Università degli Studi di Padova

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE

INDIRIZZO: Biotecnologie

CICLO XXVIII

Genome-wide patterns of genetic variation among wild and cultivated grapevines (*V. vinifera* L.)

Direttore della Scuola : Ch.mo Prof. *Paolo Bernardi*

Coordinatore d'indirizzo: Ch.ma Prof.ssa *Fiorella Lo Schiavo*

Supervisore : Ch.mo Prof. *Giorgio Valle*

Correlatore : Dott.ssa *Maria Stella Grando*

Dottorando : *Annarita Marrano*

CONTENTS

Abstract

Grapevine (*V. vinifera* L.) is one of the most important crops worldwide due to its global distribution and economic value. Two forms of grapevine still co-exist nowadays: the cultivated form *V. vinifera* subsp. *sativa* and the wild form *V. vinifera* subsp. *sylvestris*, which is considered the ancestor of present cultivars. Archeological and historical findings suggest that cultivated grapevines have been domesticated from wild populations of *V. sylvestris* circa 8,000 BP in the Near East. However, recent genetic analyses raised the outstanding question whether multiple domestication events occurred. During domestication the biology of grapes changed dramatically to guarantee greater yield, higher sugar content and more regular production. The changes in berry and bunch size as well as the transition from dioecious wild plants to hermaphrodite cultivated grapes were crucial. Additional studies on the genetic relationship between wild and cultivated grapevines are required in order to understand how this phenotypic evolution occurred and to clarify the process of adaptation to domestication in grapevine. This will be useful for the future genetic improvement of viticulture.

In this regard, we investigated the genetic and phenotypic variation within a germplasm collection of wild and cultivated grapevine accessions. The whole population was first genotyped with the commercial GrapeReSeq Illumina 20K SNP array, yielding 16K good quality single nucleotide polymorphisms (SNPs). Afterwards, a novel Restriction Associated DNA-sequencing (RADseq) procedure was developed in order to further increase the density of molecular markers across the grapevine genome. By applying this novel RAD-seq protocol to the whole population, 37K SNPs were identified, which reflected a considerable level of genetic diversity between *sativa* and *sylvestris* accessions. The two merged SNP matrices were filtered for SNP loci with a missing rate > 0.2 and a minor allele frequency (MAF) < 0.05. The final panel of 27K SNPs evenly distributed along the grapevine genome was used to investigate the population structure by using both Principal Component Analysis (PCA) and the cluster algorithm implemented in fastSTRUCTURE software. In line with previous research, both analyses highlighted a low but clear differentiation between *sativa* and *sylvestris* individuals. Therefore, the extent of Linkage Disequilibrium (LD) was evaluated within the whole grapevine population and in the two subspecies separately. LD, as measured by the classical r^2 correlation coefficient, decayed below 0.2 within 10 kb in the whole population. On the other hand, a slower LD decay was observed in the wild compartment, where r^2 reached values below 0.2 within 20 kb. This result can be related with an elevated level of inbreeding among wild individuals, linked to a small effective population size and the missing gene-flow between wild populations.

Population differentiation statistic (F_{ST}) was computed across the grapevine genomes looking for genomic regions with divergent allele frequencies between the two grapevine subspecies. An overall low level of genetic differentiation (F_{ST} = 0.12) was observed between cultivated and wild grapes, suggesting the occurrence of genetic exchange among the two subspecies. However, a non-random distribution of divergent sites was observed along the whole genome: over two thousands of SNP loci revealed a significant level of differentiation between *sativa* and *sylvestris*, validated empirically with a permutation test. 1,714 annotated genes were found in LD with these most significant SNPs, and showed an enrichment of predicted functions

related to the metabolic processes of nitrogen and carbohydrate as well as to the perception and adaptation to environmental stimuli. A slightly reduction of nucleotide diversity in the *sylvestris* ($\pi_{sylvestris}/ \pi_{sativa}$ ~0.95) was observed in almost all the identified genes involved in stress responses, suggesting that a selection is likely acting in wild populations for adaptation to several environmental changes. Therefore, these results point the attention towards *sylvestris* grapevines as valuable resources of resilience genes or alleles, which may have been lost in cultivated grapevine during the domestication process.

Genome-wide association study (GWAS) approach has been applied as an alternative strategy to identify the genes and mutations that have been targets of selection during crop domestication. Therefore, the germplasm collection of cultivated and wild grapevines has been evaluated in two years for single berry and single bunch weight, number of bunches per plant, yield and berry composition (sugar, organic acid and K^+ concentrations, titratable acidity and pH). A great phenotypic variation was observed within and between the two grapevine subspecies, notably for berry size, pH, acid contents and titratable acidity. The association test, carried out accounting for confounding factors, identified significant genotype-phenotype correlations for all traits, except for single berry weight. Genes encoding proteins related to Ca^{2+} sequestration and signalling, transcription factors and enzymes involved in the metabolism of polyamines were identified in linkage with the SNPs significantly associated to yield and bunch weight. At the same time, genes with a central role in the control of berry flesh pH and acidity were detected, such as the isocitrate lyase and V-type proton ATPase subunit a3 genes.

Therefore, the present research has proven for the first time the feasibility of population genetics and association mapping approaches for dissecting the genomic basis of phenotypic variation in a complex genetic system as grapevine. Moreover, further evidence of the relevance of wild grapevine as a model for understanding the mechanisms of adaptation to natural conditions has been provided. These results pave the way for understanding how wild and cultivated grapevines react to environmental stimuli, which will benefit the development of new breeding strategies to face the ongoing climate changes and the growing demand of a sustainable viticulture.

Riassunto

La diffusione geografica e l'importanza economica della viticoltura fanno della vite euroasiatica (*V. vinifera* L.) una delle specie più importanti per l'agricoltura mondiale. La maggior parte dei vitigni coltivati appartengono alla sottospecie *V. vinifera* subsp. *sativa*, la quale si ritiene sia stata domesticata nel vicino Oriente dalla vite selvatica (*V. vinifera* subsp. *sylvestris*) intorno al IV millenio a.C. Tuttavia, studi recenti hanno sollevato l'ipotesi di eventi di domesticazione secondaria della vite coltivata in Europa occidentale. Si pensa che il passaggio da viti selvatiche dioiche a viti con fiori ermafroditi sia stato fondamentale per la domesticazione della vite, dal momento che la capacità di produrre frutti per autofecondazione garantiva una produttività superiore e costante di uva. Altrettanto importante è stata la selezione per caratteristiche dell'uva di immediata percezione, come per esempio la dimensione della bacca ed il suo contenuto zuccherino. Studi aggiuntivi sulle relazioni genetiche tra la vite coltivata e la sua forma spontanea sono necessari allo scopo di chiarire la serie di incertezze che ancora persistono sull'origine della vite domestica ed incentivare il miglioramento genetico della viticoltura attuale.

Pertanto, il principale obiettivo del presente lavoro di tesi è stato la caratterizzazione della variabilità fenotipica e genetica di una collezione di viti coltivate e selvatiche. L'intera popolazione è stata genotipizzata con il nuovo GrapeReSeq 20K SNP chip, ottenendo una matrice finale di 16 mila marcatori SNP di alta qualità. Allo stesso tempo, un nuovo protocollo della tecnologia RAD-seq è stato messo a punto con lo scopo di incrementare la densità dei marcatori molecolari lungo il genoma di vite. In seguito all'applicazione di questa nuova procedura di RAD-seq all'intera collezione di viti, circa 37 mila marcatori SNP sono stati identificati, mettendo in evidenza una cospicua diversità genetica tra la vite coltivata ed il suo presunto progenitore. L'unione delle due matrici di marcatori SNP, seguita dalla rimozione dei loci con un tasso di dati mancanti superiore a 0.2 ed una frequenza dell'allele minore (MAF) inferiore a 0.05, ha portato alla formazione di un panel definitivo di circa 27 mila marcatori SNP, equamente distribuiti lungo il genoma di vite. Questo panel finale di marcatori SNP è stato utilizzato per analizzare la struttura della popolazione attraverso due approcci complementari, ossia l'analisi delle componenti principali (PCA) e l'approccio bayesiano implementato nel programma fastSTRUCTURE. In accordo con quanto riportato in letteratura, entrambe le strategie hanno messo in evidenza una chiara e moderata differenziazione tra le accessioni di *V. sativa* e *V. sylvestris*. Pertanto, l'estensione del Linkage Disequilibrium (LD), espresso sottoforma del classico coefficiente di correlazione r^2, è stata valutata nell'intera collezione e nei due sottogruppi separatamente. Il valore di r^2 è risultato inferiore ad una soglia di 0.2 dopo circa 10 kb nel germoplasma completo e dopo 20 kb nella sottopopolazione delle viti selvatiche. Questa discrepanza di valori di LD nelle viti spontanee può essere legata alla ridotta dimensione della popolazione effettiva ovvero alla mancanza di scambio di materiale genetico (gene-flow) tra popolazioni diverse di *V. sylvestris*.

In seguito, la differenziazione genetica tra le viti coltivate e selvatiche lungo il genoma è stata misurata sottoforma di indice di fissazione (F_{ST}) per individuare regioni genomiche con frequenze alleliche divergenti tra le due sottospecie. Il valore medio di F_{ST} pari a 0.12 ha suggerito una moderata differenziazione genetica tra le accessioni di *sativa* e *sylvestris*, indicando come tra di esse si verifichino frequenti eventi di ibridazione. Tuttavia, circa 2 mila marcatori SNP hanno mostrato un elevato livello di differenziazione tra le viti coltivate e selvatiche ($F_{ST} > 0.27$), come confermato dal test di permutazione. 1,714 geni annotati sono stati identificati in linkage con i

suddetti marcatori SNP, mostrando un significativo arricchimento in funzioni geniche predette legate al metabolismo dell'azoto e dei carboidrati, e ai meccanismi di risposta ed adattamento agli stimoli ambientali. Una lieve riduzione della diversità nucleotidica della vite selvatica ($\pi_{sylvestris}/\pi_{sativa}$ ~0.95) è stata osservata nella maggior parte delle suddette regioni geniche con un ruolo nella risposta a stress biotici ed abiotici. Pertanto, una pressione selettiva sta probabilmente operando nelle popolazioni di *V. sylvestris* per l'adattamento ai sempre più frequenti cambiamenti climatici. Questo risultato sottolinea l'importanza della vite selvatica come putativa fonte di geni e/o alleli di resilienza, i quali potrebbero essere stati persi dalla vite coltivata durante il processo di domesticazione.

L'approccio di genome-wide association study (GWAS) è stato, in seguito, applicato come strategia alternativa per l'identificazione dei geni e delle mutazioni selezionati durante la domestizatione della vite. Pertanto, l'intera collezione di viti coltivate e selvatiche è stata fenotipizzata per il peso della bacca e del grappolo, il numero di grappoli per pianta, la produttività, e la composizione chimica della bacca (contenuto in zuccheri, acidi organici e potassio, acidità titolabile e pH). Un elevata variabilità fenotipica è stata osservata tra e all'interno dei due sottogruppi di vite, soprattutto per i caratteri peso della bacca, pH, contenuto in acidi organici e acidità titolabile. Il test di associazione, corretto per la struttura della popolazione e le relazioni di parentela, ha identificato correlazioni significative marcatore-carattere per tutti i fenotipi studiati, ad eccezione del peso della bacca. Geni codificanti per fattori di trascrizione e per proteine coinvolte nel metabolismo del calcio e delle poliammine sono stati identificati in linkage con i marcatori SNP significativamente associati ai caratteri produttività e peso del grappolo. Inoltre, il test di associazione ha consentito l'identificazione di geni coinvolti nel controllo del pH e dell'acidità totale della bacca, come per esempio i geni codificanti per la subunità A3 della pompa protonica vacuolare ovvero per l'isocitrato liasi.

In conclusione, il presente lavoro di ricerca ha dimostrato per la prima volta come la genetica di popolazione e l' association mapping siano due validi approcci per individuare le basi genetiche della variabilità fenotipica osservata in un sistema genetico complesso come la vite. Inoltre, sono state fornite evidenze dell'importanza della vite selvatica come modello per lo studio dei meccanismi di adattamento agli stress ambientali. Questi risultati rappresentano la base per comprendere come le viti selvatiche e coltivate reagiscano agli stimoli ambientali, nell'ottica di sviluppare nuovi programmi di miglioramento genetico della vite ed affrontare gli attuali cambiamenti climatici e la crescente richiesta di una viticoltura sostenibile.

