Planta* **215**, 924-933.

Kobayashi, S., Goto-Yamamoto, N., Hirochika, H. (2004). Retrotransposon-Induced Mutations in Grape Skin Color. *Science* **304**, 982.

Lijavetzky, D., Ruiz-García, L., Cabezas, J.A., De Andrés, M.T., Bravo, G., *et al.* (2006). Molecular genetics of berry colour variation in table grape. *Molecular Genetics and Genomics* **276**, 427-435.

Loiselle, B.A., Sork, V.L., Nason, J., Graham, C. (1995) Spatial genetic structure of a tropical understory shrub, *Psychotri officinalis* (*Rubiaceae*). *American Journal of Botany* **82**, 1420-1425.

Long, A.D. and Langley, C.H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* **9**, 720-731.

Malosetti, M., Van Der Linden, C.G., Vosman, B., Van Eeuwijk, F.A. (2007). A mixed-model approach to association mapping using pedogree information with an illustration of resistance to *Phytophora infestans* in potato. *Genetics* **175**, 879-889.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209-220.

Marchini, J., Donnelly, P., Cardon, L.R. (2005). Genome-wide strategies

for detecting multiple loci that influence comple diseases. *Nature Genetics* **37**, 413-417.

Martin, C. and Gerats, T. (1993). Control of pigment biosynthesis genes during petal development. *Plant Cell* **5**, 1253-1264.

Matus, J.T., Loyola, R., Vega, A., Peña-Neira, A., Bordeu, E., *et al.* (2009). Post-veraison sunlight exposure induces MYB-mediated transcriptional regulation of anthocyanin and flavonol synthesis in berry skins of *Vitis vinifers. Journal of Experimental Botany* **60**, 853-867.

Matus, J.T., Poupin, M.J., Cañón, P., Bourdeu, E., Alcalde, J.A., *et al.* (2010). Isolation of WDR and bHLH genes related to flavonoid synthesis in grapevine (*Vitis vinifera* L.). *Plant Molecular Biology* **72**, 607-620.

Moore, J.H., Gilbert, J.C., Tsai, C.-T., Chiang, F.-T., Holden, T., *et al.* (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology* **241**, 252-261.

Mueller, L.A., Goodman, C.D., Silady, R.A. and Walbot, V. (2000). *AN9*, a petunia Glutathione S-transferase required for anthocyanin sequestration, is a flavonoid-binding protein. *Plant Physiology* **123**, 1561-1570.

Organisation Internationale de la Vigne et du Vin (OIV). (2009). *Recueil des méthodes internationales d'analyse des vins et des moûts.* OIV: Paris.

Payne, C.T., Zhang, F., Lloyd, A.M. (2000). *GL3* encodes a bHLH protein that regulates trichome development in *Arabidopsis* through intercation with *GL1* and *TTG1*. *Genetics* **156**, 1349-1362.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., *et al.* (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909.

Pritchard, J.K. and Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **6**, 220-228.

Pritchard, J.K., Stephens, M., Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Pritchard, J.K., Wen, X., Falush, D. (2009). *Documentation for structure software: Version 2.3.* University of Chicago: Chicago.

Quattrocchio, F., Wing, J.F., Van der Woude, K., Mol, J.N.M., Koes, R. (1998). Analysis of bHLH and MYB domain proteins: species-specific regulatory differences are caused by divergent evolution of target anthocyanin genes. *Plant Journal* **13**: 475-488.

Ramsay, N.A., Walker, A.R., Mooney, M. (2003). Two basic-helix-loop-helix genes (*MYC-146* and *GL3*) from *Arabidopsis* can activate anthocyanin biosynthesis in a white-flowered *Matthiola incana* mutant. *Plant Molecular Biology* **52**, 679-688.

Ribéreau-Gayon, P. (1982). The anthocyanins of grapes and wines. In *Anthocyanins as food colors.* Markakis, P., Ed. Academic Press: New York, pp 209-242.

Risch N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517.

Risch, N. (2000). Searching for genetic determinants in the new millenium. *Nature* **405**, 847-856.

Ritland, K. (1996). A marker-based method for inferences about

quantitative inheritance in natural populations. *Evolution* **50**, 1062-1073.

Robbins, M.P., Paolocci, F., Hughes, J.W., Turchetti, V., Allison, G., *et al*. (2003). *Sn*, a maize bHLH gene, modulates anthocyanin and condensed tannin pathways in *Lotus corniculatus*. *Journal of Experimental Botany* **54**, 239-248.

Rozen, S. and Skaletsky, H.J. (2000). Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Krawetz, S. and Misener, S., Ed. Humana Press Totowa: New Jersey, pp 365-386.

Sainz, M.B., Grotewold, E., Chandler, V. L. (1997). Evidence for direct activation of an anthocyanin promoter by the maize C1 protein and comparison of DNA binding by related Myb domain proteins. *Plant Cell* **9**, 611-625.

Salmaso, M., Malacarne, G., Troggio, M., Faes, G., Stefanini, M., Grando, S. and Velasco, R. (2008). A grapevine (*Vitis vinifera* L.) genetic map integrating the position of 139 expressed genes. *Theoretical and Applied Genetics* **116**, 1129-1143.

Spelt, C., Quattrocchio, F., Mol, J.N.M., Koes, R. (2000). *anthocyanin1* of petunia encodes a basic helix-loop-helix protein that directly activates transcription of structural genes. *The Plant Cell* **12**, 1619-1631.

Stuber, C.W., Polacco, M., Lynn, M. (1999). Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield. *Crop Science* **39**, 1571-1583.

Terrier, N., Torregrossa, L., Ageorges, A., Vialet, S., Verriès, C., *et al.* (2009). Ectopic expression of *VvMybPA2* promotes proanthocyanidin biosynthesis in grapevine and suggest additional targets in the

pathway. *Plant Physiology* **149**, 1028-1041.

This, P., Lacombe, T., Cadle-Davidson, M. and Owens, C. (2007). Wine grape (*Vitis vinifera* L.) color associates with allelic variation in the domestication gene *VvmybA1*. *Theoretical and Applied Genetics* **114**, 723-730.

Troggio, M., Malacarne, G., Coppola, G., Segala, C., Cartwright, D., *et al.* (2007). A dense sigle-nucleotide polymorphism-based genetic linkage map of grapevine (*Vitis vinifera* L.) anchoring pinot noir bacterial artificial chromosome contigs. *Genetics* **176**, 2637-2650.

Walker, A.R., Lee, E., Bogs, J., McDavid, D.A.J., Thomas, M.R., *et al.* (2007). White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant Journal* **49**, 772-785.

Yang, Y., Houle, A.M., Letendre, J., Richter, A. (2008). *RET* Gly691Ser mutation is associated with primary vesicoureteral reflux in the French-Canadian population from Quebec. *Human Mutation* **29**, 695-702.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Masanori, Y., *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208.

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., *et al.* (2007). An *Arabidopsis* example of association mapping in structured samples. *PLoS Genetics* **3**, 71-82.

Zhu, C.G.M., Buckler, E.S., Yu, J. (2008). Status and prospects of association mapping in plants. *The Plant Genome* **1**, 5-20.

Zielenski, J. and Tsui, L. (1995). Cystic fibrosis - genotypic and phenotypic variations. *Annual Review of Genetics* **29,** 777–807.

6. Acknowledgements

I must thank Nikolas Maniatis (UCL) for his assistance with association mapping experimental design and data analysis and Winston Lau (UCL) for his assistance with data handling. I would also like to acknowledge Isabel Spranger, Conceição Leandro and Baoshan Sun (INIA - Dois Portos) for assistance on anthocyanin extraction and measurement; Margarida Santos for ELISA tests (INIA - Oeiras) and Flávia Moreira (IASMA) for performing SSR genotyping.

CHAPTER V

GENERAL DISCUSSION

1. General Discussion

This thesis contributes to a better understanding of the genetic variation underlying colour and anthocyanin content of grape berries. This trait is of great importance in grape and wine industry. Anthocyanin accumulation in berry skin determines their colour and organoleptic characteristics of grapes and wine, and has an important beneficial effect on human health. Although great attention has been given to the study of anthocyanins in grapevine there is still a lack of information concerning quantitative variation among coloured cultivars and clones.

In Chapter II, variation underlying total anthocyanin concentration on berry skin between clones was investigated. Clonal selection is a major process of quality improvement in grapevine. The use of molecular markers for clonal discrimination has so far yielded contradictory results. The reasons underlying phenotypic variation among clones have been explained with chimerical state; however, some cases remain to be explained.

In this study, clones of Negra Mole and Aragonez cultivars showed identical DNA sequences to the remaining clones of the same cultivar. A small number of SNPs was identified between Aragonez and Negra Mole. Variation on gene expression level was studied between clones of Aragonez showing contrasting concentrations of total anthocyanins in berry skin. The results suggest that subtle differences in the expression of several genes may influence these colour variations, which is typical of quantitative traits. These genes included genes encoding flavonoid metabolism enzymes and mostly transcription factors. This group of transcription factors comprised members of the Myb, Myc, zinc finger, homeodomain and WRKY families. Other genes from the *Myb* and *Myc* families have been shown to be important in anthocyanin regulation in

grapevine and other plant species (Bogs *et al.*, 2007; Deluc *et al.*, 2006, 2008; Dooner, 1991; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2006; Matus *et al.*, 2009, 2010; Terrier *et al.*, 2009; This *et al.*, 2007). *Zinc finger* and *WRKY* families have also been observed to influence proanthocyanidin synthesis (Sagasser *et al.*, 2002; Johnson *et al.*, 2002) and Kubo *et al.* (1999) identified a homeobox gene to be involved in anthocyanin accumulation in *Arabidopsis*.

The importance of some of the differentially expressed genes has been supported by the results of association mapping in Chapter IV. However, other techniques of differential gene expression with higher sensitivity to small expression differences such as RT-PCR and Northern blot could be used to assess the importance of the genes detected in this study. Moreover, further work on DNA variants in other regions of the genome may identify other variants.

In Chapter III, it was aimed to compare grapevine cultivar characterisation with measures of anthocyanin concentration, relative abundance and visual assessment of colour. The use of these compounds on cultivar characterisation has been successfully applied (Ryan, 2003; Careno, 1997; Arozarena, 2002). However, association mapping has focused only on categorical visual characterisation or been limited to total anthocyanin concentration, disregarding specific types of anthocyanins proportion and concentration. The results here presented showed that different measures of concentration are highly related with each other. Relative abundance of anthocyanins and concentration are uncorrelated in many cases. Characterisation of cultivars using specific types of anthocyanins concentration and relative abundance separated cultivars effectively according to acylation pattern and anthocyanidin types. However, using concentration data in multivariate analysis

separated cultivars by anthocyanidin methylation while using relative abundance data separated them according to hydroxylation of the B ring. These observations suggest that for investigating the genetic basis of the colour trait, the two measures of concentration and relative abundance of anthocyanins should be used.

It was also aimed to compare the characterisation of cultivars using SSR markers and anthocyanins. The joint utilisation of characterisations based on phenotypic and molecular markers data has been advised as an effective germplasm characterisation approach since the latter would exclude environmental effects. Anthocyanin concentration and molecular variance of SSR markers were found to be unrelated. However, relative abundance of anthocyanins was observed to be weakly correlated with SSR molecular variance. In agreement with previous works in maize (Rebourg, 2003; Hartings, 2008), these results support a higher accuracy obtained with molecular markers than with anthocyanin data. This difference is most likely explained by environmental effects affecting the phenotype observations.

Finally, the impact of other variables such as viral infections and maturity state on total skin anthocyanin concentration was assessed. Virus infections and maturation parameters were found not to affect significantly the concentration of total anthocyanins in grape berries skin on the sample of cultivars studied. This information is important to consider on association mapping, as correction for these variables is not necessary in association tests with this sample.

In future work, an evaluation of the effect of anthocyanin types on other organoleptic properties besides colour would add valuable information to quality improvement of grapes. Dufour and Sauvaitre (2000) have already shown volatile compounds to interact with malvidin-

3,5-*O*-diglucoside. Similar studies may be performed for other anthocyanins.

The associations between candidate genes and grape colour using SNP markers and a wide range of colour related phenotypes was the aim explored in Chapter IV. Genetic studies on grape colour have focused on qualitative variation of colour or had a simplistic approach looking only at total anthocyanin concentration. For this reason there is a great gap in knowledge about the genetic control of such an important phenotype. The importance of population structure and relatedness in grapevine was also investigated in detail. The latter has received great attention in plant association mapping studies because sampling from germplasm collections often gathers samples with high levels of relatedness.

In this study, structure and relatedness effects were examined. It was found that for more than 90 % of the SNPs, the simplest model was significantly different to the full model that considered both effects. However, the analysis of total skin anthocyanin concentration showed associations with polymorphisms in the same gene set with both the simple and the full model. The associations with the simple model were examined both empirically and nominally. So the question that arises is whether the use of a less parsimonious model could lead to type II errors.

In Chapter IV, two types of relatedness measures were examined. The method of relatedness inference has been widely debated. Zhao (2007) showed that a matrix based on the proportion of shared alleles between individuals effectively corrects for cryptic relatedness. Kang (2008) showed that this matrix guarantees positive semidefiniteness and convergence when there is no missing data. Nevertheless, this matrix does not take into account allele frequencies. Yu *et al.* (2006) proposed

the use of a matrix based on Ritland's kinship coefficient. This matrix takes into account allele frequencies; however, it is a relative measure of relatedness as it considers the average relatedness in the population sample. This matrix was found to be sensitive to problems with convergence and non-positive definiteness. Our results showed that there are no differences in the association models using the matrix based on the proportion of alleles shared or based on Ritland's kinship coefficient.

Polymorphisms in three genes coding transcription factors of the Myb and Myc families (*MYB11*, *MYBCC* and MYC_B) were found to be associated with concentration of total anthocyanins in berry skin. None of the three SNPs showing these associations were non-synonymous. However, they are likely to have regulatory roles as one is located in the promoter region and the other two are in a sequence context possibly important for epigenetic regulation by cytosine methylation. Another two genes, *UFGT* and *MRP* showed polymorphisms associated with specific types of anthocyanins, especially peonidin derivatives and with phenotypes that were visual classifications of colour. This suggests that these genes are important for relative abundance of anthocyanins and less influent on concentration of anthocyanins. These genes are involved in the biosynthetic pathway and in the transport of anthocyanins to the vacuole, respectively. Pulp colour and skin and pulp colour jointly were associated with a large number of genes (11 and 9, respectively) including genes involved in the anthocyanin biosynthetic pathway, in the transport and coding transcription factors.

The association results showed variation on association between the same SNP and different phenotypes. Despite all phenotypes being related to the colour of berries, results on Chapter III showed correlation estimates between phenotypes to vary substantially. However, other

factors could also be important. Different degrees of penetrance, genetic heterogeneity and environmental factors could all lead to differences in association. Variation between different SNPs for the same phenotype was also observed. The different levels of association could be due to differences in minor allele frequency for the genetic variant and the set of markers.

Statistical interactions between a high proportion of SNPs were observed for genes coding transcription factors (*MYBCC*, *MYC$_B$* and *MYB11*) and genes involved in the biosynthetic pathway of anthocyanins (*CHI* and *LDOX*). SNPs within the three transcription factors showed significant interactions with each other. *MYB11* showed significant interactions with two genes involved in the biosynthetic pathway (*CHI* and *LDOX*) while *MYC$_B$* was observed to interact with *CHI* only. These results suggest that these three transcription factors functionally interact to regulate anthocyanin biosynthesis. This agrees with previous findings where transcription factors of the Myb and Myc families were found to regulate genes encoding anthocyanin biosynthetic enzymes in maize, petunia, *Arabidopsis* and grapevine (Bogs *et al.,* 2007; Cutanda-Perez *et al.*, 2009; Deluc *et al.,* 2006, 2008; Dooner, 1991; Kobayashi *et al.*, 2002, 2004; Lijavetzky *et al.*, 2007; Matus *et al.*, 2009; Terrier *et al.*, 2009; This *et al.*, 2007).

In grapevine, Deluc *et al.* (2006; 2008) showed *MYB5a* and *MYB5b* to regulate *CHI* expression. Also *LDOX* promoter was found to be activated by *MYBPA1* (Bogs *et al.*, 2007; Terrier *et al.*, 2009), *MYBPA2* (Terrier *et al.,* 2009), *MYB5a* (Matus *et al.*, 2009), *MYB5b* (Deluc *et al.*, 2008) and *MYBA1* (Cutanda-Perez *et al.*, 2009). Differential expression analyses and transient expression assays have shown the interaction between transcription factors to be essential for their ability to activate the

expression of genes involved on the biosynthetic pathway (Baudry *et al.*, 2004; Bogs *et al.*, 2007; Goff *et al.*, 1992; Spelt *et al.*, 2000).

The transcription factors selected as candidate genes as a consequence of the transcription analysis performed on Chapter II showed polymorphisms significantly associated with concentration of total skin anthocyanins and involved in significant interactions. This fact supports the importance of the microarray analysis (Chapter II) and the need to obtain clearer means of assessing differential expression.

The sample used for association mapping in this study was relatively small, especially when compared to studies performed in humans. Nevertheless, this is one of the largest samples used for association mapping in *Vitis vinifera* L.. Besides sample size, power calculations are based on the assumption of strong LD between the marker and the causal variant. A fine mapping approach, with an average distance between SNPs of near 300 bp, increased the power achieved in this study. However, multiple testing is a concern. Bonferroni correction is highly conservative, since all the markers are in strong LD. Consequently, the significance of the results obtained must be verified by replication and functional studies. Although recently genomic resources for grapevine have increased rapidly, these are still limited when compared to other species like humans. The availability of larger germplasm collections with genomic and phenotypic data would greatly contribute to successful association studies in the future.

Overall, this study contributed to the knowledge of grape colour trait and of the genetic mechanisms affecting it. On the clone level the results provide evidence that the phenotypic differences observed are influenced by small differences in gene expression, especially concerning genes coding transcription factors. It was also shown that it is advantageous to

study the colour phenotype on a broader perspective considering global measures as well as more detailed information on concentration and proportion of specific anthocyanin types. Association mapping provided information on three transcription factor genes (*MYBCC, MYCB* and *MYB11*) with major importance in controlling concentration of total skin anthocyanins. Another two genes, *UFGT* and *MRP* were shown to be relevant on specific anthocyanin content. Pulp colour and pulp and skin colour jointly were found to be influenced by a wide group of genes coding transcription factors, and enzymes for the transport and biosynthesis of anthocyanins. Finally, functional interactions between the transcription factors and the pathway genes were suggested by statistical interactions.