Chapter 1

INTRODUCTION

Grapevine (*Vitis vinifera* L.) is one of the most economically important fruit crop in the world, growing mainly in climates with warm dry summers and cool wet winters [1]. Grapes are widely used as fresh (table grapes) or dried (raisins) fruits as well as for wine, juice and spirits production. In addition, recent trends have also focused on antioxidants and healthful products derived from grapes [2]. According to the Organisation International de la Vigne et du Vin (OIV, 2015), 7,5 million hectares are cultivated worldwide with grapevine, yielding 73,7 million tons of grapes in 2014 [3]. 41% of total world grape is produced in Europe, with France, Italy and Spain as the leading countries, followed by Asia (29%) and America (21%). Out of total grape production 55% is used for wine-making, 35% as table grape, 8% for raisin production and the remaining 2% for other products. Due to the global wine exports in volume increased to 104 million hectoliters along with a value of 26 billion Euros in 2014 [4], high priority must be given to grapevine breeding in order to improve economically important traits, such as yield and berry composition, in view of a "sustainable viticulture".

1.1. Taxonomy and origin of the grapevines

1.1.1. The family of *Vitaceae*

Grapevine is a member of the *Vitaceae* family which consists of perennial plants distributed in temperate and inter-tropical climates as woody or herbaceous climbers or rarely shrubs [5]. About 900 species from 15 genera are documented in the *Vitaceae* family, from which only the genus *Vitis* produces edible fruits [6]. Molecular phylogenenies based on the complete plastid genome of grapevine place the *Vitaceae* into the earliest diverging lineage of rosids [7]. Moreover, several plastid (rbcL, trnL-F intron and spacer, atpB-rbcL spacer, rps16, trnC-petN spacer; [8][9][10]) and nuclear (ribosomal ITS, GAI1; [11][12]) genes have been used for resolving the *Vitaceae* phylogeny, identifying five major clades in the family: the *Ampelocissus-Vitis-Nothocissus-Pterisanthes* clade, the *Parthenocissus-Yua* clade, the core *Cissus* clade, the *Cayratia-Cyphostemma-Tetrastigma* (CCT) clade and the *Ampelopsis-Rhoicissus-Clematicissus* clade [13]. The family ancestor may belong to the *Cissus* genus, which is typically inter-tropical and possess 4-merous flowers and a basic chromosome karyotype of 2n = 24 [10]. On the other hand, the genera *Ampelocissus*, *Vitis*, *Ampelopsis* and *Parthenocissus* consist of plants with 5-merous flowers that are characterized by a karyotype of 2n = 40, except the subgenus *Vitis* (2n = 38) [1]. Recently, Wen et al. [14] used 417 single-copy nuclear genes from the transcriptomes of 15 *Vitaceae* species, and the grapevine reference genome [15] to reconstruct the deep phylogeny of the grape family, showing how the *Ampelopsis-Rhoicissus* clade is the earliest divergent lineage, while the *Vitis-Ampelocissus* and *Parthenocissus-Yua* clades are sister groups. In addition, this analysis revealed the close relationship between the CCT and *Cissus* clades, suggesting a single origin of 4-merous taxa in the grape family. This topology was further confirmed by using both full plastome and mitochondrial genes sequences of 27 *Vitaceae* species [16], indicating that the grape family did not exhibit significant reticultation at deep level.

1.1.2. Origin and diversification of the genus Vitis

The genus *Vitis* is composed of two subgenera: *Muscadinia* Planch. (2n = 40) and *Vitis* Planch. (2n = 38). *Muscadinia* subgenus is represented by only three species, *V. rotundifolia*, *V. munsoniana* and *V. popenoei*, mainly distributed across the southeast of USA and Mexico. The *Muscadinia* could be considered as a relictual monospecific subgenus (or genus) that could make the transition between the two sister clades *Vitis* and *Ampelocissus* [10]. On the other hand, the subgenus *Vitis* consists of ~60 species, among which the cultivated taxa *V. vinifera*. These species have been found mainly in the temperate zones of the northern hemisphere from North America to eastern Asia, except for some subtropical species (*V. caribeae*, *V. lanata*). The two subgenera are reproductively isolated, while the species within subgenus *Vitis* are interfertile [17]. All species are dioecious except for *V.vinifera* L., which has hermaphroditic flowers, and *V.rotundifolia*, which segregates for this trait [17]. Although only *V. vinifera* is cultivated for human consumption, the *Vitis* wild species are of great economic importance since they are used as rootstocks (*V. riparia*, *V. rupestri*, *V. berlandieri*) for the highly susceptible *V. vinifera* and represent a gene pool for resistance to biotic and abiotic stresses [1]. Studing the evolutionary relationships within the genus *Vitis* is complicated due to the numerous synonyms, which likely arose from the lack of agreement between systematic botanists on what can be considered a species, and because of the broad morphological variation within species [18]. Indeed, the systematics of *Vitis* have relied on morphology for a long time [19] and just recently molecular methods have been introduced to resolve this taxonomic controversy. A phylogenetic analysis with three plastid DNA regions of 48 accessions, including 30 *Vitis* species and several *V. vinifera* cultivars, supported that the genus *Vitis* is monophyletic [20]. In addition, three clades have been identified within the *Vitis* genus, reflecting the geographic distribution of *Vitis* species: Europe, Asia and North America. In particular, while the Asian clade presented high genetic diversity, low genetic variability was observed in the European and North American clades, suggesting hybridizations between cultivated grapevine and autochthonous accessions [20]. However, the use of plastid markers did not allow the assessment of hybridizations between the analyzed species. Recently, Wan et al. [17] examined 309 accessions from 48 *Vitis* species, varieties and outgroups, with 27 unliked nuclear genes. By estimating the divergence time, they showed how the splitting events between the deeper clades occurred almost simultaneously within the subgenus *Vitis*. This results was in agreement with the high degree of shared polymorphisms between North America wild grapevine species and European cultivated species observed by Myles et al. [21]. Moreover, they confirmed the origin of *Vitis* during the Paleogene in North America, followed by a progression to Asia to Europe [22]. In particular, the oldest age of *Vitis* was assigned to the Paleocene (65.5-58.8 Ma), during which Laurasia has only begun dividing into North America and Eurasia, and the climate was considerably warmer in the northern latitudes [23]. During the Pliocene and Pleistocene cooling cycles, fragmentation and isolation of some North America and Asian species occurred leading to the primary divisions within *Vitis*. After the glacial period, these species must have expanded and adapted ecologically to their large present range, acquiring a remarkable diversity in morphological characters. This diversity has been maintained by barriers of geographical, ecological or phenological nature. Therefore, *Vitis* was part of the great biogeographic phenomenon of range restrictions, survival in refugia, and diversifications, caused in many groups of organisms by the Quaternary ice ages [24].

1.1.3. The domestication of *V. vinifera*

The Eurasian grape (*Vitis vinifera* L.) exists nowadays as two forms in Eurasia and in North Africa: the cultivated form *V. vinifera* subsp. *sativa* (o *vinifera*), and the wild form *V. vinifera* subsp. *sylvestris*, which is considered the ancestor of present cultivars [25]. The wild-type and cultivated forms are sometimes referred as two separated subspecies based on morphological differences [26], even if this distinction can be debated since these differences are most likely the results of domestication by humans rather than geographical isolation. The grapevine domestication has been linked to the discovery of wine, although it is unclear which process came first [27]. During domestication the changes in berry and bunch size and from dioecious wild plants to hermaphrodite cultivated grapes were crucial. In addition, the biology of grapes changed dramatically to ensure greater yield, higher sugar content for better fermentation and more regular production. Uncertainty still remains about whether this changes occurred through sexual crosses and natural or human selection, or via mutation, selection and subsequent vegetative propagation [25]. Major questions about grapevine domestication concern the number of domestication events and their geographic locations [28]. Two opposite hypotheses have been formulated so far: (i) a restricted origin hypothesis in which domestication took place in a single location from a limited wild stock, with subsequent spreading to other regions [29]; (ii) a multiple-origin hypothesis in which domestication occurred along the entire distribution range of wild progenitor species, involving a large number of founders [30]. According to the first hypothesis, grapevine has been domesticated in the Near East region, stretching from the western Himalaya to the Caucasus, during the second half of the 4th millennium B.C. [31][32]. From the primo-domestication center, there was a gradual dispersal to adjacent regions such as Egypt and Lower Mesopotamia, and then further spread around the Mediterranean following the main civilizations (Assyrians, Phoenicians, Greeks, Romans) [27]. In particular, during the second half of the 2nd millennium domesticated grapevines made their first appearance in the Southern Italy and later in Northern Italy, Southern France, Spain and Portugal [33]. By the end of the Roman Empire, grape growing was common in most of Europe. Furthermore, the Romans were the first to assign names to cultivars, even if it is difficult to correlate them with modern varieties [25]. *V. vinifera* was introduced in America by the missionaries during the 16th century and in South Africa, Australia and New Zeland in the 19th century. A recent assessment of the genetic diversity within 950 *sativa* and 59 *sylvestris* genotypes with 5,387 SNPs provided further evidence of the origin of sativa in the Near East [34].

However, other studies on the genetic relationship between wild and cultivated grapevines have provided novel evidences supporting the multiple-origin hypothesis. Indeed, Grassi et al. [33] applied six microsatellite (SSR) loci to study the origin of some Italian cultivated grapevines from in situ direct domestication of the wild autochthonous grapevine, suggesting a second domestication event in the Sardinia island. Accordingly, Arroyo-García et al. [28] analyzed with nine chloroplast SSR loci 1,201 individual grapevine genotypes, including 513 *sativa* and 688 *sylvestris* accessions from the whole area of the grapevine distribution. They identified eight different chlorotypes, of which only four had a global frequencies greater than 5%. A similar geographic distribution of chlorotypes was observed between the *sylvestris* and *sativa* groups, suggesting the existence of at least two origins of the modern grapevine cultivars: (i) an eastern origin related to the *sylvestris* population groups located in Near and Middle East; (ii) a western

origin related to *sylvestris* individuals from Iberian Penisula, Central Europe and Northern Africa (Figure 1).

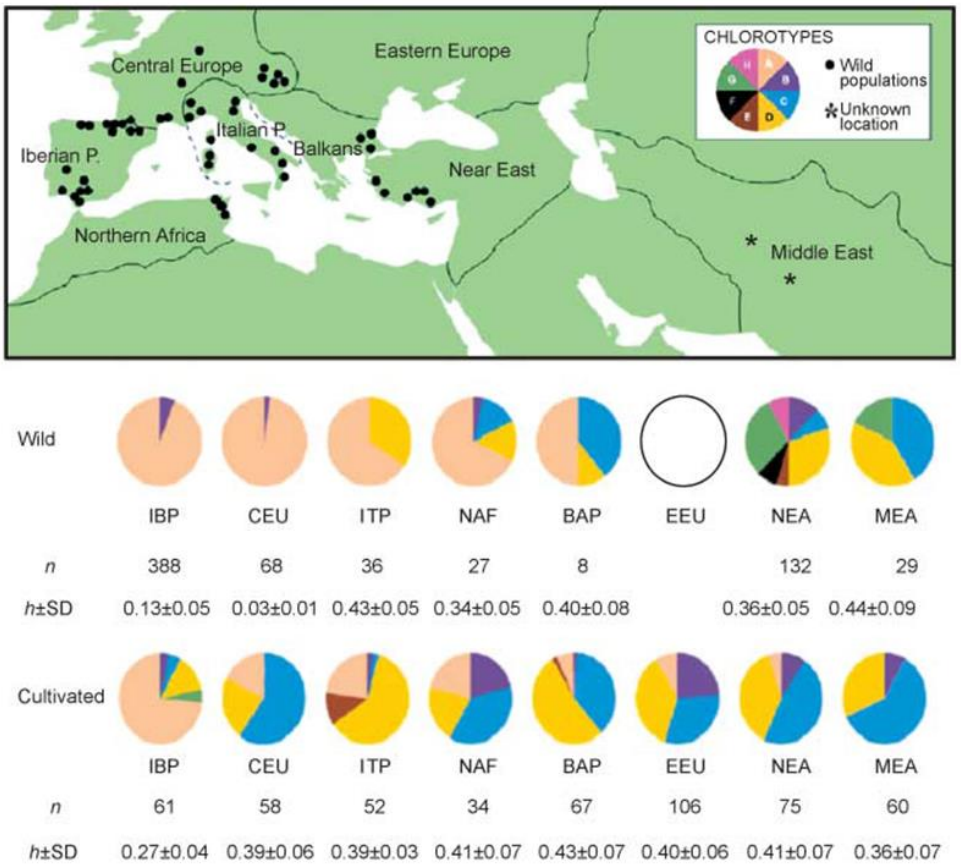

Figure 1: Chlorotype distribution in *sylvestris* and *sativa* population groups. Geographic areas considered are separated by lines when needed. Asterisks indicate that specific locations of collection in the area are unknown. From west to east: Iberian Peninsula (IBP), Central Europe, (CEU), Northern Africa (NAF), Italian Peninsula (ITP), Balkan Peninsula (BAP), Eastern Europe (EEU), Near East (NEA) and Middle East (MEA). The figure also shows the values of unbiased chlorotype diversity and the number of genotypes considered within each population group. (Source: Arroyo- García et al. [28]).