This thesis deepened the knowledge of grapevine genetics gathering data on sequence polymorphisms, phenotypes and genes of great interest for this crop. Further studies may build upon the results here presented by exploring DNA variation among clones at the genome-wide level, performing association analysis using gene expression as a phenotype or exploring functionally the polymorphisms here identified as especially interesting.

Identifying genes that control anthocyanin content is extremely valuable for grapevine germplasm management. This is a trait of great importance in grape and wine industry as it affects colour and organoleptic characteristics and has an important beneficial effect on human health. After replication and confirmation by functional assays, the findings here presented will provide genetic markers allowing for higher speed in germplasm assessment and breeding programs. This is an important step in meeting the future needs in grapevine cultivation and grape and wine industries.

2. References

Arozarena, I., Ayestaran, B., Cantalejo, M.J., Navarro, M., Vera, M., *et al.* (2002). Anthocyanin composition of Tempranillo, Garnacha and Cabernet Sauvignon grapes from high and low quality vineyards over two years. *European Food Research and Technology* **214**, 303-309.

Baudry, A., Heim, M.A., Dubreucq, B., Caboche, M., *et al.* (2004). *TT2*, *TT8* and *TTG1* synergistically specify the expression of *BANYULS* and proanthocyanidin biosynthesis in *Arabidopsis thaliana. Plant Journal* **39**, 366-380.

Bogs, J., Jaffé, F.W., Takos, A.M., Walker, A.R., Robinson, S.P. (2007). The grapevine transcription factor VvMYBPA1 regulates proanthocyanidin synthesis during fruit development. *Plant physiology* **143**, 1347-1361.

Cutanda-Perez, M., Ageorges, A., Gomez, C., Vialet, S., Terrier, N., *et al.* (2009). Ectopic expression of *VlmybA1* in grapevine activates a narrow set of genes involved in anthocyanin synthesis and transport. *Plant Molecular Biology* **69**, 633-648.

Deluc, L., Barrieu, F., Marchive, C., Lauvergeat, V., Decendit, A., *et al.* (2006). Characterization of a grapevine R2R3-MYB transcription factor that regulates the Phenylpropanoid pathway. *Plant Physiology* **140**, 499-511.

Deluc, L., Bogs, J., Walker, A.R., Ferrier, T., Decendit, A., *et al.* (2008). The transcription factor VvMYB5b contributes to the regulation of anthocyanin and proanthocyanidin biosynthesis in developing grape berries. *Plant Physiology* **147**, 2041-2053.

Dooner, H.K. and Robbins T.P. (1991). Genetic and developmental control of anthocyanin biosynthesis. *Annual Review of Genetics* **25**, 173-179.

Goff, S.A., Cone, K. C., Chandler, V. L. (1992). Functional analysis of the transcriptional activator encoded by the maize *B* gene: evidence for a direct functional interaction between two classes of regulatory proteins. *Genes & Development* **6**, 864–875.

Hartings, H., Berardo, N. Mazzinelli, G.F., Valoti, P., Verderio, A., *et al.* (2008). Assessment of genetic diversity and relationships among maize (*Zea mays* L.) Italian landraces by morphological traits and AFLP profiling. *Theoretical and Applied Genetics* **117**, 831-842.

Johnson, C.S., Kolevski, B., Smyth, D.R. (2002). *TRANSPARENT TESTA GLABRA2*, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. *The Plant Cell* **14**, 1359-1375.

Kang, H.M., Zautlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., *et al.* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-1723.

Kobayashi, S., Goto-Yamamoto, N., Hirochika, H. (2004). Retrotransposon-Induced Mutations in Grape Skin Color. *Science* **304**, 982.

Kobayashi, S., Ishimaru, M., Hiraoka, K., Honda, C. (2002). *Myb*-related genes of the Kyoho grape (*Vitis labruscana*) regulate anthocyanin biosyntesis. *Planta* **215**, 924-933.

Kubo, H., Peeters, A.J.M., Aarts, M.G.M., Pereira, A., Koornneef, M. (1999). *ANTHOCYANINLESS2*, a homeobox gene affecting anthocyanin distribution and root development in Arabidopsis. *The Plant Cell* **11**, 1217-1226.

Lijavetzky, D., Cabezas, J.A., Ibáñez, A., Rodríguez, V., *et al.* (2007). High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* **8**, 424-435.

Matus, J.T., Loyola, R., Vega, A., Peña-Neira, A., Bordeu, E., *et al.* (2009). Post-veraison sunlight exposure induces MYB-mediated transcriptional regulation of anthocyanin and flavonol synthesis in berry skins of *Vitis vinifera*. *Journal of Experimental Botany* **60**, 853-867.

Matus, J.T., Poupin, M.J., Cañón, P., Bourdeu, E., Alcalde, J.A., *et al.* (2010). Isolation of WDR and bHLH genes related to flavonoid synthesis in grapevine (*Vitis vinifera* L.). *Plant Molecular Biology* **72**, 607-620.

Rebourg, C., Chastanet, M., Gouesnard, B., Welcker, C., Dubreil, P., *et al.* (2003). Maize introduction into Europe: the history reviewed in the light of molecular data. *Theoretical and Applied Genetics* **106**, 895-903.

Ritland, K. (1996). A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* **50**, 1062-1073.

Ryan, J.M. and Revilla, E. (2003). Anthocyanin composition of Cabernet Sauvignon and Tempranillo grapes at different stages of ripening. *Journal of Agriculture and Food Chemistry* **51**, 3372-3378.

Sagasser, M., Lu, G., Hahlbrock, K., Weisshaar, B. (2002). *A. thaliana* TRANSPARENT TESTA 1 is involved in seed coat development and defines the WIP subfamily of plant zinc finger proteins. *Genes and Development* **16**, 138-149.

Spelt, C., Quattrocchio, F., Mol, J.N.M., Koes, R. (2000). *anthocyanin1* of petunia encodes a basic helix-loop-helix protein that directly activates transcription of structural genes. *The Plant Cell* **12**, 1619-1631.

Terrier, N., Torregrossa, L., Ageorges, A., Vialet, S., Verriès, C., *et al.* (2009). Ectopic expression of *VvMybPA2* promotes proanthocyanidin biosynthesis in grapevine and suggest additional targets in the pathway. *Plant Physiology* **149**, 1028-1041.

This, P., Lacombe, T., Cadle-Davidson, M., Owens, C. (2007). Wine grape (*Vitis vinifera* L.) color associates with allelic variation in the domestication gene *VvmybA1*. *Theoretical and Applied Genetics* **114**, 723-730.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Masanori, Y., *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208.

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., *et al.* (2007). An *Arabidopsis* example of association mapping in structured samples. *PLoS Genetics* **3**, 71-82.

APPENDICES

Appendix 1 Sample of 90 clones used to study SNPs on Negra Mole and Aragonez cultivars.

Cultivar	Clone	Cultivar	Clone	Cultivar	Clone
Aragonez	4001	Aragonez	3801	Negra Mole	1210
Aragonez	3507	Aragonez	1340	Negra Mole	1211
Aragonez	129	Aragonez	8201	Negra Mole	223
Aragonez	7028	Aragonez	4110	Negra Mole	1603
Aragonez	1505	Aragonez	103	Negra Mole	710
Aragonez	1607	Aragonez	107	Negra Mole	1709
Aragonez	1513	Aragonez	8303	Negra Mole	1614
Aragonez	1257	Aragonez	3911	Negra Mole	1707
Aragonez	717	Aragonez	1704	Negra Mole	205
Aragonez	8603	Aragonez	4505	Negra Mole	1203
Aragonez	8306	Aragonez	8401	Negra Mole	310
Aragonez	513	Aragonez	314	Negra Mole	2102
Aragonez	1801	Aragonez	401	Negra Mole	1404
Aragonez	1503	Aragonez	6802	Negra Mole	703
Aragonez	1702	Aragonez	8072	Negra Mole	1621
Aragonez	3311	Aragonez	4009	Negra Mole	213
Aragonez	3915	Aragonez	6407	Negra Mole	502
Aragonez	7301	Negra Mole	1615	Negra Mole	506
Aragonez	3506	Negra Mole	214	Negra Mole	714
Aragonez	1178	Negra Mole	1006	Negra Mole	305
Aragonez	501	Negra Mole	222	Negra Mole	1213
Aragonez	1507	Negra Mole	312	Negra Mole	1617
Aragonez	1285	Negra Mole	517	Negra Mole	1607
Aragonez	418	Negra Mole	2014	Negra Mole	507
Aragonez	6303	Negra Mole	2001		
Aragonez	6309	Negra Mole	1402		
Aragonez	324	Negra Mole	307		
Aragonez	8203	Negra Mole	2017		
Aragonez	8310	Negra Mole	2015		
Aragonez	506	Negra Mole	1713		
Aragonez	3903	Negra Mole	1405		
Aragonez	9101	Negra Mole	1704		
Aragonez	4010	Negra Mole	2013		

Appendix 2 List of primer pairs for study of DNA sequence among clones.

Gene	Sequence (5′ – 3′)
CHS1_F	ATTTGCATTTTCCGACGAAG
CHS1_R	ACCCACGAGAGAATCCAGGT
CHS2_F	TGACACCCACCTGGATTCTC
CHS2_R	TGTGGTGCCCTTTCCTTC
CHI_F	GGTCGAGAACGTCCTATTTCC
CHI_R	CTTCCCATCTCTCCTTCAACC
LDOX_F	CAAGCTTGCCAACAATGCTA
LDOX_R	TAGAGCCTCCTGGGTCTTCC
DFR1_F	CACAAAGTGAAACCGTGTGC
DFR1_R	GCAAGATCTGCCTTCCAGAG
DFR2_F	GCCTCCAAGCCTCATAACTG
DFR2_R	TCTTCTAGGTCTTGCCATCTACAGG

Appendix 3 List of primer pairs for validation of microarray by RT-PCR.

Gene	Sequence (5′ – 3′)
Vv4CL1_F	GCAGGATTTTACCCGATGGA
Vv4CL1_R	CTGATGCCGCTGTTGTTTCG
VvBAG6_F	CTACGGTCAACCCCATACAT
VvBAG6_R	AGAACCAGAAGGCATAGAGC
VvCCR2_F	AGTGACAAGGGGTGGATTGA
VvCCR2_R	ACAGCATGACGACTCTCTTCG
VvCHS1_F	TCTCTTCCTTCAGACCCAGTT
VvCHS1_R	GTCCCAGGGTTGATTTCCAA
VvCHS3_F	CTCGGGCTTTAGGGCTAAT
VvCHS3_R	TTTGGGCATCAAGGACTGGA
VvLOB1_F	GAAGAAGAAGAGGAAGAGGAGAC
VvLOB1_R	CAGCGGCATATTTGACGGTT
VvMybB5_F	GCAGGGTGTTGAAGCCAAAT
VvMybB5_R	AGTCCAGTCGTTCGGGTTC
VvPAL1_F	GTTCCAGCCACTGAGACAAT
VvPAL1_R	CCGAACCGAATCAAGGACTG
VvRSGTa_F	CTGACCTCGTCCACAAACTC
VvRSGTa_R	GCGGAGCTGAAGGAAAACAC
VvRSGTc_F	CGAACCTCGTCGACAAAACC
VvRSGTc_R	GCGGAGTTGAAGCAAAACGC
VvTCP5_F	TATCTGAGACCACGCTATGC
VvTCP5_R	GTTTTGCTCCTGCTGTTCGT
VvTCP9_F	CCGTCGCCATAGTAGAGTTG
VvTCP9_R	CTCTGTTGCCTCACCTTCAG
UBI_F	AGTAGATGACTGGATTGGAGGT
UBI_R	GAGTATCAAAACAAAAGCATCG

Appendix 4 List of 106 probesets with significant differential expression between four clones with high and low total skin anthocyanin concentration ($P < 0.01$).

Probeset ID	*P*-value	FC	UniProt ID	Annotation	Function
VVTU19893_at	1.3E-04	1.10	Q9SP55	Vacuolar ATP synthase subunit G related cluster	Metabolite transport facilitation
VVTU21727_s_at	1.6E-04	1.79	Q8L799	Inositol oxygenase 1 related cluster	Coenzyme and prosthetic group metabolism
VVTU25628_at	5.4E-04	1.14	Q8LD14	Hypothetical protein related cluster	Hypothetical protein
VVTU165_at	7.0E-04	1.21	Q1S9P5	cAMP response element binding (CREB) protein related cluster	Nucleic acid metabolism
VVTU16993_x_at	1.4E-03	1.16	Q84NG9	2S albumin related cluster	Storage protein
VVTU20600_at	1.8E-03	1.27	Q5XQC7	BON3 related cluster	Signal transduction
VVTU23659_at	3.0E-03	-1.11	Q1SAS4	RNase H putative related cluster	Unclassified
VVTU40312_s_at	3.6E-03	-1.15	P62577	Chloramphenicol acetyltransferase related cluster	Unclassified
VVTU15844_at	4.1E-03	1.63	O04390	Nuclear matrix constituent protein 1 related cluster	Cell structure and motility
VVTU1520_s_at	4.4E-03	-1.13	Q39242	Thioredoxin reductase 2 related cluster	Stress response
VVTU12208_at	4.5E-03	-1.29	Q6Q2Z9	Phosphoenolpyruvate carboxylase related cluster	Carbohydrate metabolism
VVTU2768_at	5.2E-03	1.09	Q94JZ8	Hypothetical protein T5E21.7 related cluster	Hypothetical protein
VVTU18932_at	1.9E-04	1.22	Q3BKH8	Hypothetical protein related cluster	Hypothetical protein
VVTU29746_s_at	3.4E-04	1.95	Q5Z8T3	Probable inositol oxygenase related cluster	Coenzyme and prosthetic group metabolism
VVTU4102_at	4.1E-04	-1.13	Q03943	Membrane-associated 30 kDa protein, chloroplast precursor related cluster	Unclassified
VVTU31304_at	1.4E-03	-1.16	Q9XGS6	Cytosolic class II low molecular weight heat shock protein related cluster	Protein metabolism and modification
VVTU15128_at	1.8E-03	1.09	Q9M2C8	Hypothetical protein T20K12.260 related cluster	Hypothetical protein
VVTU5365_at	1.8E-03	-1.16	O23063	A_IG005I10.6 protein related cluster	Nucleic acid metabolism
VVTU22795_at	2.3E-03	1.25	Q94BV6	AT5g27210 T21B4_120 related cluster	Unclassified
VVTU40706_s_at	2.4E-03	-1.10	Q6YZ89	Putative postsynaptic protein CRIPT related cluster	Unclassified
VVTU15324_at	2.6E-03	1.26	Q9FND6	Selenium-binding protein-like related cluster	Pentatricopeptide repeat
VVTU38811_at	2.7E-03	-1.09	Q8TGE7	Hypothetical protein AfA14E5.29 related cluster	Hypothetical protein
VVTU17051_at	3.2E-03	-1.27	Q6H515	Hypothetical protein OSJNBa0073A21.9 related cluster	Hypothetical protein
VVTU18661_at	3.2E-03	-1.13	Q5K4K8	Putative papain-like Cysteine proteinase related cluster	Protein metabolism and modification
VVTU34340_s_at	3.7E-03	1.41	Q9FXS1	WRKY transcription factor NtEIG-D48 related cluster	Nucleic acid metabolism

Probeset ID	*P*-value	FC	UniProt ID	Annotation	Function
VVTU4145_s_at	4.6E-03	-1.14	Q6ZL92	Hypothetical protein OJ1065_B06.22 related cluster	Hypothetical protein
VVTU14608_s_at	5.1E-03	1.21	Q9LK55	*Arabidopsis thaliana* genomic DNA, chromosome 3, P1 clone:MJG19 related cluster	Unclassified
VVTU22548_at	5.2E-03	1.16	Q9AXQ2	Mitochondrial processing peptidase beta subunit related cluster	Protein metabolism and modification
VVTU27190_x_at	5.4E-03	1.14	Q9ATF5	60S ribosomal protein L18a related cluster	Protein metabolism and modification
VVTU39365_at	5.6E-03	1.17	Q9SY57	F14N23.3 related cluster	Unclassified
VVTU27059_s_at	6.6E-03	1.47	Q9SSC3	F18B13.24 protein related cluster	Unclassified
VVTU35756_s_at	6.7E-03	1.16	Q1RZC4	Cyclin-like F-box related cluster	Protein metabolism and modification
VVTU37681_at	6.7E-03	1.20	Q5M9X0	Hypothetical protein orf160 related cluster	Hypothetical protein
VVTU4165_s_at	6.7E-03	-1.17	Q4JLS4	Hypothetical protein related cluster	Hypothetical protein
VVTU8226_at	7.3E-03	-1.16	Q32SF8	Serine threonine kinase related cluster	Signal transduction
VVTU9012_at	7.4E-03	1.37	Q9LMN0	F22L4.5 protein related cluster	Unclassified
VVTU3406_at	7.8E-03	1.93	Q45RS3	AlaT1 related cluster	Amino acid metabolism
VVTU29594_x_at	8.1E-03	-1.12	P35681	Translationally-controlled tumor protein homolog related cluster	Cell structure and motility
VVTU14082_at	8.1E-03	1.15	Q38885	Preprotein translocase secY subunit, chloroplast precursor related cluster	Protein metabolism and modification
VVTU2806_at	8.4E-03	-1.13	Q9C835	Hypothetical protein T8E24.14 related cluster	Hypothetical protein
VVTU3789_at	8.6E-03	1.17	Q8LBK0	Hypothetical protein related cluster	Hypothetical protein
VVTU37419_x_at	8.6E-03	1.35	Q1RXY0	Hypothetical protein related cluster	Hypothetical protein
VVTU38419_s_at	6.3E-04	-1.16	Q9SSV4	Inositol-3-phosphate synthase related cluster	Myo-Inositol metabolism
VVTU3597_at	6.5E-04	1.51	Q8L799	Inositol oxygenase 1 related cluster	Coenzyme and prosthetic group metabolism
VVTU570_at	8.3E-04	1.25	Q9LXU9	Hypothetical protein T24H18_70 related cluster	Hypothetical protein
VVTU38369_at	8.7E-04	1.08	Q10MC5	*O*-acetyltransferase, putative, expressed related cluster	Unclassified
VVTU2923_at	1.1E-03	1.26	Q8GTD8	Hypothetical protein 275 related cluster	Hypothetical protein
VVTU14444_at	1.6E-03	1.20	Q1SMF6	Esterase lipasethioesterase related cluster	Lipid, fatty acid, steroid metabolism
VVTU1501_at	2.0E-03	-1.13	Q6SRZ8	YABBY2-like transcription factor YAB2 related cluster	Nucleic acid metabolism
VVTU108_at	2.1E-03	1.13	Q944H6	At2g47760 F17A22.15 related cluster	Metabolism
VVTU37343_at	2.2E-03	1.41	Q9AXR6	ATP:citrate lyase related cluster	Lipid, fatty acid, steroid metabolism
VVTU15396_at	2.3E-03	-1.21	Q9LSR1	*Arabidopsis thaliana* genomic DNA, chromosome 5, BAC clone:F24B18 related cluster	Unclassified
VVTU18617_x_at	2.5E-03	1.21	Q8L7H3	Probable xyloglucan endotransglucosylase hydrolase protein 29 precursor related cluster	Cell wall metabolism

Probeset ID	*P*-value	FC	UniProt ID	Annotation	Function
VVTU26495_at	2.5E-03	-1.14	Q41251	Calmodulin-binding heat-shock protein related cluster	Signal transduction
VVTU4349_s_at	2.7E-03	1.21	O64639	Hypothetical protein At2g45590 related cluster	Signal transduction
VVTU35012_at	2.9E-03	1.15	Q9ATD1	GHMyb9 related cluster	Nucleic acid metabolism
VVTU763_at	3.9E-03	-1.10	Q5K4K7	Cysteine proteinase related cluster	Protein metabolism and modification
VVTU12956_at	4.1E-03	-1.14	Q1S084	NmrA-like related cluster	Nitrogen metabolism
VVTU5020_at	4.1E-03	-1.23	Q8W2K0	Forever young oxidoreductase related cluster	Metabolism
VVTU17_at	4.2E-03	1.13	Q9SD89	Hypothetical protein F13G24.150 related cluster	Hypothetical protein
VVTU9692_at	4.3E-03	1.21	Q9FHS6	Gb|AAC80613.1 related cluster	Unclassified
VVTU21464_at	4.6E-03	-1.11	Q28EK6	CHEK1 related cluster	Signal transduction
VVTU185_at	4.6E-03	1.41	O81366	Late embryogenesis-like protein related cluster	Stress response
VVTU4947_at	4.7E-03	1.13	Q0JQB6	Os01g0171800 protein related cluster	Unclassified
VVTU5733_s_at	4.7E-03	-1.17	Q6K4B9	Hypothetical protein OJ1509_C06.25 related cluster	Hypothetical protein
VVTU7903_at	4.7E-03	1.08	Q1SZR5	DOMON related cluster	Unclassified
VVTU2427_at	4.8E-03	-1.11	Q2R3D7	Exonuclease family protein, expressed related cluster	Nucleic acid metabolism
VVTU15326_at	4.8E-03	1.12	Q3E907	Protein At5g27550 related cluster	Unclassified
VVTU28941_x_at	5.1E-03	1.28	Q40480	C-7 protein related cluster	Stress response
VVTU15862_at	5.4E-03	-1.14	UPI00005DC273	Cluster related to UPI00005DC273, hydrolase, acting on glycosyl bonds	Carbohydrate metabolism
VVTU9070_s_at	5.9E-03	-1.13	Q41393	E24 ASN related cluster	Unclassified
VVTU32637_at	6.0E-03	-1.13	Q9M4H3	Putative metallothionein-like protein related cluster	Metabolite transport facilitation
VVTU15130_at	6.2E-03	-1.17	Q1SXZ7	Disease resistance protein, AAA ATPase related cluster	Stress response
VVTU8301_at	6.5E-03	1.11	Q6U7H9	Pectate lyase related cluster	Cell wall metabolism
VVTU16792_at	6.5E-03	1.23	O64855	Expressed protein related cluster	Unclassified
VVTU20715_at	6.6E-03	-1.10	Q9SNE9	Hypothetical protein F11C1_20 related cluster	Hypothetical protein
VVTU40262_at	6.7E-03	-1.08	Q1SD84	Integrase, catalytic region related cluster	Unclassified
VVTU3408_at	6.9E-03	-1.14	Q9ZVW2	Expressed protein related cluster	Nucleic acid metabolism
VVTU39432_s_at	7.0E-03	1.27	Q8W4H5	Hypothetical protein T18E12.18, At2g03150 related cluster	Hypothetical protein
VVTU35538_at	7.2E-03	1.17	Q1S835	2OG-Fe(II) oxygenase related cluster	Secondary metabolism

Probeset ID	*P*-value	FC	UniProt ID	Annotation	Function
VVTU6932_at	7.2E-03	-1.92	Q0PNH1	Cytochrome P450 related cluster	Cytochrome P450
VVTU9050_at	7.3E-03	1.48	O24542	Auxin-induced protein 22D related cluster	Hormone metabolism
VVTU31017_at	7.4E-03	-1.10	Q8H6Q8	CTV.20 related cluster	Unclassified
VVTU35174_at	7.5E-03	-1.57	O24548	Type IIIa membrane protein cp-wap13 related cluster	Cell wall metabolism
VVTU3160_s_at	7.7E-03	1.45	VVU68144	*Vitis vinifera* beta-1,3-glucanase mRNA, partial cds.	Cell wall metabolism
VVTU239_at	7.8E-03	-1.16	Q9LU17	Cell division protein FtsH-like related cluster	Cell growth and death
VVTU12433_s_at	7.9E-03	1.22	Q1S4P9	Isopenicillin N synthetase, KH, type 1 related cluster	Nucleic acid metabolism
VVTU37524_at	7.9E-03	-1.19	Q8VZF3	At2g47390 T8I13.23 related cluster	Unclassified
VVTU19089_at	8.0E-03	-1.14	Q10RK9	40S ribosomal protein S9, putative, expressed related cluster	Protein metabolism and modification
VVTU15630_at	8.1E-03	2.12	Q1S4X7	Berberine and berberine like, putative related cluster	Secondary metabolism
VVTU9588_x_at	8.1E-03	-1.12	Q43607	Prunin precursor related cluster	Storage protein
VVTU10737_at	8.1E-03	-1.16	Q9LJE2	Lysyl-tRNA synthetase related cluster	Protein metabolism and modification
VVTU6270_at	8.2E-03	-1.09	Q7EY72	Putative myrosinase related cluster	Stress response
VVTU4553_at	8.5E-03	-1.13	Q9C9Z4	Hypothetical protein F17O14.9 related cluster	Hypothetical protein
VVTU2631_at	8.5E-03	1.22	Q1XAN1	Sucrose responsive element binding protein related cluster	Nucleic acid metabolism
VVTU6439_at	8.5E-03	-1.14	Q1T3W0	Helix-hairpin-helix motif related cluster	Nucleic acid metabolism
VVTU20119_at	8.9E-03	-1.15	P28186	Ras-related protein ARA-3 related cluster	Signal transduction
VVTU11499_at	9.1E-03	-1.12	Q1SRF6	GRAS transcription factor related cluster	Nucleic acid metabolism
VVTU8867_at	9.4E-03	1.10	Q64MA8	Putative hASNA-I related cluster	Metabolite transport facilitation
VVTU18102_at	9.5E-03	-1.14	Q9M4H0	Putative ripening-related protein related cluster	Unclassified
VVTU38737_at	9.5E-03	-1.11	Q9SA17	F28K20.17 protein related cluster	Unclassified
VVTU24910_at	9.6E-03	-1.17	Q7XAS3	Beta-*D*-glucosidase related cluster	Carbohydrate metabolism
VVTU40436_x_at	9.6E-03	1.10	Q1T4Y6	Reverse transcriptase (RNA-dependent DNA polymerase), putative related cluster	Unclassified
VVTU16402_at	9.6E-03	1.20	UPI00005DC222	Cluster related to UPI00005DC222, TTN8 (TITAN8), ATP binding	Nucleic acid metabolism
VVTU29643_at	9.7E-03	1.10	Q1SD84	Integrase, catalytic region related cluster	Unclassified
VVTU18147_at	9.9E-03	-1.17	Q3E9C4	Protein At5g19200 related cluster	Unclassified

Appendix 5 List of 10 probesets with significant differential expression between four clones with high and low total skin anthocyanin concentration ($P < 0.05$) and fold-change higher than two.

Probeset ID	*P*-value	FC	UniProt ID	Annotation	Function
VVTU5985_s_at	2.9E-02	-2.00	O23787	Thiazole biosynthetic enzyme, chloroplast precursor related cluster	Coenzyme and prosthetic group metabolism
VVTU6962_at	3.3E-02	2.04	Q2TE76	Coat protein related cluster	Unclassified
VVTU14456_at	4.2E-02	2.08	Q3KU27	Nectarin IV related cluster	Stress response
VVTU1996_x_at	3.0E-02	2.12	Q8LA49	Globulin-like protein related cluster	Storage protein
VVTU15630_at	8.1E-03	2.12	Q1S4X7	Berberine and berberine like, putative related cluster	Alkaloid metabolism
VVTU13914_at	4.7E-02	2.14	AY043233	*Vitis vinifera* polygalacturonase mRNA, complete cds.	Cell wall metabolism
VVTU14594_s_at	4.7E-02	2.22	Q94B15	Polygalacturonase PG1 related cluster	Cell wall metabolism
VVTU3450_at	4.4E-02	2.54	Q9XEJ7	Galactinol synthase related cluster	Carbohydrate metabolism
VVTU39853_s_at	2.8E-02	3.01	Q9SQ57	Caleosin related cluster	Storage protein
VVTU3165_s_at	4.1E-02	1.09	Q3LHL3	Myb-CC type transfactor related cluster	Transcription factor

Appendix 6 Probeset with significant differential expression between four clones with high and low total skin anthocyanin concentration ($P < 0.01$) and fold-change higher than two.

Probeset ID	*P*-value	FC	UniProt ID	Annotation	Function
VVTU15630_at	8.1E-03	2.12	Q1S4X7	Berberine and berberine like, putative related cluster	Alkaloid metabolism

Appendix 7 List of 149 cultivars sampled for association mapping.

Cultivar ID	Cultivar Name	Cultivar ID	Cultivar Name
50615	Água Santa	50901	Casculho
51305	Ahmeur bou Ahmeur	53107	Castelão
41504	Alcoa	53608	Chasselas Roxo
51107	Alfrocheiro	51308	Cidadelhe
52608	Alicante Bouschet	51404	Cidreiro
52204	Alvarelhão	51102	Coarna Neagra
52604	Aragonez	50201	Complexa
50105	Aramis	50902	Concieira
53704	Aramon Noir	51304	Coração de Galo
52104	Arjunção	53803	Corinthe Noir
53705	Aspiran Noir	52004	Cornifesto
52607	Baga	53508	Cot
50203	Bandeirante	41707	Deliciosa
52101	Barca	41607	Unknown 1
52802	Bastardo	51408	Unknown 2
51607	Bastardo Tinto	51501	Unknown 3
51203	Bombalino	51601	Unknown 4
41601	Bonvedro	50904	Doçal
52807	Borraçal	50905	Doce
50106	Briosa de Oeiras	53305	Dolcetto
50208	C 19	52306	Donzelinho Tinto
50801	Cabernet Franc	51008	Engomada
53606	Cabernet Sauvignon	52904	Espadeiro
53103	Cabinda	51604	Espadeiro Mole
50301	Cabora Bassa	41502	Fepiro
53306	Cadarca	53208	Ferreira
50102	Caladoc	52708	Folgasão Roxo
52402	Camarate	53904	Gewürztraminer
41806	Campanário	50802	Gonçalo Pires
53304	Canaiolo	51204	Gorda
50603	Cardinal R	41305	Gouveio Preto
52605	Carrasquenho	50804	Grand Noir de la Calmette
52902	Carrega Burros	51602	Grangeal

Cultivar ID	Cultivar Name
51106	Grenache
51603	Grossa
50207	Imperial Rojo
50202	Joao Baga
50708	Lourela
41503	Lusitano
50608	Malandra
50601	Malvasia Fina Roxa
53205	Malvasia Preta
52002	Marufo
52908	Melhorio
52503	Mencía
51711	Molar
50702	Mondet
51804	Monvedro
41508	Moscargo
51701	Mourisco
51402	Mourisco de Semente
53303	Mourvèdre
53407	Mulata
50701	Muscat à Petits Grains Rouge
52202	Negra Mole
51303	Negro Amaro
52005	Nevoeira
50806	Padeiro
52702	Parreira Matias
52006	Patorra
52105	Pedral
51206	Petit Bouchet
51007	Pical
51606	Pilongo
53706	Pinot Noir
53505	Português Azul
52705	Preto Cardana
51803	Preto Martinho
53102	Primavera
50205	Quiebratinajas Tinta
52203	Ramisco
50303	Ribatejana
51103	Ricoca
51708	Rodo
50707	Roseira
52106	Rufete
52304	Santareno
51502	São Saul
51403	Sevilhão
50204	Sofala
51901	Sousão
53807	Teinturier
50703	Tinta Aguiar
52905	Tinta Barroca
50803	Tinta Caiada
52201	Tinta Carvalha
50103	Tinta da Guiné
50302	Tinta Ferreira
50706	Tinta Fontes
52502	Tinta Francisca
50602	Tinta Martins
50604	Tinta Mesquita
51706	Tinta Miúda
51208	Tinta Penajóia
50907	Tinta Pereira
51201	Tinta Pomar
50606	Tinta Riscadinha
51307	Tinta Tabuaço
52505	Tintem
51205	Tintinha
53307	Tinto Cão

Cultivar ID	Cultivar Name
52506	Tinto Pegões
50305	Tinto Velasco
50705	Touriga Fêmea (Dão)
52205	Touriga Franca
52206	Touriga Nacional
53004	Trincadeira
53605	Trollinger
50903	Uva Moranga
53206	Valbom
51608	Valdosa
50808	Varejoa
51513	Verdelho Roxo
51806	Verdelho Tinto
51802	Vinhão
50908	Vinhateira

Appendix 8 Summary statistics for SSR markers calculated with PowerMarker ver.3.0 (http://www. powermarker.net/**).**

Marker	Major Allele Frequency	Genotype No.	Sample Size	No. of obs.	Allele No.	Gene Diversity	Heterozygosity	PIC
VVMD5	0.2517	32	149	149	12	0.8314	0.8523	0.8105
VVMD7	0.4060	35	149	149	12	0.7769	0.7987	0.7553
VVMD25	0.2852	22	149	149	10	0.7750	0.7718	0.7394
VVMD28	0.2013	50	149	149	14	0.8785	0.8926	0.8666
VVMD32	0.2685	29	149	149	13	0.8033	0.8523	0.7749
VVS2	0.2315	37	149	149	13	0.8295	0.8054	0.8074
VrZAG62	0.4698	24	149	149	9	0.7258	0.7315	0.7002
VrZAG79	0.3345	36	149	148	11	0.8105	0.6892	0.7889
VMC1B11	0.2383	30	149	149	10	0.8374	0.8926	0.8173
VMC4F3_1	0.2114	44	149	149	13	0.8541	0.8658	0.8375
VMC4F8	0.3356	25	149	149	9	0.7637	0.7785	0.7272
VVIB01	0.4291	11	149	148	6	0.6847	0.7703	0.6281
VVIH54	0.3826	14	149	149	7	0.6868	0.7181	0.6244
VVIN16	0.4730	13	149	148	6	0.6316	0.6757	0.5621
VVIN73	0.8490	9	149	149	6	0.2709	0.2550	0.2589
VVIP31	0.2081	41	149	149	12	0.8440	0.8658	0.8254
VVIQ52	0.5777	10	149	148	5	0.5811	0.5878	0.5212
VVIV37	0.4060	31	149	149	10	0.7789	0.7852	0.7578
VVMD21	0.5777	18	149	148	8	0.6178	0.6014	0.5841
VVMD24	0.5336	19	149	149	7	0.6607	0.7181	0.6281
Mean	0.3835	26.5	149	148.75	9.65	0.7321	0.7454	0.7008

PIC stands for polymorphism information content.

Appendix 9 Correlation matrix between different concentration measures of anthocyanins.

Please see the attached CD, file "Appendix 9.xls".

Appendix 10 Correlation matrix between visual assessment of berry colour, relative abundance and concentration (mg/kg) of anthocyanins.

Please see the attached CD, file "Appendix 10.xls".

Appendix 11 Graphical representation of cultivars relative abundance of anthocyanins (%).

Please see the attached CD, file "Appendix 11.xls".

Appendix 12 *P*-values for stepwise regression testing viruses and berry maturation parameters on total skin anthocyanins (mg/kg).

Viruses	***P*-value**
GFLV	0.0469
ArMV	n.s.
GFKL	0.0244
GLRaV1	n.s.
GLRaV2	0.0198
GLRaV3	n.s.
GLRaV7	n.s.
GVB	n.s.

Maturation	***P*-value**
Brix degree (% m/m)	n.s.
Sugars (g/l)	n.s.
Volumic mass (g/cm^3)	n.s.
Probable alcohol (% v/v)	n.s.
Total acidity (g/l tartaric acid)	n.s.

n.s. stands for non-significant *P*-value (*P>0.05)*.

Appendix 13 List of 22 cultivars used for SNP identification and previous data on skin and pulp colour and total skin anthocyanin concentration.