Furthermore, the genetic analysis of a Israeli grapevine population of *sativa* and *sylvestris* genotypes against European and Asian grapevine datasets with 22 SSRs revealed how a large bulk of Israeli *sylvestris* and *sativa* populations are genetically proximal, supporting an autonomous domestication in Israel [35]. This result was further strengthened by the full genomic sequencing of nine Israeli grapevine individuals, including for the first time 3 *sylvestris* accessions [36]. The genome-wide comparison of these genomic sequences with the SNP profiles gained with the Vitis18kSNP array for Georgian and European populations confirmed the close genetic homology between Israeli *sativa* and *sylvestris* accessions.

These studies regarding the genetic relationship between cultivated and wild grapevines represent a step towards the elucidation of the grapevine domestication process. However, several doubts are left on how, where and when the cultivated grapevine arose from its wild relative. The huge progress made in plant genetics and genomics represents a great opportunity to better understand the domestication process of *V. vinifera*.

1.2. Phenotypic and genetic diversity of *V. vinifera*

1.2.1. *V. vinifera* subsp. *sativa*

Substantial phenotypic and genetic diversity has been maintained in the cultivated grapevine, whose number of cultivars available today is estimated from 6,000 to 11,000 [37]. This considerable variation of cultivated grapevines is the results of three main processes during the long history of viticulture: sexual reproduction, vegetative propagation and somatic mutations [25]. Indeed, since the high heterozygosity of grapevine genotypes, the sexual crosses produce any progeny with a novel combination of parental alleles resulting in phenotypic variation. However, due to the long juvenile period of grapevine plants, vegetative propagation is a common agronomical practise in viticulture to preserve and multiple highly desirable genotypes. In addition, cuttings are a convenient method of moving cultivars from one region to another. During this long process of vegetative propagation, somatic mutations may occur leading to morphological and agronomical differences. It is thought that the appearence of hermaphrodite flowers, which was crucial during the grapevine domestication, resulted from a mutation [25]. Moreover, a putative causal SNP responsible for the substitution of a lysine with an asparagine at position 284 of the 1-deoxy-D-xylulose 5-phosphate synthase (VvDXS) seems to be involved in muscat flavor in grapevine [38]. In this regard, transposon and retrotransposon based mutations have played a central role in promoting phenotypic variation in grapevine [39]. For instance, it has been shown that the insertion of a gypsy-type retroelement (Gret1) in the promoter region of a regulatory gene of the Myb family causes the loss of black berry colour in homozygous individuals [40]. In addition, insertion of a haT transposable element in the promoter of the TFL1A gene was shown to cause an early phenotypic alteration affecting cluster ramification and development, delay in flower meristem specification as well as both flower and flower organ reiterations [41]. Several efforts have been devoted to explore and characterize the phenotypic variation of *V. sativa*, notably for traits of interest such as berry weight [42] and composition [43][44], bunch weight [1], leaf shape [45], fertility and phenology [46]. However, the analysis of large sets of genetic resources at the morphological level are still missing because of the complexity of the methods available so far or the fact that phenotyping grape is expensive, time consuming and requires a lot of space [1].

The morphological and agronomical differences of cultivated grapevines could also arise from the adaptation to different ecological conditions across the whole geographical distribution of grapevine during the long history of viticulture. Indeed, Negrul [47] classified the *V. vinifera* cultivars into three large eco-geographical groups, called proles, based on morphological similarities. The wine grape varieties with small berries widespread in western Europe (France, Spain, Germany, Italy and Portugal) were included in the proles occidentalis, whereas the table grapes varieties with large berries, mainly cultivated in the wide area extending from Central Asia to Near East, were placed in the proles orientalis. In particular, Negrul recognised two sub-proles within the proles orientalis: (i) caspica, composed of ancient vines used for vinification before the advent of Islam (from AD 500-1100), and (ii) the antasiatica, including cultivars for table grape and raisins of more recent origin [48]. Finally, the proles pontica, probably the most ancient group, was identified by wine grape varieties cultivated around the Black Sea and in eastern Europe. Furthermore, varietal ecotypes found from Georgia to the Balkans were designated as proles

pontica sub-proles georgica and sub-proles balkanica, respectively. The Negrul's classification of grapevine cultivars has been confirmed by recent studies of genetic diversity within grapevine germplasm collections by using molecular markers. Emanuelli et al. [49] investigated the patterns of molecular diversity at 22 SSRs and 384 SNPs in 2,273 accessions of *V. sativa*, its wild relative *V. sylvestris*, interspesific hybrid cultivars and rootstocks. Out of the 1,085 non-redundant genotypes 733 were *sativa* accessions, which revealed a deep population stratification in four groups. The first cluster (vv1) represented maily Italian/Balkan wine grapes, which resemble the proles pontica, whereas the second groups (vv2) was more heterogeneous including both table grape varieties related to 'cv Sultanina' (proles orientalis sub-proles antasiatica) and some Spanish wine grapes with unknown origin. The Muscat table and wine cultivars (vv3) belong to proles orientalis sub-proles caspica, while the French and German wine cultivars (vv4) were part of the group occidentalis. This broad genetic variability allowed the construction of core collections to maximize the allelic diversity among the *sativa* accessions and make it easily accessible for future studies of gene mapping and functional genomics. Further evidence of how the genetic structure in cultivated grapevines is linked to geography and human selection was provided by Bacilieri et al. [50], which analyzed a dataset of 2,096 cultivated genotypes by 20 SSR loci. Three main genetic groups were identified: a) wine cultivars from western regions (proles occidentalis), b) wine varieties from Balkans, and East Europe (proles orientalis), and c) a group mainly composed of table grape cultivars from Eastern Mediterranean, Caucasus, Middle and Far East countries (proles pontica).

However, the extent of morphological and genetic diversity found today among cultivated grapevines might be a narrow reflection of what existed before the introduction of disease-causing agents (*Phylloxera*) from America at the end of the 19th century. Moreover, during the last 50 years the globalization of wine companies and markets caused further reduction of diversity, because of the emergence of a few popular grapevine cutivars, such as Chardonnay, Cabernet Sauvignon, Syrah, and Merlot [51]. Due to the constant evolution of disease-related agents and climate conditions, the exploration of new cutivated and wild genetic resources is required to design novel breeding programs. In this regard, several efforts have been recently devoted to investigate the genetic diversity within focal regions of grapevine development. For instance, Marrano et al. [52] (Appendix A, page 120) reported the first assessment of genetic diversity, relationships and structure of 80 grapevine cultivars and 21 *V. sylvestris* accessions originated from the regions of Uzbekistan, Tajikistan and Kyrgyzstan, revealing a significant amount of genetic variation. Similarly Basheer-Salimia et al. [53] characterized 43 putative cultivars grown mainly for local table grape consumption at Palestine with 22 common SSR markers, revealing an evaluable level of genetic diversity in a region of immense historical importance for viticulture. These genotype-based diversity analysis, coupled with other studies regarding the genetic diversity level in more grapevine germplasm collections [54, 35, 55, 49, 56, 57], agreed upon the high degree of molecular diversity in grape. The nuclear SSR diversity revealed for cultivated grapevines ranged from 0.6 to 0.85, averaging 0.77, with a mean number of alleles per locus equal to 16.9 [57]. This diversity is comparable or slightly lower than the one observed in natural population of *Arabidopsis* (14.4 alleles/locus, gene diversity = 0.83) [58], in wild populations of wild rice in China (gene diversity = 0.86) [59], and in collection of maize (14.8 alleles/locus, gene diversity = 0.79) [60]. Diversity values (expected heterozygosity) for SNP are generally low due to their bi-allelic nature. In grapevine, SNP diversity values ranged from 0 to 0.66 with a mean value of 0.30 [61, 62], which is slightly higher than the mean value reported for maize (0.26) [63].

Therefore, the exploitation of this high genetic diversity in grape will be helpful to understand the functioning of grape genome and to discover the genetic basis of important agronomical traits in order to support new breeding programs.

1.2.2. *V. vinifera* subsp. *sylvestris*

V. sylvestris is considered the putative ancestor of the cultivated grapevine and represents the only endemic taxon of the *Vitaceae* in Europe and Maghreb [64]. Wild grapevines have been identified in France [65], Spain [66], Italy [67], Germany, Switzerland, Austria, Romania [33] and Tunisia, as well as in other European countries (Figure 2) [25]. Apparently, Spain and Italy harbor the highest number of recorded wild populations and they were proposed to work as shelters for *V. vinifera* during the last glaciation as well as putative sources of postglacial colonization and diversification [68].

Figure 2: Localization of wild grapevine population in the Mediterranean basin. (Source: Heywood and Zohary [64])

However, it has been questioned if the current wild vines are real *sylvestris* individuals that have never been cultivated, or if they are naturalized cultivated forms escaped from vineyards as well as hybrids derived from spontaneous hybridizations among cultivated and wild forms [69]. Currently, wild grapevine is endangered throughout all its distribution range [70], with small and isolated population in Europe and temperate regions along deep river banks. Indeed, the distribution of the wild grapevine has dramatically been reduced over the last centuries with the introduction of pathogens from North America (phylloxera, oidium, mildew). Most of them died, except in floodplain forests as the root–host homoptera phylloxera was sensitive to flooding [71]. Moreover, while American resistant rootstocks were introduced in the vineyards to face phylloxera pest, this insect continued to infect populations of wild grapevines in regions of floodplain forests where the water table lowered. Intensive river management, starting in the middle of the 19th century, enhanced this process. In addition, the replacement of the floodplain forest by arable crops and meadows as well as the intensification of forest management with the

removal of the vines, considered detrimental to tree growth, led to a further fragmentation of wild grapevine habitats. This had an enormous impact on gene exchanges between populations, leading to a bottleneck, especially in gyno-dioicious plants [69]. Therefore, the reduction of wild grapevine populations by human actions led to a decrease of genetic diversity within most of the analyzed population of *V. sylvestris*. De Andrés et al. [66] performed a wide search of wild grapevine populations in Spain, collecting 237 individuals in 61 different locations. The amount and distribution of their genetic diversity was assessed using 25 nuclear SSR loci. The number of alleles per locus ranged from 3 to 17, with an average of 9.0, and 17 alleles showed a frequency lower than 1% (rare alleles). A slighly reduction of observed heterozigosity (H_o = 0.6) was observed compared to the expected heterozigosity (H_e = 0.7), pointing to the existence of inbreeding in some wild grape populations (Fixation Index (F) from 0,04 to 0,54). A comparable result was obtained by Emanuelli et al. [49], which observed average values of 10.6 and 1.9 alleles per locus for the SSR and SNP loci respectively in a wild grape population of 139 genotypes. In addition, a lower heterozigosity was observed within the *sylvestris* group than the cultivated population: the H_o evaluated with SSR and SNP markers was equal, respectively, to 0.63 and 0.25 in the former, and 0.76 and 0.35 in the latter. These results have been supported by other surveys of the level of genetic diversity in wild grape populations [52, 36, 28, 22]. A different scenario was described by Ergül et al. [72], which observed greater genetic diversity in wild grapes from Anatolia than the one of authoctonous grape cultivars. This result was expected as Anatolian populations are located at the primary center of diversity and thus are more diverse than in the peripheral populations. Accordingly the genetic diversity analysis of wild grape samples from different geographic locations of Georgia at four polymorphic microsatellite loci revealed high level of polymorphism [73]. Therefore, the wild forms still conserves an overall important genetic diversity, which can be explored to avoid the loss of biodiversity affecting the viticulture [69]. In this regard, the maintenance of genetic variability and the phenotypic characterization within wild grape populations has become a priority. Revilla et al. [74] have characterized the anthocyanin profile of 126 mostly Spanish wild grapevine accessions during several years. Considerable variability in the anthocyanin fingerprints was observed, leading to distinguish three groups: (i) in the first group (23 accessions), grapes did not contain acylated anthocyanins [75], occurring primarily in Pinot Noir and its mutants, in some grey and rosé cultivars or white grapes [76]; (ii) in the second group (17 accessions), grapes contained acylated anthocyanins and a high proportion of cyanidin-derived monoglucosides, occurring rarely in cultivated grapevines; (iii) in the third group (86 accessions), grapes contained acylated anthocyanins and a large proportion of delphinidin-derived monoglucosides, as do most grapevine cultivars [76]. Therefore, there is a considerable genetic variability related to anthocyanins in Spanish wild grapevine populations, higher than those reported for cultivated varieties commonly considered of Spanish origin [74]. Bodor et al. [77] compared 45 wild grapevine accessions from Germany, Italy and Turkey for 36 ampelometric traits using digital image analysis. The investigation of leaf morphological characters among the wild grape accessions revealed how geographic origin, sex of the flowers and vintage have significant effect on the broad diversity of leaf morphology in wild grapes. Particular interest has been raised on the genetic variability related to the resistance against pests and disease within *V. sylvestris*. Recently, Guan et al. [78] conducted a broad screen to evaluate the susceptibility levels to Botryosphaeria dieback, an important grapevine trunk disease, within a large selection of accessions from the family *Vitaceae*, including also *V.sylvestris* individuals. Large variation of resistance levels was found, with good performance in several accessions from *V. sylvestris*, whose resistance correlated with earlier and higher induction of some defence genes,