Cultivar ID	Cultivar Name	Skin colour[1]	Pulp colour[1]	TSA (mg/kg)[2]
41702	Gouveio-Roxo	Light skin colour	Non-coloured	_
50207	Imperial Rojo	Light skin colour	Non-coloured	_
50601	Boal Roxo	Light skin colour	Non-coloured	_
51513	Verdelho-Roxo	Light skin colour	Non-coloured	_
53904	Gewürztraminer	Light skin colour	Non-coloured	_
54005	Moscatel Roxo	Light skin colour	Non-coloured	_
52202	Negra-Mole	Dark skin colour	Non-coloured	30.71
52306	Donzelinho-Tinto	Dark skin colour	Non-coloured	267.07
51402	Mourisco-de-Semente	Dark skin colour	Non-coloured	356.49
50706	Tinta Míúda de Fontes	Dark skin colour	Non-coloured	363.21
52206	Touriga-Nacional	Dark skin colour	Non-coloured	665.08
51002	Castelã	Dark skin colour	Non-coloured	733.39
52205	Touriga-Franca	Dark skin colour	Non-coloured	879.44
53307	Tinto Cão	Dark skin colour	Non-coloured	973.17
50806	Padeiro de Basto	Dark skin colour	Non-coloured	1157.94
52004	Cornifesto	Dark skin colour	Non-coloured	1510.68
52502	Tinta-Francisca	Dark skin colour	Non-coloured	1533.05
52103	Pau Ferro	Dark skin colour	Non-coloured	2271.94
50804	Grand Noir de la Calmette	Dark skin colour	Coloured	_
51206	Petit-Bouchet	Dark skin colour	Coloured	_
53704	Aramon Noir	Dark skin colour	Coloured	_
53807	Teinturier	Dark skin colour	Coloured	_

[1]Eiras-Dias, J.E., personal communication. [2]Garcez, R.M. (1997). BSc Thesis. Caracterização de cultivares de videira através dos perfis antociânicos (HPLC). Instituto Politécnico de Santarém. Escola Superior Agrária de Santarém.

Appendix 14 Sequences of primer pairs successfully used on 22 cultivars.

Gene Code	Primer pair	Sequence (5′ – 3′)
CHI	1_2F	CCGAATTGCAAAATTTGGTG
CHI	1_2R	GAAAATCTCGCCAAAATCCA
CHI	1_4F	GGGTCGCCAGTATTCAGACA
CHI	1_4R	CAATATTTAATTGGGATGGTTTTT
CHSA	2A_4F	ACTTGTGAAGGCCATTTTCG
CHSA	2A_4R	CTTCCTCCTCATCTCGTCCA
CHSC	2C_5F	AAATTGAACGCCCACTGTTC
CHSC	2C_5R	AGTCTTGGTAAGGCGGGATT
DFR	3_2F	ATGGATTGACTCATAGTGGAGTTGA
DFR	3_2R	AATGCACTTTGTAATTGGATTGAGA
DFR	3_4F	GCGGATCAGATAAGAAATTAATCGT
DFR	3_4R	GCTCCAGGAGCCTCATGAC
DFR	3_5F	CACAAAGTGAAACCGTGTGC
DFR	3_5R	GCAAGATCTGCCTTCCAGAG
DFR	3_6F	ACTGTTTTGTGCTCAGCTAACGT
DFR	3_6R	AGTTAATCATGAACAAACAGCCATT
DFR	3_7F	GTCCACAAATGAAATGGCTGTT
DFR	3_7R	GGTATTTTTCTCTAAGCATTTTTGCA
DFR	3_8F	TCCCACGATTGTATCATTCTCG
DFR	3_8R	TCTTCTAGGTCTTGCCATCTACAGG
DFR	3_9F	GCTGACAGATTTGGGGTTTG
DFR	3_9R	CCCTATGCAAACACAAACGA
F3H	4_2F	TCATCATCCATTTATAGTCTTTGATCTC
F3H	4_2R	AGTCGTTGTAGGCCACCTTG
F3H	4_3F	CTCTTGCAGGCGAGAAGACT
F3H	4_3R	AGGTTGAACGGTGATCCAAG
F3'H$_B$	5B_2F	AAACGTTGACTCTGAAGGAGCTA
F3'H$_B$	5B_2R	GCGCTGTTGTAGCTGAAAAAT
F3'H$_B$	5B_4F	CCACTCTCTCTGTATACACTACACATC
F3'H$_B$	5B_4R	TTCTGCCATGTCAACAGACG
F3'H$_B$	5B_5F	TTTTCTGTCACTGTTTCGTCTACT
F3'H$_B$	5B_5R	GAAGCTACTCCTTCTTCCCTGA
LDOX	6_2F	GCGGTTTTCTCTTCTAAGTTCTAGCC
LDOX	6_2R	CAAAATAGTTTAATCAAAGACCACAAA
LDOX	6_3F	TGTCATGAATAAATACCAAAAACATT
LDOX	6_3R	CCGCACCAGATTCAACACTCA
LDOX	6_4F	CGGTCTAAATCTCACAAGGGTTAGAAG
LDOX	6_4R	TGTCTTAGGCCAGATGGTCA
LDOX	6_5F	CAAGCTTGCCAACAATGCTA
LDOX	6_5R	TAGAGCCTCCTGGGTCTTCC
LDOX	6_6F	GACGGTGTCTGAGACTGAGC
LDOX	6_6R	AGCTTTCTTCCCACCACTCA
UFGT	7_2F	AAGTTAAAATACAGTTTTTGGGTCTTT
UFGT	7_2R	TGGTTTGAGACATGGTTGGA
UFGT	7_3F	TCAACAGCCAAAACCCAAAT

Gene Code	Primer pair	Sequence (5′ – 3′)
UFGT	7_3R	GCAGTCGCCTTAGGTAGCAC
UFGT	7_4F	AAGACGAGCTGCTCAATTTCA
UFGT	7_4R	TTAGGACCAACTGCCCTGTC
UFGT	7_5F	GGATGCTTTGGAGATTGGAG
UFGT	7_5R	TGTATTTTGCTGTCTGGTAAGAGC
MRP	8_2F	GATTGTTCACCCATTTGTGG
MRP	8_2R	CGTCCTCCATAGTTTTTGTCC
MRP	8_3F	CCTGTGCCCCTGAATCTTAT
MRP	8_3R	GGGACAGATGTTCCTAAAAGCA
MRP	8_4F	GCCATGGCTCGTGAACTT
MRP	8_4R	AAAGAGCTCCAAGTCCCACA
MRP	8_5F	TGGTCGATGCTTATCGAATG
MRP	8_5R	TGCAACTCACAGTTCCTGAAA
MRP	8_6F	ACAGTACAACGGGGGCATAA
MRP	8_6R	GCTCAAATGCTGTCCCTGAT
MRP	8_7F	TCTGCCTTCTGCATTTCTAGG
MRP	8_7R	GGCCATAATTCCAATGGTTG
MRP	8_8F	TTCAGTGTTGTGGATTTCGAT
MRP	8_8R	CCTTCAGATTTTGCAACTCTATCC
MRP	8_9F	AGTTCATGCGCATACCACCT
MRP	8_9R	GGAAGACTGCTGATTGTTGC
MRP	8_10F	TCACATATCTGATTGGTTTGGTTC
MRP	8_10R	GTTGCCTGCTCATTCCCTAC
MRP	8_11F	GGCCATCTTTTTCATTGCTT
MRP	8_11R	GAGAGGGAGACTTATCCTGGTG
GST	9_4F	GCCACGACATCTTTTTCTGG
GST	9_4R	CTTGCCCAAAAGGCTACAAG
GST	9_5F	TGAGCTCAATTTTCTTGTTATGTTG
GST	9_5R	AATAAATGAGACTCGTTGAATGGA
MYC$_A$	10_1F	TCTTAGAATTGAAGCATGGTGTG
MYC$_A$	10_1R	GGACGTTACGAGTCCAATCAA
MYC$_A$	10_3F	AGAGCGCAGAAACAAACCTC
MYC$_A$	10_3F	CACCATCGCTTTATCCTCCT
MYC$_A$	10_4F	TTTCTGGGGCGTTTTTATTG
MYC$_A$	10_4F	CCCCTGATGAATGGCAAATA
MYC$_A$	10_5F	GCTGGGTTCTGTGAGGTCAT
MYC$_A$	10_5F	TTCTTCTGGAGCTCGGTGAT
MYC$_A$	10_6F	CAAGATGGACAAAGCCTCCT
MYC$_A$	10_6F	TTGCAAGATGAATACTTCCTTATGA
Myb11	11_2F	TTGATGGTCAATAATAAGGGATAGATG
Myb11	11_2R	AAAACCCTGGTTGGATGTGG
Myb11	11_3F	TGCGAACCAGAGATTGCTTT
Myb11	11_3R	CACTTACAAGGGAAAGTTGAAAATGA
Myb11	11_4F	GGGAACAGGTGTGCTTCGTT
Myb11	11_4R	ACCAGTCCATTTCGCAATCA

Gene Code	**Primer pair**	**Sequence (5′ – 3′)**
Myb11	11_5F	AAGCCCCACTGCTGATGAAT
Myb11	11_5R	TCAATCCCTTCATGAACATTTGAC
Myb9	12_2F	CCATGATGTGAGCCAATATGAGTT
Myb9	12_2R	TTAGTGTGAGCTTTCTCACAACAGG
Myb9	12_3F	ATGGGTAGGTCTCCCTGTTGTG
Myb9	12_3R	CCAAATTTCTGTAATCCAAAACACC
MYC_B	13_1F	GGTTGTGATTCACGCCTTATG
MYC_B	13_1R	TGCTGCACACGTGTCAATTC
MYC_B	13_2F	TGCAGATGTACCCAGTCAAAGC
MYC_B	13_2R	AGCACTTTCTTCCGGTGCTC
MYC_B	13_3F	TAGGATGGGGTGACGGCTAC
MYC_B	13_3R	ACTCGGGCTTCAGGGAATG
MYC_B	13_4F	CAGCTACTGGAGCTTCCAATCC
MYC_B	13_4R	GGAGGTGGTCGAGATGAACC
MYC_B	13_5F	TTCACAGTACTCCGGTTCATCTC
MYC_B	13_5R	TGGATTTATGGTAACTGCAGAAAGA
MYBCC	14_3F	GCATAAGGGTCTCATGTCAAGC
MYBCC	14_3R	AATGTCTTATGACAGCTGAGGAACTC
MYBCC	14_4F	TGCCTGTTCATAAGCGTAGTGG
MYBCC	14_4R	ACAGAGGGTTCTATTCTAAGCCATC
MYBCC	14_5F	TACCTTCCGGACTCATCATCTG
MYBCC	14_5R	CCTCATTTTACCATACTGGGATTTC
MYBCC	14_6F	GAAATCCCAGTATGGTAAAATGAGG
MYBCC	14_6R	TTCCTTCGCAGCCTTATCTAGG
MYBCC	14_7F	GACAAGACTGACCCAGCAACTC
MYBCC	14_7R	TTTCAACATGCTCTCTCTAGCAAC

Appendix 15 List of 445 polymorphisms identified among 22 cultivars.

Gene	Marker ID
CHI	48_2129343_S
CHI	48_2129749_S
CHS_A	168_495863_S
CHS_A	168_496247_S
CHS_A	168_496274_S
CHS_A	168_496343_S
CHS_c	9_1264427_S
CHS_c	9_1264448_S
CHS_c	9_1264516_S
CHS_c	9_1264521_S
CHS_c	9_1264529_S
CHS_c	9_1264555_S
CHS_c	9_1264611_S
CHS_c	9_1264687_S
CHS_c	9_1264694_S
CHS_c	9_1264782_S
CHS_c	9_1264820_S
CHS_c	9_1264886_S
CHS_c	9_1264955_S
CHS_c	9_1265027_S
CHS_c	9_1265081_S
F3H	83_460830_S
F3H	83_461117_S
F3H	83_461217_S
F3H	83_461238_S
F3H	83_461373_S
F3H	83_461378_S
F3H	83_461386_S
F3H	83_461395_S
F3H	83_461484_S
F3H	83_461643_S
$F3'H_B$	12_4318608_S
$F3'H_B$	12_4318658_S
$F3'H_B$	12_4318662_S
$F3'H_B$	12_4318838.5_I
$F3'H_B$	12_4318842_S
$F3'H_B$	12_4319929_S
$F3'H_B$	12_4319966_S
$F3'H_B$	12_4320007_S
GST	30_2386123_S
GST	30_2386277_S
GST	30_2386316_S
GST	30_2386326_S
GST	30_2386339_S
GST	30_2386811_S
GST	30_2386852_S
GST	30_2387078_S
GST	30_2387101_S
GST	30_2387122_S
GST	30_2387138_S
GST	30_2387242_S
GST	30_2387258_S
MYB9	83_144212_S
MYB9	83_144279_S
MYB9	83_144380_S
MYB9	83_144881_S
MYB9	83_145083_S
MYB_{CC}	342_112073_S
MYB_{CC}	342_112113_S
MYB_{CC}	342_112421_S
MYB_{CC}	342_112436_S
MYB_{CC}	342_112440_S
MYB_{CC}	342_112473.5_I
MYB_{CC}	342_112501_S
MYB_{CC}	342_112645_S
MYB_{CC}	342_112650_S
MYB_{CC}	342_112718_S
MYB_{CC}	342_112719_S
MYB_{CC}	342_112732_S
MYB_{CC}	342_112788_S
MYB_{CC}	342_112815_S
MYB_{CC}	342_112834_S
MYB_{CC}	342_112888_S
MYB_{CC}	342_112910_S
MYB_{CC}	342_113010_S
MYB_{CC}	342_113158_S
MYB_{CC}	342_113169_S
MYB_{CC}	342_113232_S
MYB_{CC}	342_113348_S
MYB_{CC}	342_113353_S
MYB_{CC}	342_113354_S
MYB_{CC}	342_113439_S
MYB_{CC}	342_113440_S
MYB_{CC}	342_113453_S
MYB_{CC}	342_113550_S
MYB_{CC}	342_113699_S
MYB_{CC}	342_113799_S
MYB_{CC}	342_113821_S
MYB_{CC}	342_113848_S
MYB_{CC}	342_113919_S
MYB_{CC}	342_114151_S
MYB_{CC}	342_114160_S
MYB_{CC}	342_114222_S
MYB_{CC}	342_114241_S
MYB_{CC}	342_114261_S
MYB_{CC}	342_114338_S
MYB_{CC}	342_114450_S
MYB_{CC}	342_114488_S
MYB_{CC}	342_114543_S
MYB_{CC}	342_114565_S
MYB_{CC}	342_114648_S
MYB_{CC}	342_114684_S
MYB_{CC}	342_114688_S
MYB_{CC}	342_114753_S
MYB_{CC}	342_115316_S
MYB_{CC}	342_115359_S
MYB_{CC}	342_115367_S
MYB_{CC}	342_115408_S
MYC_A	203_203671_S
MYC_A	203_203694_S
MYC_A	203_203755_S
MYC_A	203_203787_S
MYC_A	203_203803_S
MYC_A	203_203807_S

Gene	Marker ID
MYC$_A$	203_204285_S
MYC$_A$	203_204329_S
MYC$_A$	203_204582_S
MYC$_A$	203_204708_S
MYC$_A$	203_205101_S
MYC$_A$	203_205150_S
MYC$_A$	203_205413_S
MYC$_A$	203_205425_S
MYC$_A$	203_205540_S
MYC$_A$	203_205599_S
MYC$_A$	203_205727_I
MYC$_A$	203_205967_S
MYC$_A$	203_205968_S
MYC$_A$	203_206032_S
MYC$_A$	203_206115.5_I
MYC$_A$	203_206754.5_I
MYC$_A$	203_206797_S
MYC$_A$	203_206884_S
LDOX	112_321532_S
LDOX	112_321571_S
LDOX	112_321661_S
LDOX	112_321729_S
LDOX	112_321732_S
LDOX	112_323166.5_I
LDOX	112_323391_S
LDOX	112_323489_S
LDOX	112_323523_S
LDOX	112_323698_S
LDOX	112_323745_S
LDOX	112_323777_S
LDOX	112_323789_S
LDOX	112_323794_S
LDOX	112_323802_S
LDOX	112_323865_S
LDOX	112_323868_S
LDOX	112_323878_S
LDOX	112_323998_S
UFGT	10_2334592_S
UFGT	10_2334638.5_I
UFGT	10_2334801_S
UFGT	10_2334873_S
UFGT	10_2334901_S
UFGT	10_2334914_S
UFGT	10_2334953_S
UFGT	10_2334981_S
UFGT	10_2334986_S
UFGT	10_2335000_S
UFGT	10_2335012_S
UFGT	10_2335015_S
UFGT	10_2335024_S
UFGT	10_2335092_S
UFGT	10_2335132_S
UFGT	10_2335191_S
UFGT	10_2335203_S
UFGT	10_2335296_S
UFGT	10_2335303_S
UFGT	10_2335305_S
UFGT	10_2335330_S
UFGT	10_2335357_S
UFGT	10_2335384_S
UFGT	10_2335483_S
UFGT	10_2335510_S
UFGT	10_2335527_S
UFGT	10_2335529_S
UFGT	10_2335546_S
UFGT	10_2335548_S
UFGT	10_2335584_S
UFGT	10_2335586_S
UFGT	10_2335591_S
UFGT	10_2335621_S
UFGT	10_2335663_S
UFGT	10_2335687_S
UFGT	10_2335704_S
UFGT	10_2335721_S
UFGT	10_2335770_S
UFGT	10_2335779_S
UFGT	10_2335836_S
UFGT	10_2335873_S
UFGT	10_2335880_S
UFGT	10_2335881_S
UFGT	10_2335888_S
UFGT	10_2335905_S
UFGT	10_2335907_S
UFGT	10_2335925_S
UFGT	10_2335938_S
UFGT	10_2335940_S
UFGT	10_2336006_S
UFGT	10_2336047_S
UFGT	10_2336055_S
UFGT	10_2336072_S
UFGT	10_2336217.5_I
UFGT	10_2336241.5_I
UFGT	10_2336248.5_I
UFGT	10_2336289_S
UFGT	10_2336293_S
UFGT	10_2336324.5_I
UFGT	10_2336330_S
UFGT	10_2336339_S
UFGT	10_2336396_S
UFGT	10_2336423_S
UFGT	10_2336457_S
UFGT	10_2336459_S
UFGT	10_2336468.5_I
UFGT	10_2336527_S
UFGT	10_2336550_S
UFGT	10_2336557_S
UFGT	10_2336588_S
UFGT	10_2336592_S
UFGT	10_2336603_S
UFGT	10_2336647_S
MRP	7_2189043_S
MRP	7_2189049_S
MRP	7_2189133_S
MRP	7_2189378_S
MRP	7_2189422_S
MRP	7_2189451.5_I
MRP	7_2189451_I
MRP	7_2189482_S
MRP	7_2189510_S
MRP	7_2189517_S