both in green and necrotic wood. Moreover, leaves of several *V. sylvestris* accessions were also less susceptible to necrosis induced by treatment with a culture filtrate of *Botryosphaeriaceae*, compared to commercial cultivars of *V. vinifera*. Furthermore, Riaz et al. [79] screened 306 *V. vinifera* cultivars, 40 accessions of *V. sylvestris*, and 34 accessions of *Vitis* species from northern Pakistan, Afghanistan and China, with 34 SSR loci, which included markers in linkage to the known powdery mildew (*Erysiphe necator*) resistance loci Ren1, Run1, Run2 and Ren4 [80, 81]. Two mildew resistant genotypes of *V. sylvestris* were identified, which presented the sequences previously identified in two mildew resistant *V. vinifera* cultivars: 'Kishmish vatkana' and 'Karadzhandal'. Accordigly, Tisch et al. [82] analyzed a collection of the European wild grape, representing a complete copy of the genetic variation still present in Germany, revealing that many genotypes show good tolerance against several grapevine diseases, such as downy mildew (*Plasmopara viticola*), powdery mildew (*E. necator*), and black rot (*Guignardia bidwelli*). In addition, Duan et al. [83] investigated the potential genetic variation in *V. sylvestris* with respect to their output of stilbenes and potential use for resistance breeding. Considerable variation in stilbene inducibility was identified in wild grapes, which splitted in two clusters of stilbene 'chemovars': one cluster showed quick and strong accumulation of stilbenes, almost exclusively in the form of non-glycosylated resveratrol and viniferin, while the second cluster accumulated fewer stilbenes. A screen of the population with respect to susceptibility to downy mildew of grapevine revealed that the subpopulation of genotypes with high stilbene inducibility was significantly less susceptible than low stilbene genotypes. On the other hand, Ocete et al. [84] observed in 53 (25 females and 28 males) wild individuals from Spain a wide range of leaves morphologies and a remarkable low incidence of pests and diseases. In addition, some of these wild genotypes produced wines with high acidity and intense color. Therefore, a broad phenotypic and genotipic variation can be observed in the grapevine wild-relative, which may shift into the centre of the attention of plant breeding and evolutionary biology, as a valuable genetic resource for breeding and sustainable viticulture [85].

1.2.3. Genetic relationship between cultivated and wild *V.vinifera*

Since the advent of molecular markers several analyses have been focused on the genetic relationship between cultivated and wild grapes, outlining a low but clear distinction among the two forms of *V.vinifera*. The analysis of population structure within a grapevine collection of wild and cultivated accessions from Spain identified four main clusters: the first (C1) and second (C2) clusters were mainly composed by wild individuals, instead of clusters 3 and 4 consisting notably of cultivated accessions [66]. In particular, the two genetic groups C1 and C2 mirrored the geographic origin of wild accessions from respectively Northern and Southern regions of Spain. A clear genetic differentiation was detected between wild and cultivated grapevine forms (F_{ST} = 0.12), even if the existence of a restraint genetic exchange between them was suggested. Indeed, as expected for an outcrossing dioecious subspecies, 10 spontaneous hybrids (4% of the collected samples) between wild and cultivated forms were detected. This result was in agreement with the low pollen flow between vineyards and wild plants reported by Di Vecchi et al. [70], which tested a direct paternity-based approach for the characterization of pollen-mediated gene flow between wild and cultivated populations of grapevine. The pollen flow resulted strongly correlated to the distance between individuals, with an estimation of pollen immigration in the wild populations from the cultivated compartment ranging from 4.2% to 26%. However, most of the fertilizing

pollen could be assigned to wild males growing nearby. This result could explain the positive values of the inbreeding coefficient (F_{IS}) observed in wild grapevine accessions from Spain [66].

A clear genetic differentiation between the two *V. vinifera* subspecies has also been reported by Emanuelli et al. [49]. Indeed, the Principal Component Analysis (PCA) approach clearly differentiated along the PC2 the *sativa* from the *sylvestris*, accounting for 5% and 8% of the total genetic variability for SSRs and SNPs respectively. However, a clear overlapping zone was observed, highlighting the occurence of gene flow between wild and cultivated grapes. The overall F_{ST} value equal to 0.16 between cultivated and wild grapevines strongly supported this probability. Much more resolution in the distinction between wild and cultivated grapes was gained by applying a hierarchical population structure analysis with the sofware STRUCTURE [86]. Indeed, some wine grapes related to Pinot Noir and Traminer, two ancient cultivars [87], were clearly distinguished from wild grapevine accessions (Figure 3).

The relationship between wild and cultivated grapevine has been recently investigated using high-throughput SNPs discovered with Next Generation Sequencing (NGS) technologies. The analysis of relatedness among *sativa* and *sylvestris* genotypes by using 5K SNPs provided strong support for a clear differentiation between the two forms of *V. vinifera* (F_{ST} ~0.1) [34]. However, relatedness among geographically diverse sample of wild and cultivated revealed how all *sativa* accessions were genetically closer to *sylvestris* populations from Near East than to wild populations from Western Europe. On the other hand, Western European cultivars were more closer related to western *sylvestris* than other *sativa* accessions, remarking the occurence of gene flow between wild and cultivated grapes in Western Europe. Two main groups of *sativa* and *sylvestris* were also identified in a grapevine collection from Georgia by De Lorenzis et al. [54] through the latest Vitis18KSNP array. The F_{ST} value, accounting 0.1, meant that the two groups have a moderate differentiation, in agreement with the gene flow between the wild and cultivated compartments [70].

In symmary, the picture arising today about the genetic relationship between *V. sativa* and its wild-relative *V. sylvestris* is of a clear differentiation between the two subspecies. Evidence of genetic introgressions between wild and cultivated compartments have been provided, highlighting how the hybridization has played a central role in the domestication and diversification of modern cultivars. Indeed, the analysis of genetic diversity within *sativa* cultivars have defined a large complex pedigree resulting from a number of spontaneous and inter-generation crosses between cultivars that have been vegetatively propagated for centuries [34]. On the other hand, positive values of F_{IS} have been observed within the wild populations, suggesting a potential inbreeding depression likely resulted from their small size as well as the intra-population pollen flow and the absence of inter-wild population flow [70].

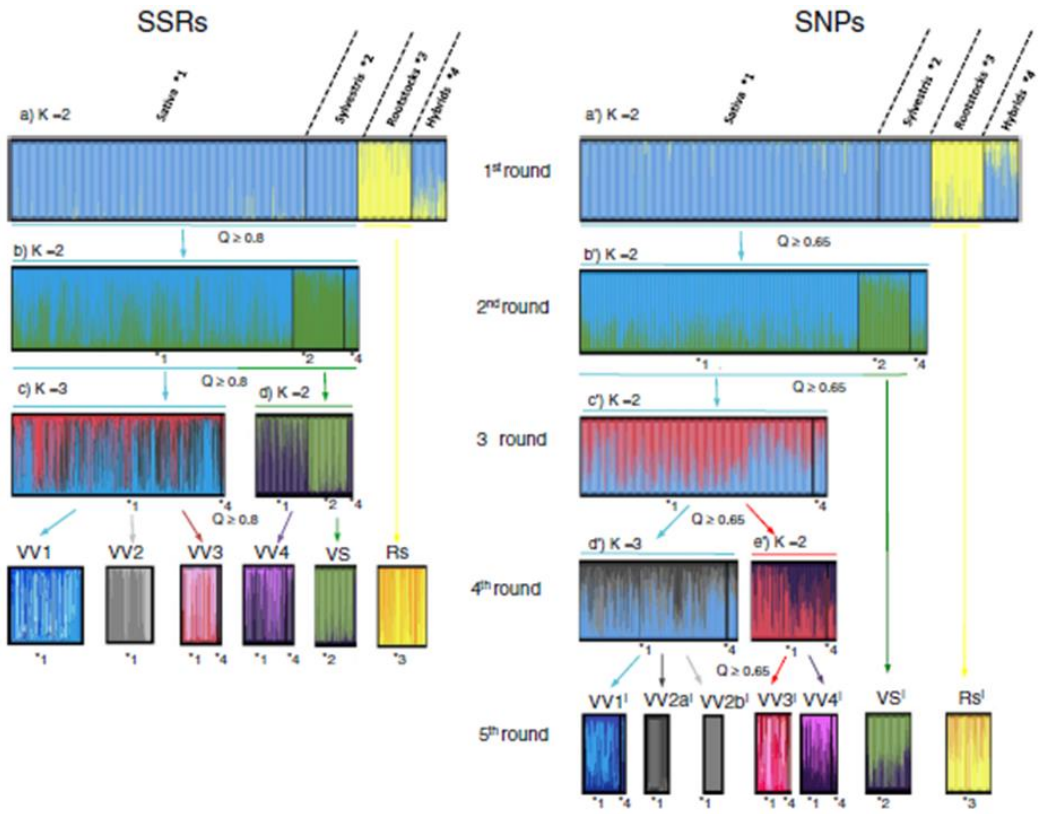

Figure 3: Flow chart of hierarchical STRUCTURE analysis of the Vitis germplasm composed by 1,085 unique accessions using 22 SSRs and 384 SNPs. In the first chart, samples of the four predefined groups are separated by black lines, while in subsequent charts, populations found by previous rounds of analysis are separated. Ultimately for the SSR and SNP data, respectively, there are: 1 cluster of rootstocks (Rs/RsI), 1 cluster of Vitis vinifera sylvestris (VS/VSI) and 5 subclusters of cultivated grapevine: VV1, VV2, VV3, VV4/ VV1I, VV2I, VV3I, VV4I. Q – membership coefficient. (Source: Emanuelli et al. [49]).

1.3. Grapevine genomics and genetics

1.3.1. Whole genome sequences

The first reference sequence of the grapevine genome has been reported in 2007 by Jaillon et al. [15], being the first genome produced for fruit crop, the second for woody species and the fourth for flowering plants. The nearby full homozygous line PN40024 (estimated homozigosity ~ 93%), derived from Pinot Noir by successive selfings, was sequenced through the whole-genome shotgun strategy, gaining an 8.4-fold coverage of the genome. When considering only one of the haplotypes in each heterozygous region, the assembly consisted of 19,577 contigs and 3,514 supercontigs, for an overall sequence of 487 Mb. 69% of the assembled genome was anchored along the 19 linkage groups (LGs) of the reference genetic map. Repetitive/transposable elements (TEs) constituted 41.4% of the grapevine genome, a slightly higher proportion than the one identified in the rice genome [88]. By the analysis of paralogous regions, it was concluded that the current grapevine haploid genome originated from the contribution of three ancestral genomes. The comparison of the grapevine gene regions with those of other completely sequenced plant genomes led to conclude that the paleo-hexaploidy was present in the common ancestor to grapevine, Arabidopsis and poplar. In particular, it seems that the formation of the palaeo-hexaploid ancestral genome occurred after the separation between monocotyledons and

dicotyledons and before the radiation of the Eurosids [15]. An alternative scenario was proposed by Velasco et al. [89], which presented a draft genome sequence of a cultivated clone of Pinot Noir with a size of 504.6 Mb. The variation within this clone of grape consisted largely of chromosome-specific gaps and hemizygous DNA. Indeed, the two homologous chromosomes showed either different sequences in some genomic regions (hemyzygous DNA) or gaps corresponding to sequence present in just one chromosome. These results suggested that the two homologous chromosomes of the cultivated Pinot Noir differ on average by 11.2% of their DNA sequences and that grape exists in a dynamic state mediated in part by transposable elements [90]. In addition, over 2 millions SNPs, of which 1,7 millions anchored to the 19 LGs, were discovered between the two homologous chromosomes, for an estimated SNP frequency of 4.0 polymorphisms per Kb. By the evaluation of the number of synonymous substitutions per synonymous site (K_S), a relative recent large-scale duplication in the grapevine genome was proposed. Therefore, three genome duplications were assumed to have occurred in both poplar and Arabidopsis [91, 92], one of which has been shared by all dicots, one that has been shared by Arabidopsis and poplar but not *Vitis*, and one that has been specific for Arabidopsis and poplar respectively [93]. In addition, a hybridization event might have occured in *Vitis* after the genome duplication shared by all dicots, explaining the presence of many grapevine genomic regions in triplicate. Other individual grapevine genomes have been completely sequenced so far [94, 95, 96], highlighting the complexity and high variability of grape genomes. The development of third-generation sequencing (TGS) technologies offers several advantages, such as longer read lengths (i.e. ~10 Kb with the single-molecule real-time (SMRT) sequencing, developed by Pacific BioSciences, PacBio [97]), which will benefit current grapevine genomics by closing gaps, characterizing structural variation in individual genomes and studying the grapevine methylome [98]. Indeed, new sequencing projects of other individual grapevine genomes are in progress [99]. They will open a new stage of the grapevine genomics, which will see the integration of –omics technologies to better understand deeply the functional complexity of the grapevine genome and its interaction with environmental stimuli.