Gene	Marker ID
MRP	7_2189720_S
MRP	7_2189791.5_I
MRP	7_2189816_S
MRP	7_2189842_S
MRP	7_2189864_S
MRP	7_2189907_S
MRP	7_2189914_S
MRP	7_2189918_S
MRP	7_2190152_S
MRP	7_2190290_S
MRP	7_2190477_S
MRP	7_2190688_S
MRP	7_2190710_S
MRP	7_2190713_S
MRP	7_2190745_S
MRP	7_2190751_S
MRP	7_2191068_S
MRP	7_2191168_S
MRP	7_2191189_S
MRP	7_2191250_S
MRP	7_2191309_S
MRP	7_2191312_S
MRP	7_2191471_S
MRP	7_2191540_S
MRP	7_2191654_S
MRP	7_2191720_S
MRP	7_2191883_S
MRP	7_2191947_S
MRP	7_2192024_S
MRP	7_2192079_S
MRP	7_2192648_S
MRP	7_2192993_S
MRP	7_2193140.5_I
MRP	7_2193381.5_I
MRP	7_2194080_S
MRP	7_2194107_S
MRP	7_2194111_S
MRP	7_2194113_S
MRP	7_2194117_S
MRP	7_2194155_S

Gene	Marker ID
MRP	7_2194174_S
MRP	7_2194218_S
MRP	7_2194502_S
MRP	7_2194527_S
MRP	7_2194538_S
MRP	7_2194571_S
MRP	7_2194585_S
MRP	7_2194931_S
MRP	7_2194977_S
MRP	7_2195066_S
MRP	7_2195080_S
MRP	7_2195168_S
MRP	7_2195480_S
MRP	7_2195512_S
MRP	7_2195520_S
MRP	7_2195628_S
MRP	7_2195659_S
MRP	7_2195662_S
MRP	7_2195772.5_I
MRP	7_2195860_S
MYB11	7_1171043_S
MYB11	7_1171133_S
MYB11	7_1171172_S
MYB11	7_1171190_S
MYB11	7_1171226_S
MYB11	7_1171233_S
MYB11	7_1171245_S
MYB11	7_1171268_S
MYB11	7_1171311_S
MYB11	7_1171324_S
MYB11	7_1171370_S
MYB11	7_1171521_I
MYB11	7_1171525_S
MYB11	7_1171712_S
MYB11	7_1171723_S
MYB11	7_1171762_S
MYB11	7_1171810_S
MYB11	7_1171919_S
MYB11	7_1172074_S
MYB11	7_1172084_S

Gene	Marker ID
MYB11	7_1172184_S
MYB11	7_1172194_I
MYB11	7_1172347_S
MYB11	7_1172354_S
MYB11	7_1172369_S
MYB11	7_1172444_S
MYB11	7_1172641.5_I
MYB11	7_1172655_S
MYB11	7_1172657_S
MYB11	7_1172759.5_I
MYB11	7_1172762_S
MYB11	7_1172786_S
MYB11	7_1172804_S
MYB11	7_1172977_S
MYB11	7_1173011_I
MYB11	7_1173068_S
MYB11	7_1173099_S
MYB11	7_1173102_S
MYB11	7_1173185_S
MYB11	7_1173247_S
MYB11	7_1173269.5_I
MYB11	7_1173292_S
MYB11	7_1173342_S
MYB11	7_1173378_S
MYB11	7_1173384_S
MYB11	7_1173386_S
MYB11	7_1173393_S
MYB11	7_1173407_S
MYB11	7_1173409_S
MYB11	7_1173423_S
DFR	1_2947862_S
DFR	1_2947869_S
DFR	1_2947881_S
DFR	1_2947887_S
DFR	1_2947892_S
DFR	1_2947894_S
DFR	1_2947896_S
DFR	1_2947927_I
DFR	1_2947936_S
DFR	1_2947958_S

Gene	Marker ID
DFR	1_2947966_S
DFR	1_2947969_S
DFR	1_2947989_S
DFR	1_2947990_S
DFR	1_2948003_S
DFR	1_2948041_S
DFR	1_2948089_S
DFR	1_2948091_S
DFR	1_2948123_S
DFR	1_2948136.5_I
DFR	1_2948137_S
DFR	1_2948147.5_I
DFR	1_2948153.5_I
DFR	1_2948175_S
DFR	1_2948184_S
DFR	1_2948185_S
DFR	1_2948188.5_I
DFR	1_2948213.5_I
DFR	1_2948259.5_I
DFR	1_2948286_S
DFR	1_2948370_S
DFR	1_2948381_S
DFR	1_2948391_S
DFR	1_2949038_S
DFR	1_2949122_S
DFR	1_2949136_S
DFR	1_2949171_S
DFR	1_2949207_S
DFR	1_2949211_S
DFR	1_2949225_S
DFR	1_2949293_S

Gene	Marker ID
DFR	1_2949295_S
DFR	1_2949346.5_I
DFR	1_2949352_S
DFR	1_2949385_S
DFR	1_2949410_S
DFR	1_2949431_S
DFR	1_2949441_S
DFR	1_2949447_S
DFR	1_2949550_S
DFR	1_2949581_S
DFR	1_2949596_S
DFR	1_2949864.5_I
DFR	1_2949979_S
DFR	1_2950071_S
DFR	1_2950111.5_I
DFR	1_2950119_S
DFR	1_2950132_S
DFR	1_2950145_S
DFR	1_2950160_S
DFR	1_2950179_S
DFR	1_2950189_S
DFR	1_2950204_S
DFR	1_2950238_S
DFR	1_2950245_S
DFR	1_2950316_S
DFR	1_2950327_S
DFR	1_2951032_S
DFR	1_2951059_S
DFR	1_2951348_S
DFR	1_2951857_S
DFR	1_2951866_S

Gene	Marker ID
DFR	1_2951943_S
DFR	1_2951996_S
DFR	1_2952028_S
DFR	1_2952038_S
DFR	1_2952068_S
DFR	1_2952101_S
DFR	1_2952167_S
MYC_B	11_3994483_S
MYC_B	11_3994528_I
MYC_B	11_3994752_S
MYC_B	11_3994754_S
MYC_B	11_3994908_S
MYC_B	11_3995344.5_I
MYC_B	11_3995416_S
MYC_B	11_3995500.5_I
MYC_B	11_3995556_S
MYC_B	11_3995574_S
MYC_B	11_3995740_S
MYC_B	11_3995764_S
MYC_B	11_3996076_S
MYC_B	11_3996226_S
MYC_B	11_3996452_S
MYC_B	11_3996496_S
MYC_B	11_3996646_S
MYC_B	11_3996937_S
MYC_B	11_3996949_S
MYC_B	11_3997123_S
MYC_B	11_3997795_S
MYC_B	11_3997827.5_I

The marker ID is composed by the scaffold number separated from the base pair location in this scaffold and the type of polymorphism by "_". Scaffold numbers are according to the Genoscope database with the sequencing version 8x coverage. For the type of polymorphism, *S* means SNP and *I* means INDEL.

Appendix 16 List of 140 SNPs selected for genotyping for association mapping, showing Minor Allele Frequency (MAF), Hardy Weinberg chi-square and Missing values for a sample of 22 cultivars.

SNP ID	Missing (%)	MAF	Hardy Weinberg (chi-square)
1_2947869_S	18.1818	0.2778	18.0000
1_2949122_S	22.7273	0.2353	2.0370
1_2949136_S	31.8182	0.3000	4.1157
1_2949293_S	22.7273	0.2353	2.0370
1_2949431_S	18.1818	0.4167	0.0147
1_2949447_S	4.5455	0.4286	0.5833
1_2949581_S	9.0909	0.0750	0.1315
1_2950119_S	0.0000	0.3182	0.0499
1_2950189_S	4.5455	0.3571	0.4160
1_2951348_S	4.5455	0.1190	12.5472
1_2951866_S	0.0000	0.0455	0.0499
1_2951996_S	0.0000	0.3409	0.1768
1_2952028_S	0.0000	0.4545	0.2200
1_2952038_S	0.0000	0.1136	2.2963
1_2952101_S	0.0000	0.1364	0.5485
10_2334801_S	22.7273	0.1471	1.4950
10_2334901_S	4.5455	0.1905	0.1135
10_2334953_S	4.5455	0.1190	0.3835
10_2335092_S	4.5455	0.2619	0.2472
10_2335191_S	0.0000	0.1818	3.3272
10_2335303_S	4.5455	0.1905	1.1626
10_2335305_S	0.0000	0.1591	0.7874
10_2335527_S	0.0000	0.3182	0.0499
10_2335548_S	0.0000	0.0682	0.1178
10_2335586_S	0.0000	0.3409	0.2792
10_2335663_S	0.0000	0.2045	1.4547
10_2335687_S	0.0000	0.2273	1.0992
10_2335721_S	0.0000	0.0909	0.2200
10_2335770_S	0.0000	0.3409	2.1822
10_2335779_S	0.0000	0.0682	0.1178
10_2335873_S	0.0000	0.2955	1.2225
10_2335888_S	0.0000	0.2045	0.0109
10_2335940_S	0.0000	0.2500	0.5051
10_2336055_S	0.0000	0.3182	0.5767
10_2336289_S	13.6364	0.1842	4.2806
10_2336293_S	0.0000	0.0455	22.0000
11_3994483_S	22.7273	0.0882	0.1592
11_3994752_S	59.0909	0.4444	9.0000
11_3995416_S	4.5455	0.2857	1.8900
11_3995574_S	0.0000	0.0455	22.0000
11_3995740_S	4.5455	0.1429	7.8426
11_3995764_S	0.0000	0.2955	4.5364
11_3996076_S	0.0000	0.0909	0.2200
11_3996452_S	4.5455	0.1905	0.1135

SNP ID	Missing (%)	MAF	Hardy Weinberg (chi-square)
11_3996646_S	0.0000	0.2500	0.5051
11_3997123_S	0.0000	0.2045	1.4547
112_321732_S	0.0000	0.2500	0.5051
112_323489_S	4.5455	0.2381	0.9483
112_323745_S	9.0909	0.3500	0.2922
12_4318608_S	50.0000	0.2727	11.0000
12_4318658_S	40.9091	0.3462	8.9571
12_4318662_S	40.9091	0.3462	8.9571
12_4319929_S	22.7273	0.4412	2.7679
12_4319966_S	27.2727	0.4375	3.8740
12_4320007_S	22.7273	0.4412	2.7679
168_494768_SD	_	_	_
168_495863_S	0.0000	0.1591	0.4988
168_496247_S	4.5455	0.1190	2.1359
168_496274_S	0.0000	0.4545	0.2200
168_496343_S	9.0909	0.4250	0.3141
203_203755_S	0.0000	0.0909	0.2200
203_203807_S	0.0000	0.3409	0.2792
203_204329_S	0.0000	0.2955	0.0066
203_205101_S	4.5455	0.1190	2.1359
203_205150_S	9.0909	0.4500	3.1038
203_205968_S	4.5455	0.4762	21.0000
203_206884_S	68.1818	0.5000	0.1429
30_2386811_S	13.6364	0.1579	0.6680
30_2387101_S	13.6364	0.4211	10.0496
30_2387122_S	13.6364	0.4211	10.0496
30_2387138_S	36.3636	0.3929	5.8616
30_2387242_S	31.8182	0.4000	6.6667
30_2387258_S	36.3636	0.3929	5.8616
342_112113_S	50.0000	0.0909	11.0000
342_112719_S	0.0000	0.3864	2.3810
342_112815_S	0.0000	0.0682	9.0754
342_113232_S	4.5455	0.2619	0.2472
342_113440_S	4.5455	0.2143	0.0021
342_114160_S	0.0000	0.1818	0.1528
342_114222_S	0.0000	0.0909	0.2200
342_114338_S	0.0000	0.3182	0.0499
342_114488_S	0.0000	0.3182	0.0499
342_114543_S	4.5455	0.4524	0.0687
342_114684_S	0.0000	0.3409	0.1768
342_114688_S	0.0000	0.1818	0.1528
48_2128153_SD	_	_	_
48_2128188_SD	_	_	_
48_2128606_SD	_	_	_
48_2129343_S	13.6364	0.4737	0.0586
48_2129749_S	13.6364	0.0526	0.0586
7_1171190_S	0.0000	0.3182	0.0499
7_1171233_S	0.0000	0.5000	0.1818
7_1171268_S	0.0000	0.1136	0.3616
7_1171324_S	0.0000	0.2955	1.2225

SNP ID	Missing (%)	MAF	Hardy Weinberg (chi-square)
7_1171370_S	0.0000	0.4773	0.0001
7_1171525_S	9.0909	0.0750	8.1828
7_1171762_S	0.0000	0.3409	10.6741
7_1171919_S	0.0000	0.2273	22.0000
7_1172074_S	0.0000	0.3636	8.1097
7_1172184_S	9.0909	0.2750	7.7831
7_1172354_S	4.5455	0.0952	0.2327
7_1172369_S	4.5455	0.4286	2.7384
7_1172444_S	4.5455	0.2857	0.0933
7_1172657_S	18.1818	0.1111	0.2813
7_1172786_S	0.0000	0.2045	0.0109
7_1173068_S	18.1818	0.1944	12.1819
7_1173102_S	22.7273	0.2059	11.4338
7_1173247_S	27.2727	0.2188	10.6836
7_1173409_S	40.9091	0.1923	7.3590
7_2189043_S	9.0909	0.1750	0.8999
7_2189133_S	9.0909	0.2500	0.8000
7_2190290_S	4.5455	0.2381	2.0508
7_2190713_S	18.1818	0.4722	7.9753
7_2191068_S	22.7273	0.0882	0.1592
7_2191250_S	4.5455	0.3571	0.4160
7_2191312_S	4.5455	0.2143	0.0021
7_2191654_S	4.5455	0.3333	1.7143
7_2192024_S	13.6364	0.4737	1.3511
7_2192648_S	4.5455	0.0714	0.1243
7_2192993_S	0.0000	0.0455	0.0499
7_2194080_S	0.0000	0.0909	0.2200
7_2194931_S	36.3636	0.1429	2.4306
7_2195659_S	9.0909	0.1500	0.9304
83_144212_S	63.6364	0.3125	1.6529
83_144380_S	63.6364	0.1250	0.1633
83_144881_S	0.0000	0.2500	2.4444
83_145083_S	0.0000	0.3182	1.4547
83_459657_SD	_	_	_
83_460830_S	0.0000	0.3409	2.1822
83_461117_S	18.1818	0.1389	10.6097
83_461217_S	22.7273	0.2059	3.6022
83_461386_S	77.2727	0.2000	5.0000
83_461643_S	68.1818	0.3571	0.0311
9_1264427_S	13.6364	0.4737	19.0000
9_1264687_S	0.0000	0.3636	22.0000
9_1264886_S	0.0000	0.1364	8.2949
9_1265027_S	0.0000	0.0909	4.4550
9_1265081_S	4.5455	0.0952	4.2029
9_1266000_SD	_	_	_
9_1266031_SD	_	_	_

The SNP ID is composed by the scaffold number separated from the base pair location in this scaffold and the type of polymorphism by "_". Scaffold numbers are according to the Genoscope database with the sequencing version 8x coverage. For the type of polymorphism, *S* means SNP and *I* means INDEL.

Appendix 17 List of 124 SNPs used for association mapping after filtering according to quality control criteria on genotype data on 149 individuals.

SNP ID_1	SNP ID_2	Scaffold	Location	MAF	Hardy Weinberg (chi-square)	Missing (%)
1_2949136_S	s1	1	2949.136	0.3286	5.4866	6.04
1_2949293_S	s2	1	2949.293	0.1926	0.0732	0.67
1_2949431_S	s3	1	2949.431	0.4078	0.2552	5.37
1_2949447_S	s4	1	2949.447	0.4757	1.4349	3.36
1_2950119_S	s5	1	2950.119	0.2621	1.6142	2.68
1_2950189_S	s6	1	2950.189	0.3169	0.0822	4.70
1_2951348_S	s7	1	2951.348	0.2347	0.0020	1.34
1_2951866_S	s8	1	2951.866	0.0608	0.6205	0.67
1_2951996_S	s9	1	2951.996	0.2551	2.3961	1.34
1_2952028_S	s10	1	2952.028	0.3964	0.1252	6.04
1_2952038_S	s11	1	2952.038	0.0342	0.1836	2.01
1_2952101_S	s12	1	2952.101	0.1267	1.0107	2.01
10_2334801_S	s13	10	2334.801	0.1074	0.3767	0.00
10_2334901_S	s14	10	2334.901	0.2349	0.1259	0.00
10_2334953_S	s15	10	2334.953	0.1047	0.2986	0.67
10_2335092_S	s16	10	2335.092	0.2959	0.0026	1.34
10_2335191_S	s17	10	2335.191	0.2041	0.0039	1.34
10_2335303_S	s18	10	2335.303	0.2034	0.2617	2.68
10_2335305_S	s19	10	2335.305	0.1014	0.1863	4.03
10_2335527_S	s20	10	2335.527	0.3058	7.8357	6.71
10_2335548_S	s21	10	2335.548	0.0884	1.3836	1.34
10_2335586_S	s22	10	2335.586	0.3231	3.0767	1.34
10_2335663_S	s23	10	2335.663	0.1632	0.2597	3.36
10_2335687_S	s24	10	2335.687	0.3322	0.0070	4.03
10_2335721_S	s25	10	2335.721	0.0586	0.5623	2.68
10_2335770_S	s26	10	2335.77	0.3681	1.5787	3.36
10_2335873_S	s27	10	2335.873	0.3425	0.1038	2.01
10_2335888_S	s28	10	2335.888	0.2517	0.6383	2.68
10_2335940_S	s29	10	2335.94	0.3380	0.0844	4.70
10_2336055_S	s30	10	2336.055	0.3681	0.2915	3.36
10_2336289_S	s31	10	2336.289	0.2061	0.1289	0.67
11_3994483_S	s32	11	3994.483	0.0634	0.6502	4.70