The latest updated gene prediction, called v2 [100], counts 31,922 genes and 55,649 transcripts in the grapevine genome. Indeed, the incorporation of RNAseq data allowed to add 2,258 new coding genes and 3,336 putative long non-coding RNAs to the previous gene predictions [101]. 80% of the new genes were found to have at least one gene ontology annotation, enriching the list of functional categories with functions that were previously under-represented, such as those related to nucleotide binding site. The v2 gene prediction showed longer transcripts and coding sequences (CDS), with an average length of 1,207 and 247 bp respectively, and a number of exons per gene equal to 5,3. 30% of v2 predicted genes undergo alternative splicing producing 32,395 different isoforms. In particular, 64% of the alternative spliced genes produced more than two isoforms and a total number of 21,632 alternative splicing events were identified. The comparison of alternative splicing in different tissues, genotypes and stress conditions led to conclude that the extent of change in alternative splicing due to stress is similar to that seen in different tissues, clearly indicating its role in stress response.

1.3.2. Challenges of grapevine genetics and Linkage Disequilibrium

As other tree species, grapevine is a challenging genetic system compared to herbaceous species such as Arabidopsis and cereals [102]. The grapevine plant presents several physiological constraints, such as its deciduous perennial nature, vineyard space requirements, an annual reproductive cycle and a generation time varying between 2 and 5 years, depending on genotype and growing conditions [103]. A novel grapevine system suitable for rapid genetic studies in small controlled environments has been described [104]. It is based on the mutant allele in the grapevine GA insensitive gene (VvGAI1) which confers a dwarf stature, short generation cycles and continuous flowering ('microvine') [104]. Recently, Chaïb et al. [105] demonstrated how the 'microvine' can be used for rapid plant transgenic studies and for rapid genetic mapping and trait dissection beyond an initial F1 generation.

Due to the cultivated grapevine derived from the domestication of dioecy wild plants followed by extensive vegetative propagation, current grapevine genomes are highly polymorphic [28]. Therefore, the extent of linkage disequilibrium (LD) is generally low in the short range when a sample of genetically distant genotypes is analyzed [62]. LD is a measure of the degree of non-random association between alleles at different loci [106]. It can be considered as a historically reduced level of the recombination of specific alleles at different loci controlling particular genetic variations in a population. The basic measurement of LD is determined by calculating the difference between observed haplotype frequency and that expected based on allele frequency [107]. Usually LD is measured by two related statistics D' [108] and r^2 [109], which both can have values ranging from 0 to 1. If they are equal to zero the presence of alleles at different loci are completely indipendent of one another (linkage equilibrium), while if D' and r^2 are equal to 1 the presence of alleles at different loci are totally correlated. The main difference between D' and r^2 is that the latter accounts for both recombination and mutations events, while the former takes in consideration just the recombination events [110]. Indeed, LD is a sensitive indicator of the population forces that structure the genome, such as mutation, genetic drift, population structure and selection [111]. In particular, the bottlenecks associated to the domestication led to reduce genetic diversity and to increase the extent of LD by eliminating recombinant lineages [112]. Even when loci remain polymorphic, the number of allelic combinations across loci can be much reduced, leading to extensive haplotype structure (Figure 4) [112].

The studies published to date on the extent of LD in grape suggest that LD decays to background levels within a small number of kilobases. Lijavetzky et al. [62] characterized over 200 random gene fragments, representing circa 1 Mb of total sequence and over 1,500 SNPs, within 11 genotypes corresponding to ancient unrelated cultivars as well as wild plants. r^2 values close to 0.2 were observed along genetic distances of 100-200 bp between pairs of SNP loci. Accordingly Myles et al. [21] evaluated the LD decay in 10 cultivated genotypes of *V. vinifera* with the Vitis9KSNP array, concluding that, while LD is generally low across all distances, it remains above background levels to ~10 kb. Moreover, this rapid LD decay appeared unchanged in 59 accessions of *V. sylvestris* genotyped with the same SNP technology [34].

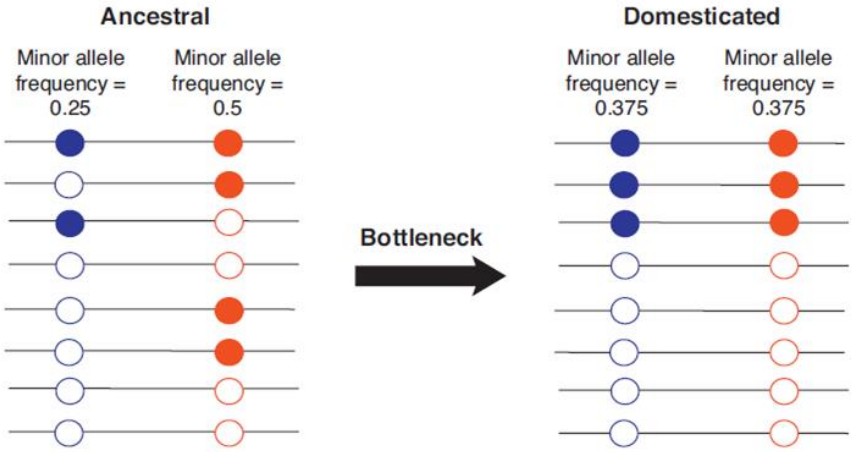

Figure 4: During a bottleneck, lineages are lost from the population. This leads to lost one or more of the gametic types with a consequent increase of LD. Indeed, only two of the four possible gametic types remain after a bottleneck, resulting in a situation of 'perfect LD' between SNPs. (Source: Hamblin et al. [112]).

Recently, Nicolas et al. [113] assessed LD extent by genotyping 372 SNPs over four genomic regions and 129 SNPs distributed over the whole genome in a diversity panel of 279 cultivars. LD, measured by r^2 corrected for kinship, reached 0.2 for a physical distance between 9 and 458 kb depending on genetic pool and genomic region. In addition, different values of LD were observed across the four genomic regions between wine eastern cultivars, wine western variaties, eastern table grapes and wild grapevine individuals. In particular, LD extent in the wild panel ranged from 31 to 127 kb. Further studies on the pattern of LD across the whole genome are still necessary to design suitable grapevine collections for genome-wide association studies (GWAS) and genome selection (GS). Indeed, differences in the extent of LD have a very important effect on the marker density required for GWAS and GS, and the potential gene mapping resolution. Moreover, the assessment of LD size in cultivated and wild populations of grapevine will help to understand which evolutionary forces have been operating and whether some genomic regions have been subjected to selective pressures during the long history of viticulture [114].

1.3.3. How to identify genes responsible for natural genetic variation in grapevine

The identification of genes underlying the natural genetic variation for specific traits as well as the perception of the nature and effects of their allelic differences represent a major challenge in grapevine genetics [103]. Since the common quantitative nature of genetic variation, quantitative trait loci (QTL) mapping approaches are frequently applied to identify the genomic regions responsible for the phenotypic variation at different traits in grapevine. QTL mapping studies in grape usually rely on the use of F1 progenies obtained by crossing cultivars [115] or in selfed progenies [116]. QTL mapping has been extensively used to identify genomic regions contributing to resistance traits in crosses between *V. vinifera* cultivars and other *Vitis* species resistant to several grapevine diseases. This is the case for the Run1 [117] and Ren1 [81] loci responsible for dominant resistance to powdery mildew (*Erisyphe necator*), and the Rpv1 locus for the resistance to downy mildew (*Plasmopara viticola*) [118]. Recently, the inheritance of powdery

mildew resistance and susceptibility of wild *V. rupestris* B38 and cultivated *V. vinifera* 'Chardonnay' has been studied by using 17K SNPs identified with genotyping-by-sequencing (GBS) approach [119]. Linkage maps of over 1,000 SNPs were constructed for the two parents and the 'Chardonnay' locus named Sen1 (Susceptibility to Erysiphe necator 1) was corroborated, providing the first insight into the genetics of susceptibility to powdery mildew from *V. vinifera*. Regarding plant growth and physiology, the genetic structure of the traits seems to be complex and controlled by many QTLs of small effects. Exceptions are the control of plant sex and berry colour, that seems to be regulated by single loci, both located at LG2 [120, 121]. Other QTL mapping analysis have focused on the genetic control of berry-related triats, such as seedlessness and berry size [122][123]. A major seed development inhibitor (SDI) locus was detected on LG18 with a dominant effect on seedlessness and pleiotropic effects on berry size. In addition to this major QTL on LG18, Doligez et al. [42] have recently identified five new QTLs for berry weight on LGs 1, 8, 11 and 17, and four new loci for seed traits on LGs 4, 5, 12 and 14. Several QTLs have also been identified for sugar and organic acid composition of grape fruits. Chen et al. [43] reported 14 QTLs at ten LGs for berry sugar content, and 8 QTLs for berry malic acid content, total acidity and tartaric acid-to-malic acid ratio on LGs 6, 13 and 18. Houel et al. [124], by constructing a mapping population of 129 microvines derived from Picovine x Ugni Blanc flb, identified seven major and minor QTLs for malate and tartrate contents at green lag phase of grape berries, of which four co-localize with the seed number and berry weight QTLs on LG 7. Even though QTL mapping analysis have successfully provided a list of candidate genes putatively underlying the investigated traits in grapevine, the final demonstration of the role of a specific gene in the determination of a given phenotypic trait is still missing. Indeed, QTL mapping has limited mapping resolution and relatively low power in accurately estimating the number and size of QTLs [125]. In addition, the results of QTL analysis often depend on the environment as well as the parental lines used in the cross [126].

An alternative to mapping traits in segregant populations is to performe LD- or association mapping, which uses a population of unrelated individuals [127]. Indeed, LD mapping approach is applied on samples of individuals from germplasm collections or natural populations, leading to explore a broader genetic variations with wider background for marker-trait correlations (i.e., many alleles evaluated simultaneously) [128]. Therefore, association mapping relies on the utilization of majority recombination events from a large number of meiosis throughout the germplasm development history [129]. As a result, the phenotype of interest may be associated with a much smaller chromosomal segment than in a classical bi-parental QTL mapping, providing in theory greater mapping resolution. LD mapping can be separated in two types, each focusing on a different level of genetic analysis. The first, called "candidate-gene association mapping", focuses on the genetic variation in one or few candidate genes, putatively involved in the phenotypic variation of specific traits [130]. The second type of association analysis, called "genome-wide association mapping (GWAS)", aims to identify genome-wide variation that associates with phenotypic variation. Therefore, GWAS requires measures of genetic variabiity in markers representing most of the genome and tests phenotype-genotype association for each marker [131]. However, one of the main limitations of LD mapping is the detection of spurious phenotype-genotype associations due to population structure [107]. Usually, population structure is geographic because crops were moved to a much broader range of environments, where natural selection drove genetic adaptation to these new habitats. Equally important is genetic structure associated with end-use or cultural preferences, such as table and wine grapes [113]. In

such cases, the phenotypic variation within subpopulations will strongly correlate with the differences of their allele frequencies, leading to false-positive marker-trait associations. In this regards, several models have been built so far to account for confounding factors in LD mapping [132, 133]. To date, GWAS has been mainly applied in cereals and other herbaceous species [134, 135, 136]. For instance, the phenotypic variation in malting quality in barley was successfully linked to haplotype variation at the β-amylase2 gene, a locus involved in starch hydrolysis [137]. In maize, a GWAS approach was applied to identify QTLs and underlying candidate genes for leaf metabolite variation [138]. Recently, association mapping studies have been carried out also in perennial species, such as apple [139] and banana [140], revealing how LD mapping is a valuable genetic tool to dissect the genomic basis of main agronomical traits in complex genetic systems.