SNP ID_1	SNP ID_2	Scaffold	Location	MAF	Hardy Weinberg (chi-square)	Missing (%)
11_3994752_S	s33	11	3994.752	0.3163	0.2424	1.34
11_3995416_S	s34	11	3995.416	0.3188	0.4910	0.00
11_3995574_S	s35	11	3995.574	0.2000	0.3879	2.68
11_3995740_S	s36	11	3995.74	0.1493	0.6303	3.36
11_3995764_S	s37	11	3995.764	0.3087	0.0060	0.00
11_3996076_S	s38	11	3996.076	0.0805	1.1432	0.00
11_3996452_S	s39	11	3996.452	0.2047	0.0150	0.00
11_3996646_S	s40	11	3996.646	0.3074	0.0000	0.67
11_3997123_S	s41	11	3997.123	0.1711	3.7747	0.00
112_321732_S	s42	112	321.732	0.1905	0.0318	1.34
112_323489_S	s43	112	323.489	0.3129	0.0229	1.34
112_323745_S	s44	112	323.745	0.2774	0.5313	2.01
12_4318608_S	s45	12	4318.608	0.3912	6.7620	1.34
12_4319966_S	s46	12	4319.966	0.3370	7.8033	7.38
12_4320007_S	s47	12	4320.007	0.4132	4.8610	3.36
168_494768_SD	s48	168	494.768	0.4261	1.2238	4.70
168_495863_S	s49	168	495.863	0.2215	1.6351	0.00
168_496247_S	s50	168	496.247	0.1107	0.0207	0.00
168_496274_S	s51	168	496.274	0.3381	0.5142	6.71
168_496343_S	s52	168	496.343	0.4595	0.5512	0.67
203_203755_S	s53	203	203.755	0.1520	4.7571	0.67
203_203807_S	s54	203	203.807	0.2692	0.0356	12.75
203_204329_S	s55	203	204.329	0.2568	4.1981	0.67
203_205101_S	s56	203	205.101	0.0676	0.1791	0.67
203_205150_S	s57	203	205.15	0.3759	0.2762	2.68
203_205968_S	s58	203	205.968	0.2669	1.1407	0.67
30_2386811_S	s59	30	2386.811	0.1473	4.3540	2.01
30_2387242_S	s60	30	2387.242	0.2310	1.6401	2.68
30_2387258_S	s61	30	2387.258	0.2448	2.7491	2.68
342_112113_S	s62	342	112.113	0.3732	0.1969	7.38
342_112719_S	s63	342	112.719	0.3562	0.0353	2.01
342_112815_S	s64	342	112.815	0.0448	1.8889	2.68
342_113232_S	s65	342	113.232	0.2081	2.9419	0.00
342_113440_S	s66	342	113.44	0.1520	0.0719	0.67
342_114160_S	s67	342	114.16	0.1294	0.0853	4.03
342_114222_S	s68	342	114.222	0.0612	0.6252	1.34

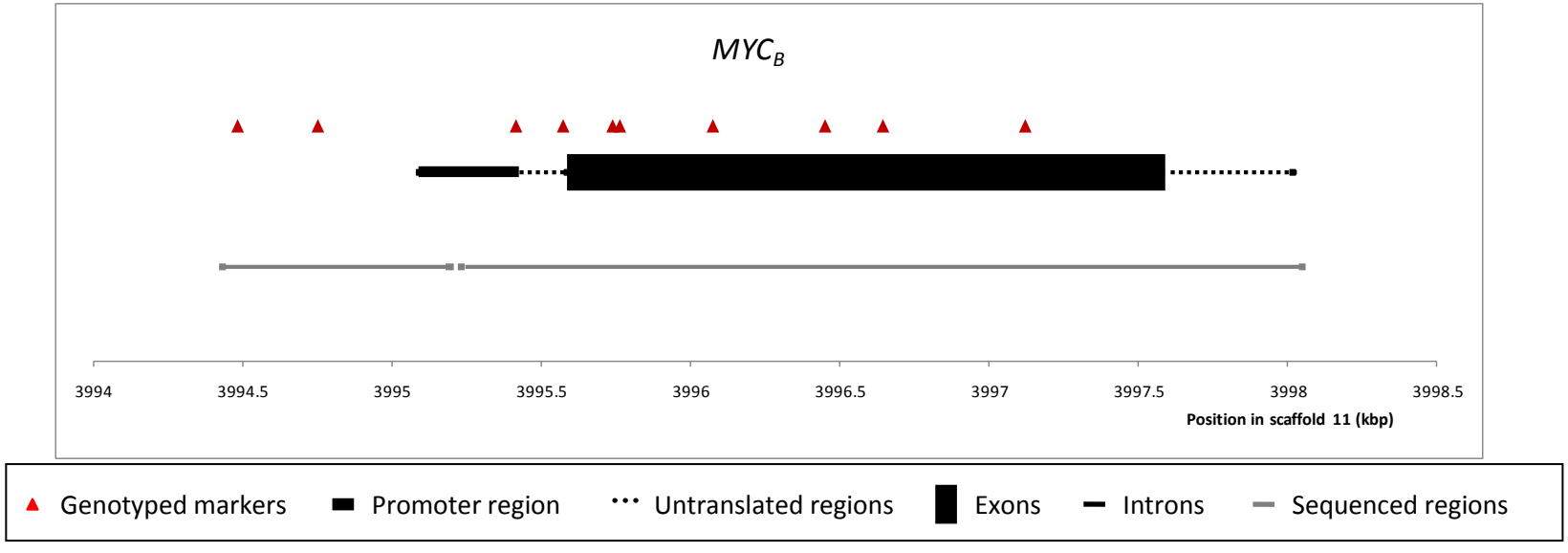
SNP ID_1	SNP ID_2	Scaffold	Location	MAF	Hardy Weinberg (chi-square)	Missing (%)
342_114338_S	s69	342	114.338	0.3557	0.0928	0.00
342_114488_S	s70	342	114.488	0.2206	0.0364	8.72
342_114543_S	s71	342	114.543	0.4760	0.4042	2.01
342_114684_S	s72	342	114.684	0.3537	0.0203	1.34
342_114688_S	s73	342	114.688	0.1486	0.6808	0.67
48_2128606_SD	s74	48	2128.606	0.0270	0.1142	0.67
48_2129343_S	s75	48	2129.343	0.4388	1.2224	1.34
48_2129749_S	s76	48	2129.749	0.0486	0.3759	3.36
7_1171190_S	s77	7	1171.190	0.2657	0.1495	4.03
7_1171233_S	s78	7	1171.233	0.4401	0.7300	4.70
7_1171268_S	s79	7	1171.268	0.1275	0.0969	0.00
7_1171324_S	s80	7	1171.324	0.4027	1.7096	0.00
7_1171370_S	s81	7	1171.370	0.3844	0.8962	1.34
7_1171525_S	s82	7	1171.525	0.3389	0.1045	0.00
7_1171762_S	s83	7	1171.762	0.3826	0.3921	0.00
7_1171919_S	s84	7	1171.919	0.4865	0.0001	0.67
7_1172074_S	s85	7	1172.074	0.4000	0.0766	2.68
7_1172184_S	s86	7	1172.184	0.4000	0.7143	6.04
7_1172354_S	s87	7	1172.354	0.0680	0.7832	1.34
7_1172369_S	s88	7	1172.369	0.3592	0.2305	4.70
7_1172444_S	s89	7	1172.444	0.3345	2.6934	0.67
7_1172657_S	s90	7	1172.657	0.0594	0.5711	4.03
7_1172786_S	s91	7	1172.786	0.3252	9.0164	4.03
7_1173068_S	s92	7	1173.068	0.0621	0.3964	2.68
7_1173102_S	s93	7	1173.102	0.0705	0.1058	0.00
7_1173247_S	s94	7	1173.247	0.3542	4.6794	3.36
7_2189043_S	s95	7	2189.043	0.2585	0.8774	1.34
7_2189133_S	s96	7	2189.133	0.3309	3.3487	6.71
7_2190290_S	s97	7	2190.290	0.1846	0.2533	0.00
7_2190713_S	s98	7	2190.713	0.4191	4.6317	8.72
7_2191068_S	s99	7	2191.068	0.0604	0.4340	0.00
7_2191250_S	s100	7	2191.250	0.2987	0.0129	0.00
7_2191312_S	s101	7	2191.312	0.1973	0.8035	1.34
7_2191654_S	s102	7	2191.654	0.3054	0.5352	0.00
7_2192024_S	s103	7	2192.024	0.4896	0.7038	3.36
7_2192648_S	s104	7	2192.648	0.1042	1.9470	3.36

SNP ID_1	SNP ID_2	Scaffold	Location	MAF	Hardy Weinberg (chi-square)	Missing (%)
7_2192993_S	s105	7	2192.993	0.0338	0.1809	0.67
7_2194080_S	s106	7	2194.080	0.1000	0.1724	2.68
7_2194931_S	s107	7	2194.931	0.2133	0.0634	4.03
7_2195659_S	s108	7	2195.659	0.2158	0.0566	6.71
83_144212_S	s109	83	144.212	0.1667	0.3600	3.36
83_144380_S	s110	83	144.380	0.0816	1.1615	1.34
83_144881_S	s111	83	144.881	0.3221	4.2006	0.00
83_145083_S	s112	83	145.083	0.2603	0.6603	2.01
83_459657_SD	s113	83	459.657	0.3142	0.0222	0.67
83_460830_S	s114	83	460.830	0.2226	2.3930	2.01
83_461217_S	s115	83	461.217	0.3682	0.4654	0.67
83_461386_S	s116	83	461.386	0.3537	0.8836	1.34
83_461643_S	s117	83	461.643	0.1929	4.2407	6.04
9_1264427_S	s118	9	1264.427	0.3255	0.0063	0.00
9_1264687_S	s119	9	1264.687	0.3758	3.1066	0.00
9_1264886_S	s120	9	1264.886	0.0671	0.7712	0.00
9_1265027_S	s121	9	1265.027	0.0304	0.1455	0.67
9_1265081_S	s122	9	1265.081	0.0302	0.1445	0.00
9_1266000_SD	s123	9	1266.000	0.1757	0.1038	0.67
9_1266031_SD	s124	9	1266.031	0.3836	2.4584	2.01

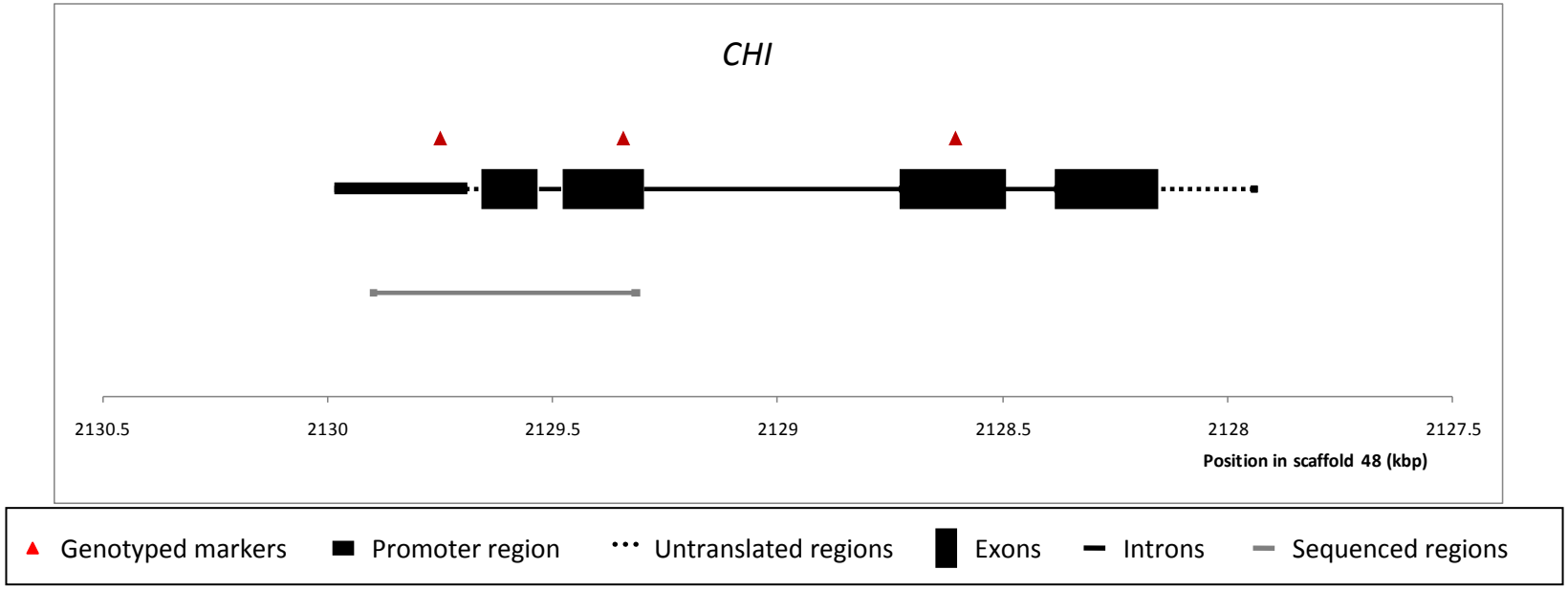
The SNP ID_1 is composed by the scaffold number separated from the base pair location in this scaffold and the type of polymorphism by "_". Scaffold numbers are according to the Genoscope database with the sequencing version 8x coverage. For the type of polymorphism, *S* means SNP and *I* means INDEL.

SNP ID_1matches the SNP ID used in previous stage of SNP selection. SNP ID_2 was created for writing convenience.

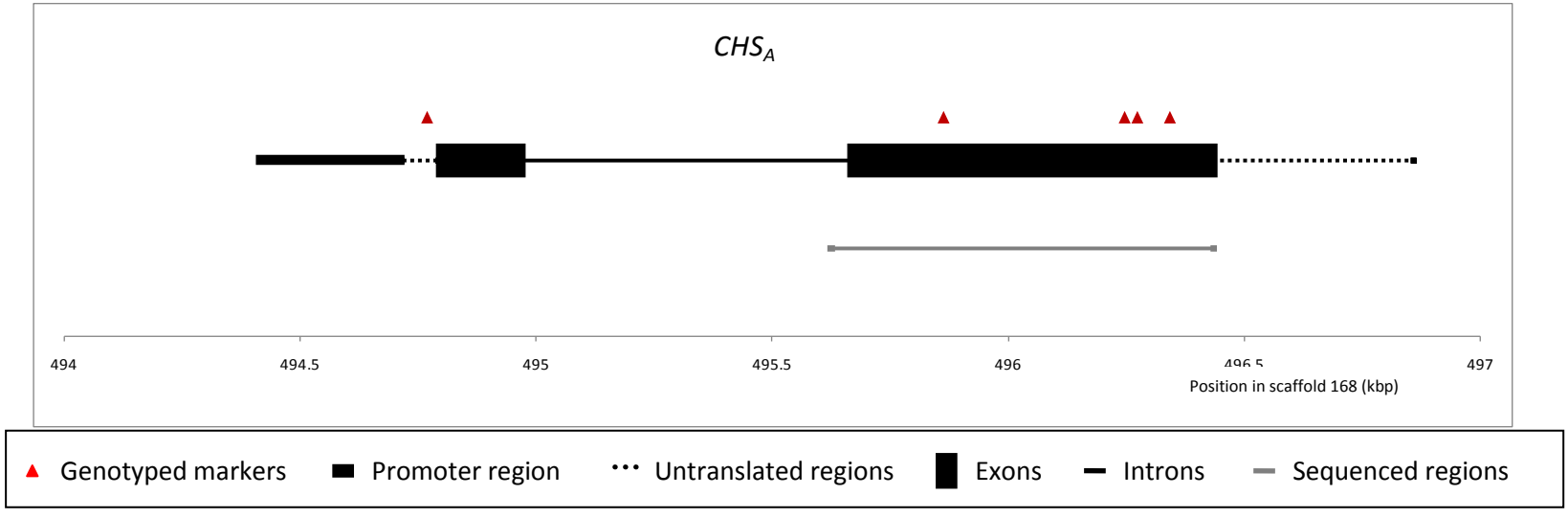
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs.



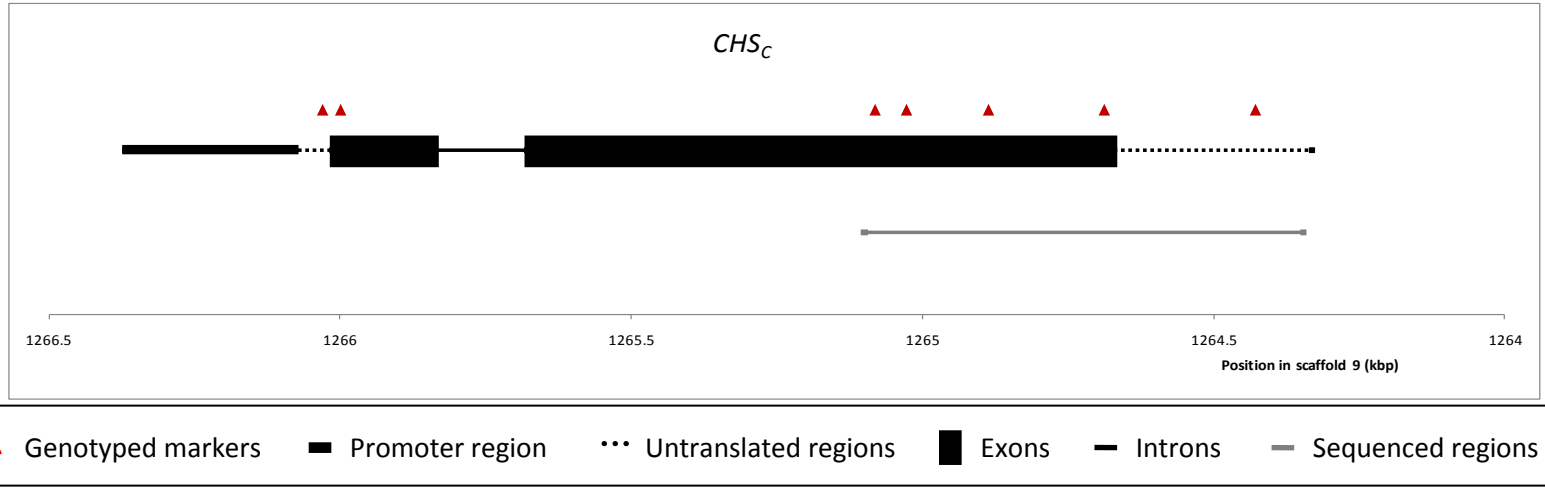
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



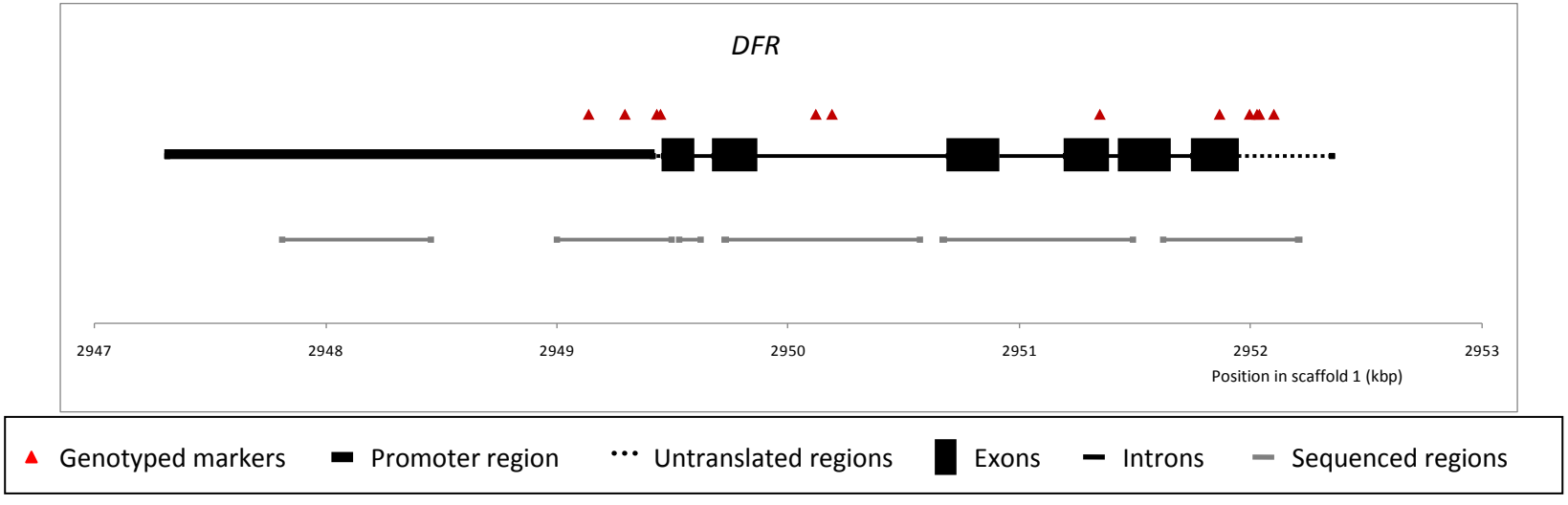
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



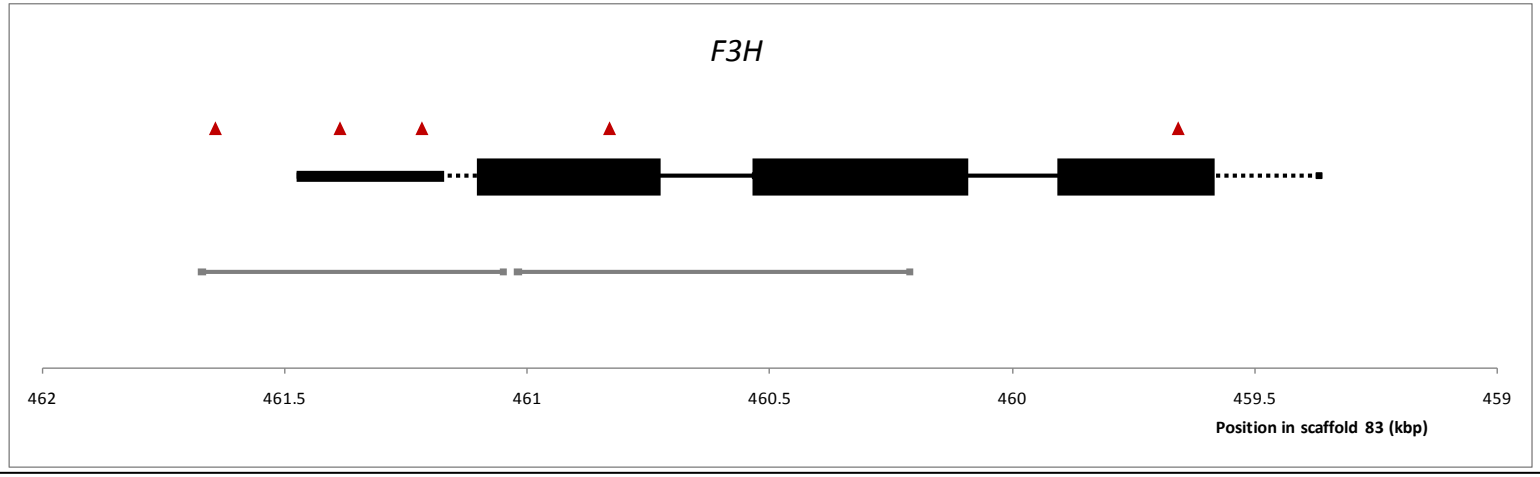
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



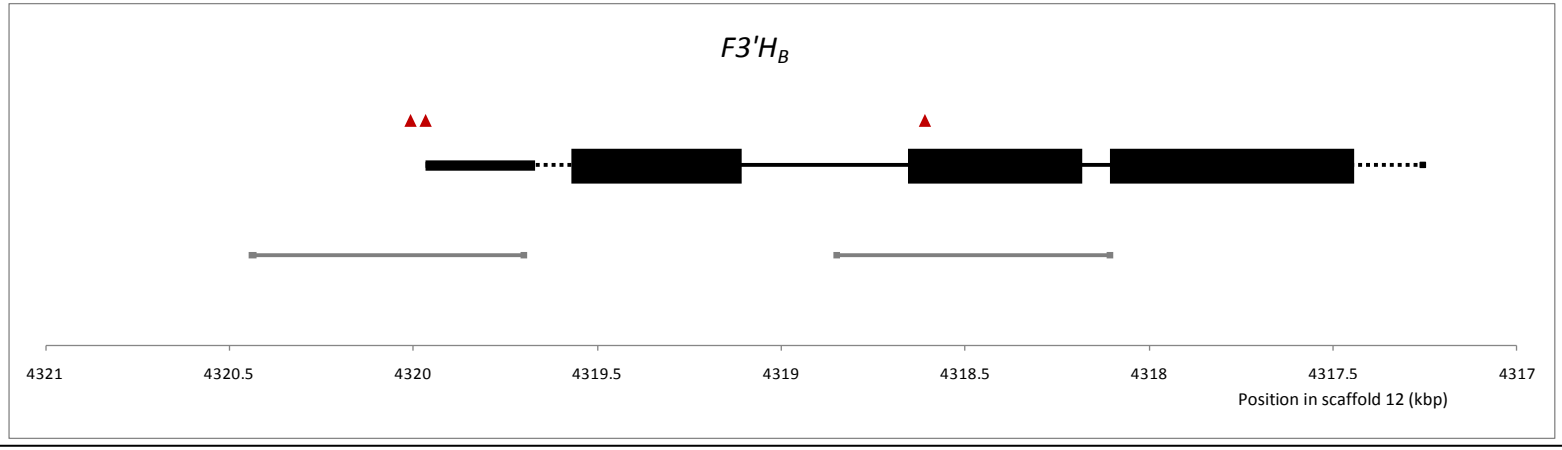
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



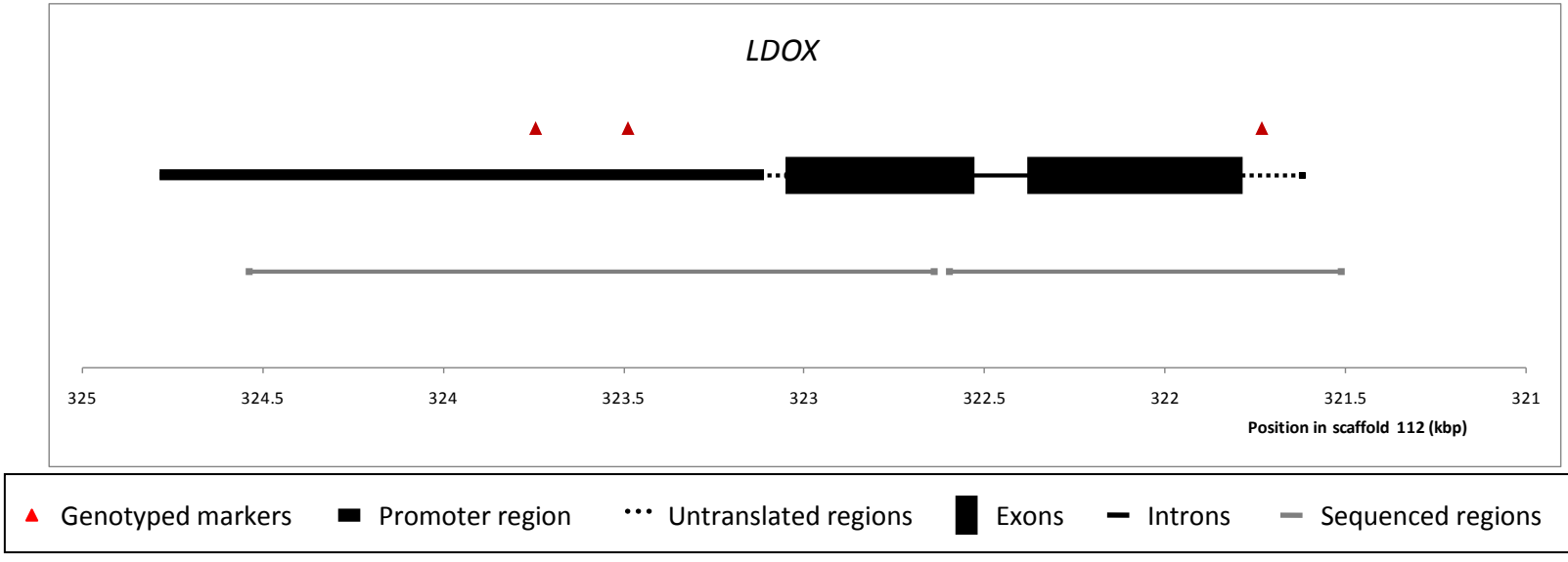
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



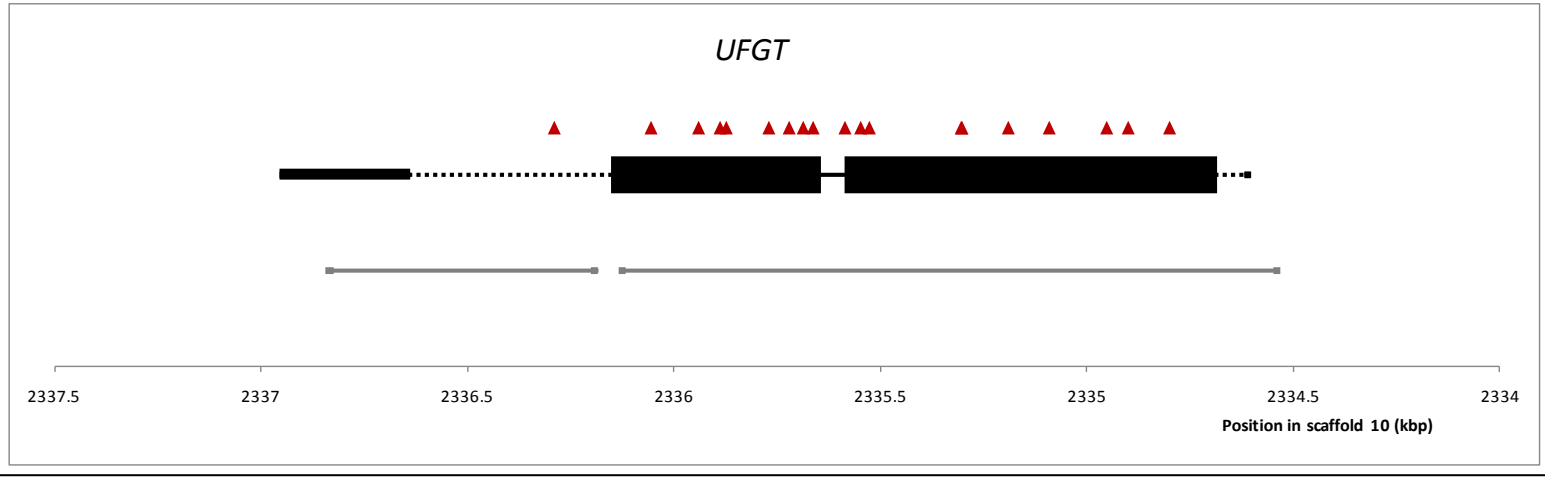
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



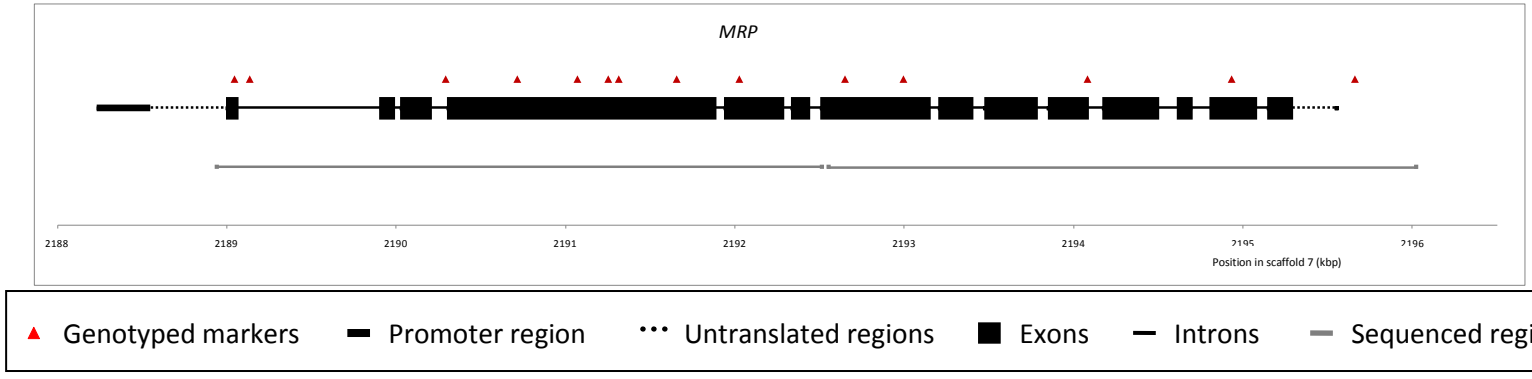
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



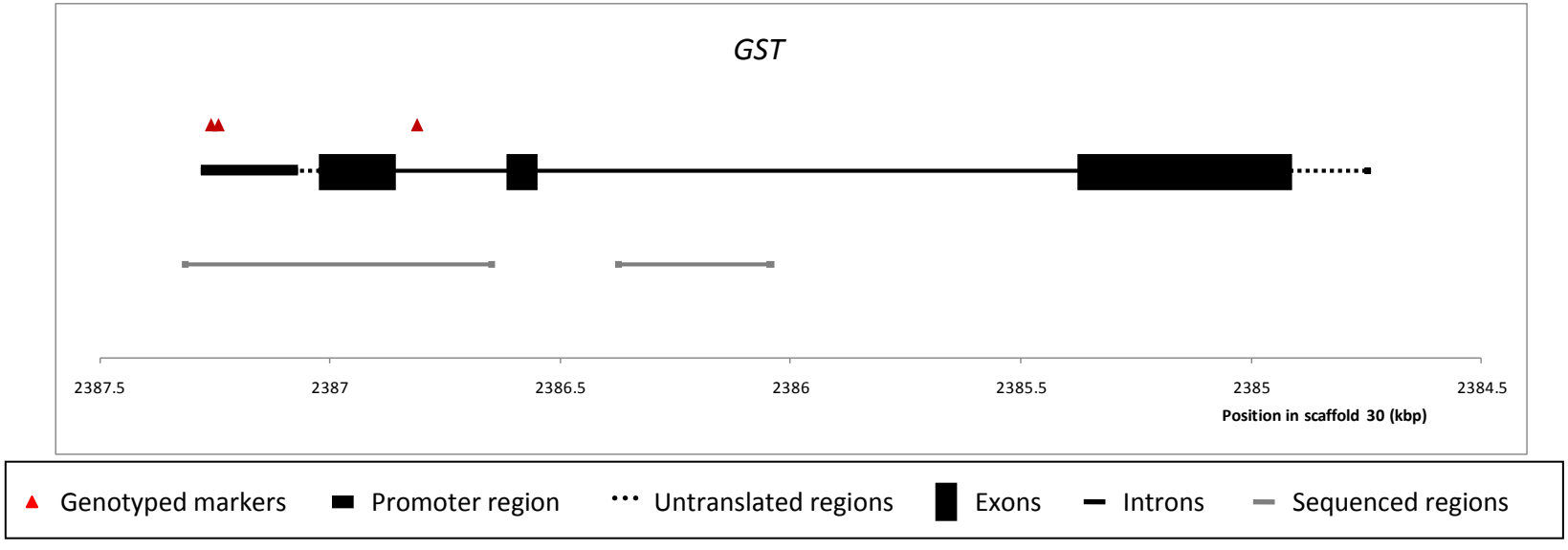
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