Both QTL mapping and GWAS approaches have been extensively used as part of the "top-down" strategy for identifying genes underlying specific traits [126]. Indeed, the top-down approach begins with the phenotype and uses genetic analyses to uncover genomic regions and candidate genes involved in the phenotype of interest. An alternative approach, named "bottom-up approach", start by using population genetics to discover "signature of selections" and than make use of other genetic tools to identify the phenotypes to which these genes contribute. Indeed, selection reduces variation at genomic regions surrounding genes controlling target phenotypes, because just a portion of the population will carry the alleles under selection. Therefore, only the selected alleles and those of genes in close linkage ("genetic hitch-hiking" [141]) will be retained [142]. This localized reduction of diversity at the selected locus and its surrounding genomic regions is well defined as "signature of selection" (Figure 5). Researchers in molecular evolution divide selection in different categories. Positive selection is defined as any type of selection in favor of new advantageous mutations. Negative selection refers to the opposite case in which selection acts against new mutations, also known as purifying selection. Balancing selection occurs when two or more extreme phenotypic values are favored simultaneously. This type of selection will often increase variability (R Nielsen 2005). Overdominance, which occurs if the heterozygote has the highest fitness, is a case of balancing selection [143].

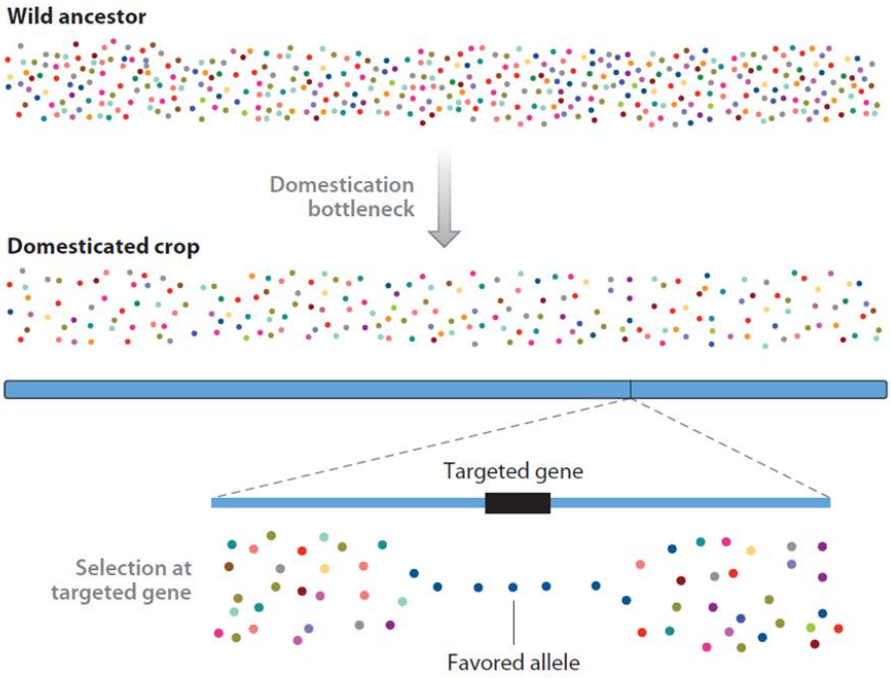

Figure 5: The impact of domestication on genetic diversity. Colored dots represent neutral allelic diversity at genes across a chromosome (blue bar) in populations of a crop's wild ancestor (top) and in the crop itself (middle). Genetic drift, acting strongly during the domestication bottleneck have caused a genome-wide reduction in genetic diversity. In contrast, selection have differentially reduced diversity at the specific genes that control the traits subject to selection. As a favored allele is driven to high frequency, much of the standing genetic variation within and around the targeted gene (black bar) is removed from the population, creating a molecular signature of selection. (Source: Olsen and Wendel [142]).

Bottom-up approaches have been used to identify domestication loci [142], which are the genomic regions underlying the main changes occurred during crop domestications [144]. Indeed, the earliest agricultural practice was to grow and harvest wild plants of a favorable species, marking the shift from the hunter-gatherer life to agricultural civilization [144]. Afterward, humans would select the individuals with the desired characteristics in the wild species populations and use the favorable seeds to resow and plant the next year. During these constant cycles of human selection and crop improvement every year, many morphological and physiological traits of the wild progenitors were reshaped. The traits under human selection in crop domestication included seed dormancy, flowering time, mating system (e.g., the change from dioecious to monoecious plants in grapevine [25]), and coloration. Understanding the genetic basis of domestication-related traits is of particular importance since they still represent a target of modern crop breeding [145] (Figure 5). Bottom-up approaches begin with whole-genome profiling of sequence variation in a diverse population sample, including domesticated varieties and its wild ancestors. Afterwards, genome scanning for selection signatures, also referred to as "selective sweeps" [146], is performed by applying population genetics methods. One way to detect selective events is the comparison of allele frequencies within and between populations (i.e. domesticated versu wild individuals) by using Wright's fixation index (F_{ST}), the most common metric for population differentiation [147]. F_{ST} is defined as the difference between the average expected heterozygosity of subpopulations and the expected heterozygosity of the total population based on the Hardy–Weinberg equilibrium [106]. Indeed, if selection is acting on

a locus within one population but not within other related populations, then the allele frequencies at that locus among the populations can differ significantly. Large values of F_{ST} at a locus indicate high differentiation between populations, which is suggestive of directional selection, a case of positive selection [148]. Small values indicate that the populations being compared are homogenous, which may be indicative of balancing or directional selection in both [114]. Signatures of selection can be also detected by measuring and comparing the level of nucleotide diversity within each population [149]. The nucleotide diversity (π) is the average number of nucleotide site differences found when each unique pair of DNA sequences in a sample is compared. π is sensitive to the frequency of each DNA sequence allele in a sample, since more frequent sequences appear in more of the pairwise comparisons. Lin et al. [150] compared the level of nucleotide diversity between populations of the wild species *Solanum pimpinellifolium* with small red-fruited tomato, *S. lycopersicum* var. *cerasiforme* with cherry tomato and the big-fruited tomato *S. lycopersicum*. 186 and 133 regions were identified as candidate domestication sweeps and improvement selective sweeps respectively, ,leading to develop a two-steps evolution hypothesis of fruit mass in tomato. The analysis of nucleotide diversity is often completed with the test of Tajima's D, which quantifies the reduction in the genetic diversity around the selected locus by comparing π with the total number of segregating polymorphisms (θ) [151]. A segregating site is any nucleotide site that maintains two or more nucleotides within the population. Due to π is sensitive to the allele frequencies in the sample, a surplus of low frequency alleles (rare alleles) inflates θ. This leads the Tajima's D to reach negative values, which indicate positive selection [151]. On the contrary, positive values of D results from an excess of intermediate-frequency alleles, which may occur in case of balancing selection [114]. Branca et al. [152] observed strongly negatively skewed distributions of Tajima's D in a diverse collection of 26 *M. truncatula* accessions, due to an excess of low-frequency SNPs. These skewed distributions of D may reflect a recent population expansion or positive selection events in *M. truncatula*. Bottom-up approaches have been applied to different species, such as rice [153], maize [154], apple [155] and sorghum [135]. Indeed, bottom-up approaches have several advantages for finding genes that contribute to adaptive traits and that will be useful in an agronomic context: (i) it is not necessary to develop segregant populations; (ii) far fewer plant samples are required compared to LD mapping; (iii) as association mapping, population genetics approaches can be applied to species with a long juvenile phase; (iv) they provide historical insights into the process of domestication [126]. Bottom-up and top-down approaches are complementary genetic tools. For instance, population genetics studies can provide candidate genes to further genetic analysis with LD mapping, or rather GWAS results can be better interpreted by implementing population genetic methods.

OBJECTIVES

Grapevine is a complex genetic system due to its perennial nature and high polymorphic genome, that slow down the identification of genes underlying important agronomical traits. In addition, the origin of modern grapevine cultivars and their genetic relationship with their own wild relatives are still controversial. The genetic and phenotypic exploration of wild populations is becoming a priority in grapevine, since they still represent a reserve of natural genetic diversity, which can be exploited in future plant breeding to face the genetic erosion occurring nowadays in viticulture. Therefore, the present research aims to provide further evidence of the relationship between cultivated and wild grapevines at both genomic and phenotypic levels. Moreover, in the present study the feasibility of new genetic tools, such as GWAS and population genomics, is explored as an alternative to the traditional strategies of gene mapping, aiming to acquire new information about the functional genomic basis of the phenotypic diversity observed in grapevine. Therefore, the present research has been structured as follows:

• **Chapter 2** describes the development of a new protocol of restriction-site associated DNA (RAD) sequencing technology, in order to discover and validate a dense panel of SNP loci throughout the grapevine genome in a germplasm collection consisting of wild and cultivated grapevine accessions.

• **Chapter 3** shows the application of population genetics approaches to characterize the differentiation between the two subspecies of *V. vinifera* and identify the genomic regions underlying the adaptation occurred during the grapevine domestication.

• **Chapter 4** is a study of linkage disequilibrium mapping in a population of *V. sylvestris* and *V. sativa* for ten domestication-related traits, including berry size and composition.

Chapter 2

SNP-DISCOVERY BY RAD-SEQUENCING IN A GERMPLASM COLLECTION OF WILD AND CULTIVATED GRAPEVINES (*V. vinifera* L.)

Abstract

Background: Grapevine genome has a high level of heterozygosity and a rapid linkage disequilibrium decay. The increase of molecular marker density throughout the genome is fundamental in order to improve the power and resolution of genetic mapping and to enable the application of population genomics methods. In this study we carried out the (high-throughput) SNP discovery in a grapevine germplasm collection of cultivars (*Vitis vinifera* subsp. *sativa*) and wild accessions (*V. vinifera* subsp. *sylvestris*) through a novel protocol of restriction-site associated DNA (RAD) sequencing based on 5500 SOLiD™ System.

Results: By resequencing 1.1% of the grapevine genome at a high coverage, we recovered 34K BamHI unique restriction sites, of which 6.8% were absent in the 'PN40024' reference genome. Moreover, we identified 37,748 single nucleotide polymorphisms (SNPs) that included 154 non-nuclear variants. 93% of markers belonged to the 19 assembled chromosomes with an average of 1.8K SNPs per chromosome. 48% of the identified SNPs fell in genic regions mostly assigned to the functional categories of metabolism and regulation which may reflect different adaptation mechanisms among wild and cultivated grapevines. The SNP validation with both Sanger sequencing and the Vitis20K array showed the ability of RAD-seq to accurately determine genotypes in a highly heterozygous species.

Conclusions: We provide a novel panel of high-quality and informative SNPs which reflects a considerable level of genetic diversity between *sylvestris* and *sativa* accessions. It will be useful in future surveys to select candidate polymorphisms contributing to domestication-related traits and to investigate the molecular pathways associated with plant response to environmental stimuli.

Background

The introduction of molecular markers in plant breeding has enabled remarkable increases in agricultural production thanks to the discovery of genes associated to major agronomic traits, the study of species diversity and evolution, and the characterization of plant genetic resources [156]. During the last ten years Single Nucleotide Polymorphisms (SNP) have become the markers most widely used due to their abundance in the genome. They compensate the biallelic nature by being ubiquitous and amenable to high-throughput automation [157]. The advent of Next Generation Sequencing (NGS) has increased the possibilities of de novo and reference SNP discovery in cost-effective and parallel manners. At the same time, huge progress has been achieved for high throughput SNP genotyping thanks to the introduction of array-based technologies, able to screen several thousands SNPs per assay [158]. SNP arrays rely on the prior production of sequence information, the identification and validation of polymorphisms and finally the array construction [159]. Myles et al. [21] designed the first SNP array for grape (Illumina Vitis9KSNP chip) which included 8,898 SNPs discovered in a panel of 17 genomic DNA samples from *V. vinifera* cultivars and wild *Vitis* species. The second highest throughput SNP array produced in grapevine as part of the GrapeReSeq Consortium [160] includes 18,775 SNPs (Illumina Vitis18KSNP array). De Lorenzis et al. [54] used this tool to investigate the genetic variability of a Georgian germplasm collection including cultivated and wild grapevine genotypes, obtaining a final panel of 12,083 polymorphic loci. These experiments have shown how the application of array-based technologies to population genetic studies may underestimate the real genetic diversity of the investigated populations, especially when the discovery panel is evolutionary divergent from the studied accessions [161].