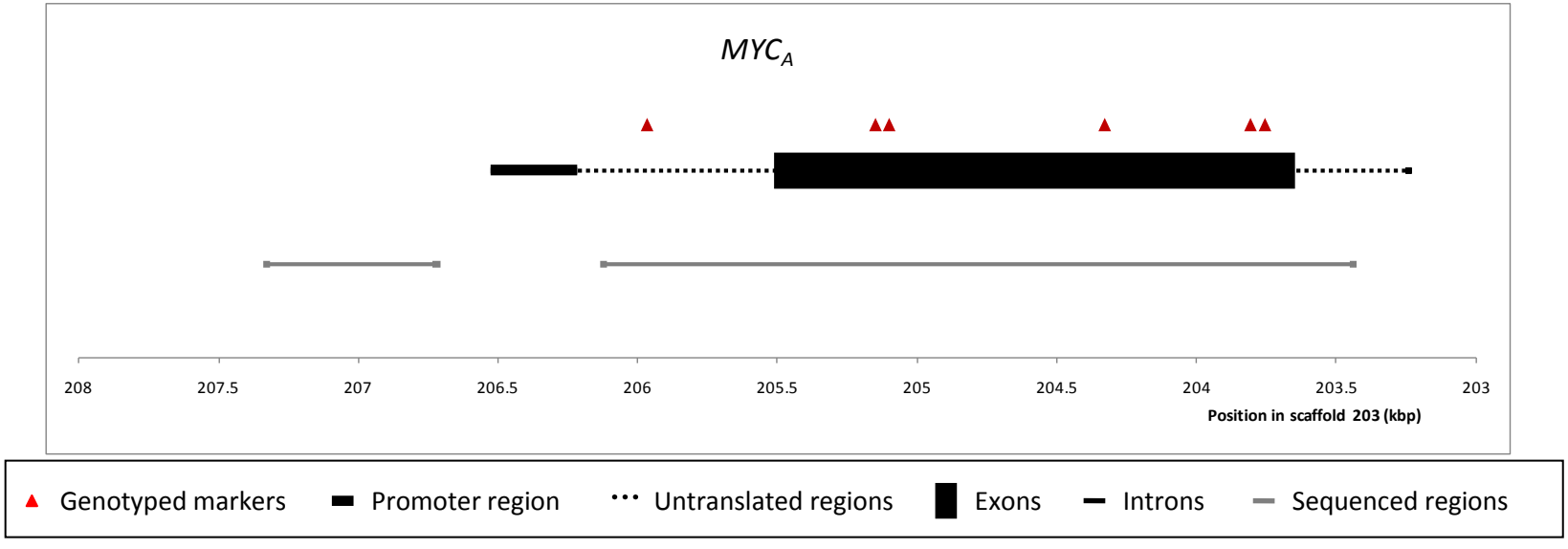
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



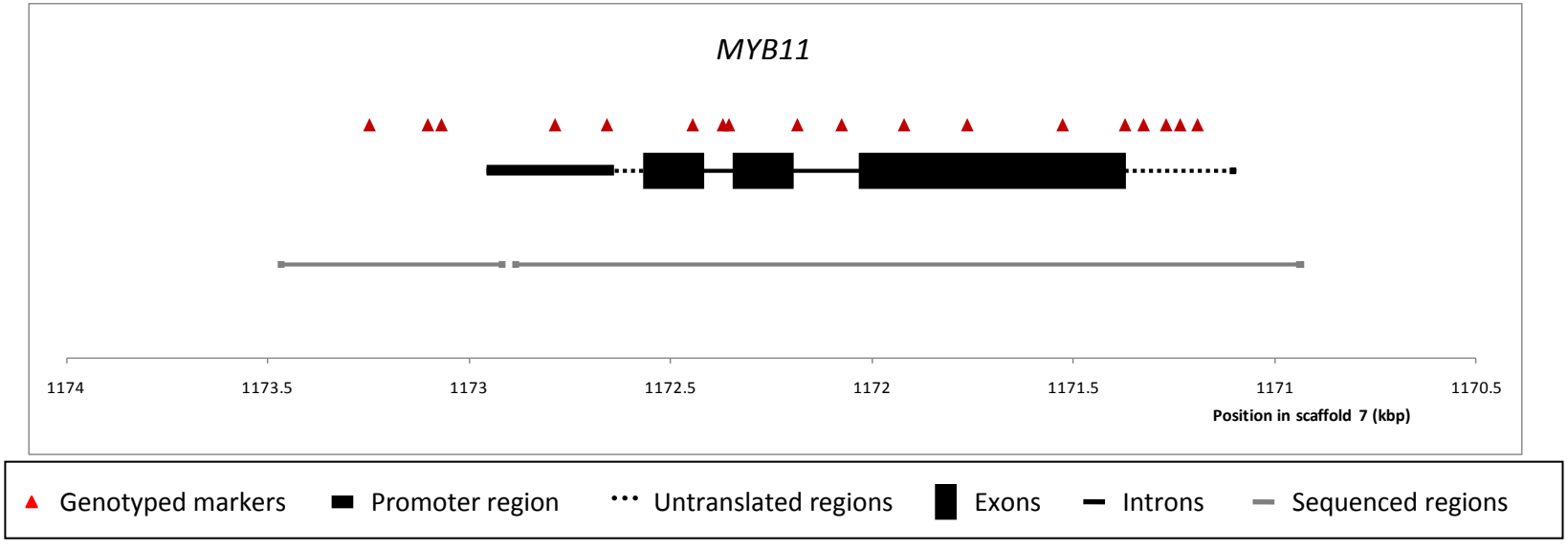
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



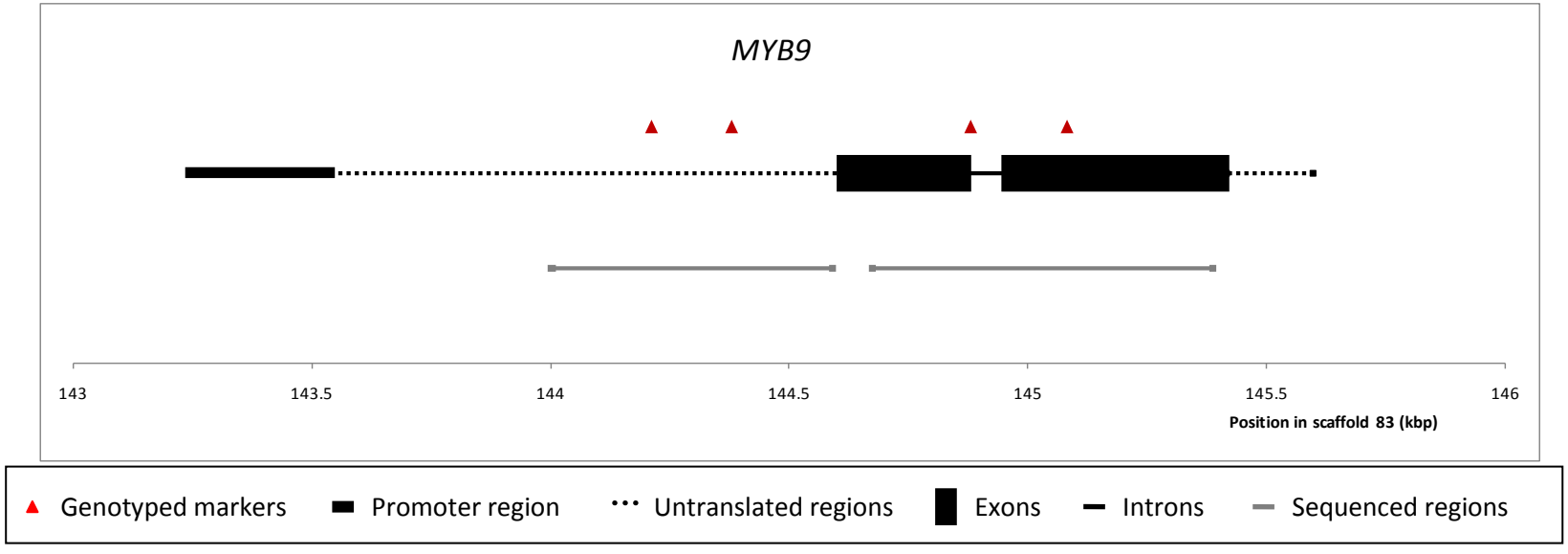
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



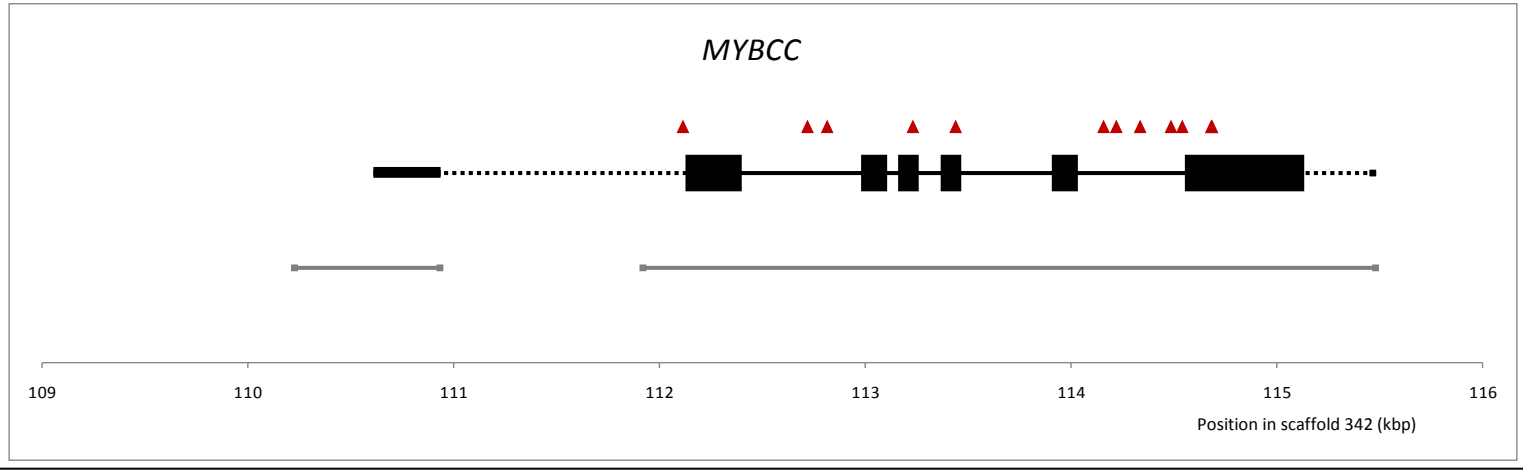
Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



Appendix 18 Schematic representation of the candidate genes and the genotyped SNPs (continuation).



Appendix 19 Pairwise LD values estimated between SNPs within each gene.

See attached CD, file “Appendix 19.xls”.

Appendix 20 Estimates of the proportion of each individual's variation that came from each subpopulation according to Pritchard's method (2000).

Cultivar ID	Q1	Q2
41305	0.088	0.912
41502	0.360	0.640
41503	0.970	0.030
41504	0.959	0.041
41508	0.627	0.373
41601	0.769	0.231
41607	0.684	0.316
41707	0.974	0.026
41806	0.522	0.478
50102	0.710	0.290
50103	0.953	0.047
50105	0.954	0.046
50106	0.528	0.472
50201	0.629	0.371
50202	0.931	0.069
50203	0.806	0.194
50204	0.960	0.040
50205	0.410	0.590
50207	0.481	0.519
50208	0.049	0.951
50301	0.978	0.022
50302	0.978	0.022
50303	0.903	0.097
50305	0.284	0.716
50601	0.943	0.057
50602	0.500	0.500
50603	0.042	0.958
50604	0.539	0.461
50606	0.790	0.210
50608	0.251	0.749
50615	0.977	0.023
50701	0.035	0.965
50702	0.122	0.878
50703	0.480	0.520
50705	0.514	0.486
50706	0.155	0.845
50707	0.943	0.057
50708	0.455	0.545
50801	0.024	0.976
50802	0.123	0.877
50803	0.235	0.765
50804	0.721	0.279
50806	0.630	0.370
50808	0.024	0.976
50901	0.873	0.127
50902	0.480	0.520
50903	0.232	0.768
50904	0.079	0.921
50905	0.130	0.870
50907	0.421	0.579
50908	0.668	0.332
51007	0.126	0.874
51008	0.511	0.489
51102	0.253	0.747
51103	0.540	0.460
51106	0.883	0.117
51107	0.919	0.081
51201	0.651	0.349
51203	0.179	0.821
51204	0.911	0.089
51205	0.963	0.037
51206	0.552	0.448
51208	0.405	0.595
51303	0.082	0.918
51304	0.039	0.961
51305	0.162	0.838
51307	0.096	0.904
51308	0.941	0.059
51402	0.335	0.665
51403	0.033	0.967
51404	0.133	0.867
51408	0.572	0.428
51501	0.400	0.600
51502	0.040	0.960
51513	0.043	0.957
51601	0.081	0.919
51602	0.032	0.968
51603	0.878	0.122
51604	0.066	0.934
51606	0.110	0.890
51607	0.397	0.603
51608	0.935	0.065
51701	0.036	0.964
51706	0.118	0.882
51708	0.040	0.960
51711	0.224	0.776

Cultivar ID	Q1	Q2
51802	0.037	0.963
51803	0.855	0.145
51804	0.415	0.585
51806	0.108	0.892
51901	0.030	0.970
52002	0.791	0.209
52004	0.953	0.047
52005	0.137	0.863
52006	0.359	0.641
52101	0.515	0.485
52104	0.084	0.916
52105	0.030	0.970
52106	0.114	0.886
52201	0.319	0.681
52202	0.534	0.466
52203	0.061	0.939
52204	0.028	0.972
52205	0.306	0.694
52206	0.069	0.931
52304	0.060	0.940
52306	0.095	0.905
52402	0.967	0.033
52502	0.121	0.879
52503	0.515	0.485
52505	0.968	0.032
52506	0.968	0.032
52604	0.449	0.551
52605	0.700	0.300
52607	0.696	0.304
52608	0.960	0.040
52702	0.752	0.248
52705	0.657	0.343
52708	0.081	0.919
52802	0.220	0.780
52807	0.046	0.954
52902	0.817	0.183
52904	0.033	0.967
52905	0.385	0.615
52908	0.049	0.951
53004	0.504	0.496
53102	0.978	0.022
53103	0.958	0.042
53107	0.980	0.020
53205	0.951	0.049
53206	0.972	0.028
53208	0.040	0.960
53303	0.109	0.891
53304	0.049	0.951
53305	0.057	0.943
53306	0.109	0.891
53307	0.101	0.899
53407	0.978	0.022
53505	0.055	0.945
53508	0.050	0.950
53605	0.136	0.864
53606	0.041	0.959
53608	0.033	0.967
53704	0.732	0.268
53705	0.033	0.967
53706	0.069	0.931
53803	0.321	0.679
53807	0.408	0.592
53904	0.067	0.933

Appendix 21 Pairwise relationship matrix based on Ritland Kinship Coefficient (RKC).

See attached CD, file “Appendix 21.xls”.

Appendix 22 Pairwise relationship matrix based on the Proportion of Shared Alleles (PSA).

See attached CD, file “Appendix 22.xls”.

Appendix 23 Association tests results for single SNP tests under Model A (*P*-values).

See attached CD, file “Appendix 23.xls”.

Appendix 24 Association tests results for single SNP tests under Model A (model parameter values).

See attached CD, file “Appendix 24.xls”.

Appendix 25 Association tests results for single SNP tests using log transformed phenotypic values under Model A (*P*-values).

See attached CD, file “Appendix 25.xls”.

Appendix 26 Association tests results for Single SNP tests under Model B (*P*-values).

See attached CD, file “Appendix 26.xls”.

Appendix 27 Association tests results for Single SNP tests using log transformed phenotypic values under Model B (*P*-values).

See attached CD, file “Appendix 27.xls”.

Appendix 28 Percentage of phenotypes showing significant associations (*P* < 0.01) for each SNP.

See attached CD, file “Appendix 28.xls”.

Appendix 29 Interactions *P*-values between SNPs in different genes.

Please see the attached CD, file “Appendix 29.xls”.

Appendix 30 Percentage of SNPs involved in significant interactions.

MYB11					
Second gene tested to interact with *MYB11*	**Number of significant interactions ($P < 0.001$)**	**SNPs within *MYB11* involved in significant interactions**		**SNPs within the second gene involved in significant interactions**	
		Number	**Percentage**	**Number**	**Percentage**
CHS_A	1	1	5.6	1	20.0
CHS_C	0	0	0.0	0	0.0
CHI	5	5	27.8	1	33.3
F3H	1	1	5.6	1	20.0
$F3'H_B$	0	0	0.0	0	0.0
DFR	0	0	0.0	0	0.0
LDOX	6	6	33.3	3	100.0
UFGT	1	1	5.6	1	5.3
MRP	10	4	22.2	6	42.9
GST	0	0	0.0	0	0.0
MYC_A	2	1	5.6	2	16.7
MYB9	1	1	5.6	1	10.0
MYBCC	21	13	72.2	5	41.7
MYC_B	21	18	100.0	3	30.0
MYB11	_	_	_	_	_

MYC_B					
Second gene tested to interact with *MYC_B*	**Number of significant interactions ($P < 0.001$)**	**SNPs within *MYC_B* involved in significant interactions**		**SNPs within the second gene involved in significant interactions**	
		Number	**Percentage**	**Number**	**Percentage**
CHS_A	5	1	10.0	5	100.0
CHS_C	7	1	10.0	7	100.0
CHI	3	1	10.0	3	100.0
F3H	5	1	10.0	5	100.0
$F3'H_B$	3	1	10.0	3	100.0
DFR	12	1	10.0	12	100.0
LDOX	7	4	40.0	3	100.0
UFGT	19	1	10.0	19	100.0
MRP	12	1	10.0	12	85.7
GST	3	1	10.0	3	50.0
MYC_A	5	1	10.0	5	41.7
MYB9	4	1	10.0	4	40.0
MYBCC	19	8	80.0	12	100.0
MYC_B	_	_	_	_	_
MYB11	_	_	_	_	_

MYBCC					
Second gene tested to interact with *MYBCC*	**Number of significant interactions ($P < 0.001$)**	**SNPs within *MYBCC* involved in significant interactions**		**SNPs within the second gene involved in significant interactions**	
		Number	**Percentage**	**Number**	**Percentage**
CHS_A	1	1	8.3	1	20.0
CHS_C	3	1	8.3	3	42.9
CHI	3	2	16.7	2	66.7
F3H	1	1	8.3	1	20.0
$F3'H_B$	2	1	8.3	2	66.7
DFR	10	1	8.3	10	83.3
LDOX	2	1	8.3	2	66.7
UFGT	8	1	8.3	8	42.1
MRP	13	2	16.7	12	85.7
GST	2	1	8.3	2	33.3
MYC_A	3	1	8.3	3	25.0
MYB9	3	2	16.7	2	20.0
MYBCC	–	–	–	–	–
MYC_B	–	–	–	–	–
MYB11	–	–	–	–	–