Several methods that combine genome-wide SNP discovery and SNP genotyping are nowadays available. They rely on the use of restriction enzymes in order to reduce the portion of the genome to be sequenced. The number and type of restriction enzyme used as well as the amount of digested DNA, the multiplexing capabilities and the final depth of SNPs coverage distinguish the different protocols of genome-wide SNP discovery. One of these approaches is the Restriction-site Associated DNA sequencing (RAD-seq) based on rare-cutter restriction enzymes (6-8 bp recognition site) for sequencing short DNA fragments surrounding a particular recognition site throughout the genome. This method derives from the RAD tag marker technique [162] adapted to NGS platforms [163, 164]. The RAD-seq approach produces two types of markers: a) co-dominant SNP markers within the flanking regions of the restriction enzyme site; b) dominant markers due to sequence variations of the restriction endonuclease cutting site. RAD-seq has been applied in several plant species, such as sorghum [164] and eggplant [165], to discover SNPs, construct genetic maps and identify quantitative trait loci (QTLs). Recently, Wang et al. [43] genotyped a biparental population of grape interspecific hybrids with the RAD-seq approach producing a rather dense genetic linkage map of 1,814 SNPs. Chen et al. [43] using the same procedure built a genetic map of 1,826 SNP markers in a wine grape cross and could localize some QTLs for berry quality traits. Several modifications of the original RAD-seq protocol have been introduced by Genotyping-by-sequencing (GBS) [166], double digest restriction-site-associated DNA sequencing (ddRAD-seq) [167] and 2b-RAD-seq [168] methods. For instance, GBS [166] used a frequent cutter enzyme to generate reduced representation libraries prior to sequencing. GBS was first applied in grape by Barba et al. [119] to investigate the inheritance of powdery mildew (*Erysiphe necator*) resistance within a segregating population of *V. rupestris* x *V. vinifera* 'Chardonnay', finally mapping 35,8% of the 47K SNPs identified. Actually, one of the major

drawbacks reported for GBS is the high rate of missing data which is currently faced by imputation programs such as LinKImpute [169] and Beagle [170](Browning and Browning 2007).

The reference genome sequence of grapevine has been available since 2007 [15] with a total size of 487 Mb. Almost two million putative SNPs were reported for the heterozygous cultivar 'Pinot Noir' with an overall rate of 4 polymorphisms per kilobase [89]. A few other individual grapevine genomes have been completely sequenced so far. Da Silva et al. [94] analyzed the genome of the cultivar 'Tannat' using a mixture of de novo assembly and iterative mapping onto the 'PN40024' reference genome. The 'Tannat' genome was 1% shorter than the reference genome and presented more than two million single-base differences compared to the latter. Di Genova et al. [95] by sequencing the ancient table grape 'Sultanina' found 1,193,566 high quality SNPs and novel genes absent in the *V. vinifera* 'PN40024' reference genome. More recently Corso et al. [96] resequenced two grape rootstocks, both interspecific hybrids, revealing a SNP frequency of one variant every 200 bases with the 'PN40024' reference genome. All the mentioned analyses evidenced the high level of heterozygosity in the grape genome. Moreover recent studies [62, 34] showed low levels of Linkage Disequilibrium (LD) in *V. vinifera*, with a decay of LD at ~10 kb inter-SNP distances, which necessitates increasing the density of molecular markers throughout the genome.

As shown in other plant species [164, 171, 172] RAD-seq is a suitable method to develop robust markers for population genetics analyses. In this study we present the (high-throughput) SNP discovery carried out in cultivated and wild forms of *Vitis vinifera* through a novel protocol of RAD sequencing based on the 5500 SOLiD™ System. Our aim was to generate a tool for further investigations of grapevine domestication.

Methods

Plant material and DNA extraction

A germplasm collection of 51 cultivated (*Vitis vinifera* spp. *sativa*) and 44 wild-type (*Vitis vinifera* spp. *sylvestris*) grapevines was sorted at the FEM grape repository (ITA362), located in San Michele all'Adige, Italy (Appendix B, page 125). The *sativa* accessions were chosen within a genetic core collection (G-110) that retains 100% of SSR and SNP loci diversity present in the source collection [49]. The wild individuals, mostly originating from the Italian Peninsula, were selected within the *sylvestris* accessions of the same repository previously clustered through a hierarchical STRUCTURE analysis [49]. Young leaf tissue of one field grown plant per accession was harvested and stored immediately in sterile tubes at -80°C for DNA extraction and successive analyses. Total genomic DNA was isolated from freeze-dried tissue after grinding with the MM 300 Mixer Mill system (Retsch., Germany) using the DNeasy 96 plant mini kit (QIAGEN, Germany). DNA concentration and purity were checked both by the Synergy HT Multi-Mode Microplate Reader (BioTek) and the NanoDrop 8000 UV-Vis Spectrophotometers (Thermo Scientific).

Choice of Restriction Enzyme and Adapter Design

RAD-seq libraries (see paragraph “Libraries construction”) were previously constructed with genomic DNA from PN40024 using three restriction enzymes (HindIII, BamHI and NcoI) separately that present a different number of recognition sites on the grapevine reference genome. The number of restriction sites recovered by each RAD-seq library at different coverage thresholds (number of RE site with coverage 4X, 8X, 16X, 24X; Supplementary Table S1) was checked in order to apply the best candidate RE to the entire grapevine population.

Two types of adapters were used. The common 5500 Series SOLiD™ P1-T adapter for Fragment Library Preparation was modified by adding a biotin on the 5’ end of the top strand, and a 4 bp overhang, complementary to the sticky ends generated by BamHI, on the 5’ end of the bottom strand (Figure 1). The sequences of the top and bottom oligonucleotides are: 5’-Biotin-CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT-3’ and 5’-Phosphate-GATCATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGGCC-3’. The P1 adapter oligonucleotides were diluted separately in Milli-Q water (100 µM each) and then annealed in a thermocycler according to the following conditions: 95°C for 3 min, ramp down to 4°C by 1°C/30 sec; 4°C hold. The second adapter type was the standard barcoded adaptor used for 5500 SOLiD Fragment libraries and has a 10 pb barcode sequence. The different oligonucleotide sequences of the standard barcoded adapters are available on the Fragment Library Preparation 5500 Series SOLiD™ Systems User Guide [173]. Both biotinylated and barcoded adapters were diluted in water to 5 µM. Moreover, the presence of the restriction site in both adapters was verified in order to avoid its regeneration after the ligation with genomic DNA.

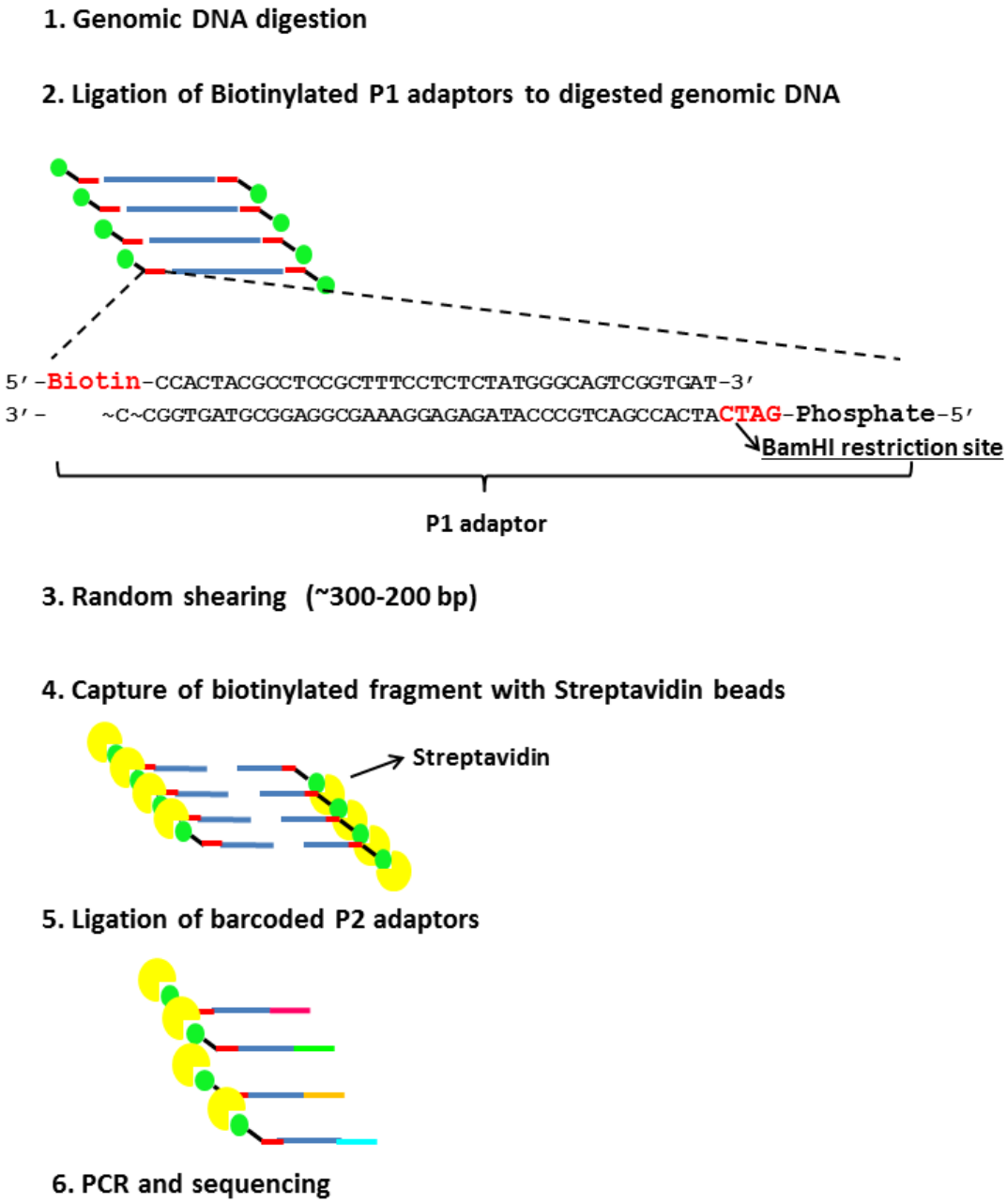

Figure 1: Main steps of the novel RAD-seq protocol: 1-2) sample genomic DNA is digested. The resulting digested DNA fragments are ligated to a P1 adaptor, that presents a biotyn group and a 4 bp overhang complementary to BamHI recognition site. **3-4-5)** Biotynilated fragments are random sheared to a target size of 300-200 bp, captured using streptavidin beads and ligated to standard barcoded adaptors for 5500 SOLiD Fragment libraries. **6)** RAD-seq libraries are amplified and purified before sequencing.

Libraries construction

DNA samples (500 ng) were digested with BamHI-High Fidelity (New England Biolabs, NEB) enzyme for 1h at 37°C in 25 µL volumes containing 1X NEB CutSmart Buffer and 5U of BamHI (Figure 1). Next 30 µL of ligation master mix, containing 4 pmols of the biotinylated P1 adapter, 1X T4 DNA ligase reaction buffer (Invitrogen™) and 1U T4 DNA ligase (Invitrogen™) were added to the digestion products, and samples were incubated at 16°C overnight. The ligation products were

purified using one volume of Agencourt AMPure XP beads (Beckman Coulter) according to the manufacturer's instructions and solubilized in 50 µL of 1X Low TE (10 mM Tris-HCl, 0.1 mM EDTA). DNA fragments were random sheared with a Covaris S220 Focused-ultrasonicator in 130 µL microTUBEs AFA Fiber Snap-Cap following the manufacturer's protocol for Target BP Peak of 200 pb. Afterwards the samples were vacuum concentrated to a final volume of 20 µL. Next 10 µL of Dynabeads® MyOne™ Streptavidin C1 (10 µg/µL), previously washed three times with 50 µL of 2X Binding and Washing (B&W) Buffer (10 mM Tris-HCl pH 7.5; 1 mM EDTA, 2 M NaCl), were added to each sample and resuspended in 20 µL of 2X B&W. Samples were incubated for 30 min at room temperature in rotation in order to capture the biotinylated fragments. Biotinylated coated beads of each sample were separated with a magnet for 2–3 min, collecting the surnatant in a clean tube to estimate the DNA recovery rate through a Qubit® 2.0 Fluorometer (dsDNA HS Assay; Life Technologies). The biotynilated coated beads were first washed with 50 µL of 1X B&W buffer and later with 50 µL of Buffer EB (Qiagen), and then resuspended in 20 µL of Buffer EB. Next 25 µL of NEBNext® End Repair Module (New England Biolabs) master mix, containing 5 µl of NEBNext End Repair Reaction Buffer (10X) and 2.5 µl of NEBNext End Repair Enzyme Mix (10,000 units/ml T4 PNK; 3,000 units/ml T4 DNA Polymerase), were added to the biotinylated beads. The End Repair mix was incubated for 15 min at room temperature in rotation. After the End Repair Enzymes inactivation at 75°C for 20 min, 50 µl of ligation master mix, containing 4 pmols of the blunt barcoded P2 adapters, 1X T4 DNA ligase reaction buffer and 10U T4 DNA ligase (Invitrogen™), were added to the biotinylated samples and incubated 1h at room temperature in rotation. The biotinylated fragments from each library were amplified in 50 µl volumes containing 25 ng DNA fragments, 1X GoTaq® Green Master Mix (Promega) and 25 pmol each of the following primers: Library PCR Primer 1, 5' -CCACTACGCCTCCGCTTTCCTCTCTATG-3' and Library PCR Primer 2, 5' -CTGCCCCGGGTTCCTCATTCT-3' [173]. The amplification was performed according to the following conditions: 95°C for 5 min, 12 cycles of 94°C for 20 sec, 62°C for 20 sec, 72°C for 50 sec, with a final Taq extension at 75°C for 3 min. PCR products were purified using 1.3 volumes of Agencourt AMPure XP beads. Each library was loaded on a Bioanalyzer (Agilent Technologies) for the evaluation of fragments size through a High Sensitivity DNA Assay. Libraries were considered suitable for sequencing if adapter dimers (99 bp in length) were minimal or absent and the majority of other DNA fragments were between 150–350 bp. If an excess of adapter dimers were present, the RAD libraries were purified again. Finally fragments sequencing (75 bp reads) was performed on a 5500 SOLiD™ System (Applied Biosystems, Life Technologies) pooling the libraries and running them in two different flow-cell lanes using the Exact Call Chemistry module (ECC).

Reads pre-processing

Reads were expected to start with the 5'-GATCC-3' sequence released by BamHI cut and corresponding to T12320 in color space format. Reads 75 bp long obtained from SOLiD sequencing were inspected for the presence of the T12320 sequence at their starting point. When there were no color errors or one color sequencing error at the beginning, the read starting sequence was replaced with the full color space BamHI restriction site (T102320). Reads with more than one color error in their starting sequence were discarded.

DNA sequence alignment

Pre-processed reads in color space were mapped on the reference 12X grape genome[15], the mitochondrial (mtDNA) [174] and the chloroplast (cpDNA) [175] DNA sequences using BFAST v0.7.0a [176] aligner. Only unique alignments with identity at least 90% were kept. All statistical analysis were performed using 'stats' v3.4.0 [177] and ggplot2 v2.1.0 [178] R packages.

SNP calling and annotation

The UnifiedGenotyper tool of the Genome Analysis Toolkit (GATK) v3.2-2 [179] was applied to call variants on unique alignments with a mapping quality score higher than 17. SNPs and indels having at least 10 reads and a quality score > 30 were retained. SNP genotypes were inferred through a Bayesian genotyper implemented in GATK that assigned genotype at each site as the genotype with the greatest posterior probability. SNP density across the *V. vinifera* 'PN40024' reference genome was evaluated by counting the number of SNPs in sliding windows of 500 kb using VCFtools [180]. Pearson's correlation (r) was used to determine the relationship between the number of SNPs per chromosome and chromosome physical size. Finally, SNPs were classified into genomic feature groups and gene classes according to the grape gene annotation v2.1 [100].

SNP validation

50 fragments were selected to validate 183 SNPs with Sanger sequencing [181]. PCR primers were designed using NCBI/Primer-BLAST [182] to yield products 266-1002 bases long. Target sequence fragments were amplified in 4 cultivated and 3 wild accessions chosen within the analyzed population. Another *V. sativa* variety, that showed an uncommon low level of genetic variation at microsatellite loci, was also included during Sanger sequencing in order to test the ability of RAD-seq markers to capture undisclosed genetic diversity. The products of Sanger sequencing were run on the 96-capillary 3730xl DNA analyzer (Applied Biosystems®). Finally, STADEN package v2.0.0 [183] was used to analyze the DNA sequences.

The grapevine population investigated in this study had previously been genotyped with the commercial GrapeReseq Illumina Vitis20KSNP chip [184]. The Infinium genotyping raw data were analyzed using the Genotyping Module v1.9 of the Illumina GenomeStudio Data Analysis software [185]. An individual locus analysis, where loci are identified by sorting on per-locus metrics such as call rate and cluster separation, was carried out to obtain a final data set of good quality SNPs. In order to assess the rate of fitted genotypes between GrapeReseq 20K chip and RAD-seq, the genetic profiles of the shared SNPs between the two data sets were compared for all samples, except for the sample GRAPE_51 which was not evaluated with the Vitis20KSNP chip (Appendix B, page 125).

Results

Sequencing summary

We selected BamHI as candidate restriction enzyme to construct RAD-seq libraries. Indeed, it showed almost a constant and high number of recovered RE sites at different levels of coverage, compared to the other two REs used to test the technical performance of the novel RAD-seq protocol (Supplementary Table S1). RAD-seq libraries were constructed separately for 95 grapevine samples and were sequenced in two lanes using the 5500 SOLiD™ System. A total of 566M reads 75 bp long were produced (Table 1) with an average of 5,102,500.3 reads per sample. The coefficient of variation (CV) for the number of reads was equal to 33.9 % among samples and 2.5% per sample among lanes. BamHI is a type II restriction endonuclease without methylation sensitivity that recognizes a 6 bp site (5'–GGATCC–3'), cleaving just after the 5'-guanine on each strand. It leaves four base-long sticky ends (GATC-C) whose sequences are equal in color space format to T12320. As shown in Figure 2, 75% of the reads started with a correct T12320 sequence and 11% presented one single color mismatch that we assumed to be a sequencing error. The remaining reads (14%) showed more than one different color at the beginning sequence and were discarded. In order to increase the alignment specificity, the retained reads were pre-processed by replacing the starting sequence with the full BamHI restriction site in color space format (T102320), yielding finally 485M correct reads (76 bp).

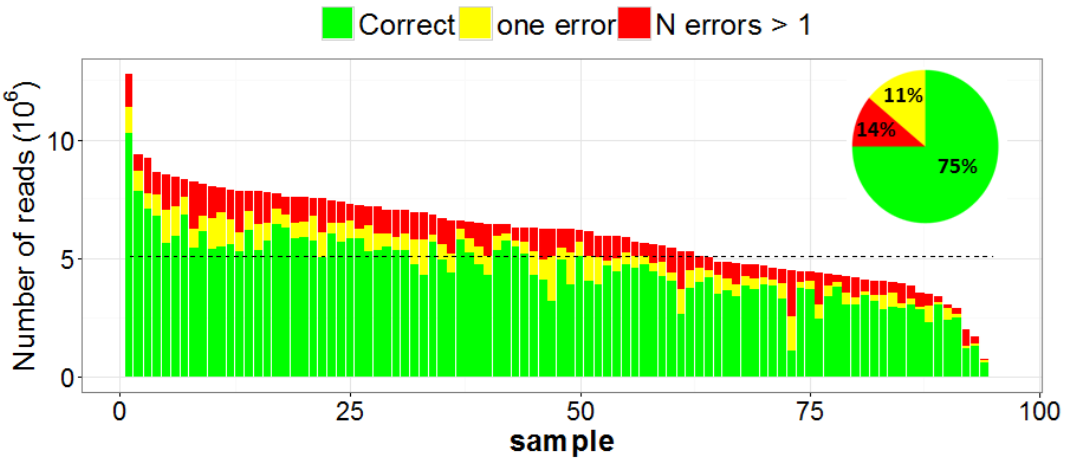

Figure 2: Summary of SOLiD sequencing errors at the starting sequence. Reads per sample with no colors errors (green); reads per sample with one color error (yellow); discarded reads per sample due to color errors higher than one (red). The black dotted lines indicates the average number of reads per sample. Inset shows the percentage of reads with no color errors, one error and more than one error.

Table 1: Number of reads and sequence produced by each filtering step during reads treatment.

Step of reads treatment	Number of reads	Sequence (Gb)
5500 SOLiD™ sequencing	566M	42.4
Pre-processing	485	36.8
Unique alignments	294M	22.3
Unique alignments with MapQ > 10	177M	13.4

Alignment

Pre-processed reads were aligned to the reference 12X grape genome including mtDNA and cpDNA sequences in order to reduce the rate of multiple alignments (Figure 3). 60.3 % unique alignments (Table 1) showed a mapping quality score higher than 10 (177,212,079 over 293,786,586 reads). Among them 8.4 % (14,963,674) accounted for not nuclear alignments.

In silico digestion of the grapevine reference genome with BamHI identified 60,733 putative restriction sites with an average distance of 7.9 kb. We recovered a total of 34K unique restriction sites with at least ten alignments, 93.2% of which were predicted and 6.8% were absent in the reference genome (Table 2). This sequence polymorphism rate at the recognition site may reflect the genetic variability within the investigated germplasm collection, consisting of cultivated and wild forms of grapevine. If we consider the number of recovered restriction sites, the length of a SOLiD read and the assumed presence of two reads going upstream and downstream from each restriction site (Number of covered RE *2*75bp), about 1.1% of the grapevine genome looks resequenced in our study at a high coverage.

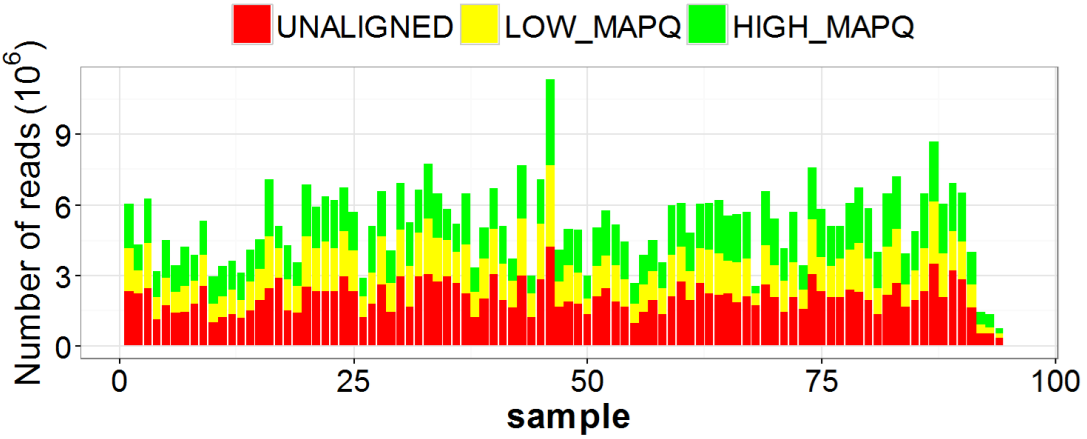

Figure 3: number of alignments per sample. High quality (MapQ > 10) alignments per sample are shown in green, low quality (MapQ < 10) alignments in yellow and unaligned and multiple aligned reads in red.

Table 2: Number of identified *BamHI* recognition sites.

Type Of Restriction Site	Total Number
PREDICTED	32,080
UNPREDICTED	2,353
NOT NUCLEAR PREDICTED	163
NOT NUCLEAR UNPREDICTED	4
Total	**34,600**

The RE sites found in the grapevine PN40024 reference genome though an *in silico* digestion, are called "PREDICTED". The RE site absent in the PN40024 genome are defined "UNPREDICTED". "Not nuclear" RE sites are those identified in mitochondrial and chloroplast DNA sequences.

We considered each up- or downstream read as a RAD locus. We expected that the read depth of each RAD locus would be similar for all the sequenced RE sites if digestion and sequencing were unbiased. However, some RE sites (16.5%) showed a huge difference in read depth among the two adjacent RAD loci. Indeed, those RE sites presented high depth (number of reads aligned to a locus > 10) in more than 80% of the samples at either upstream or downstream RAD loci. The correlation between read depth and the logarithm of restriction fragment length for 69,525 unique RAD loci covered by at least one read was very small ($r = 0.08$; p-value < 2.2e-16). We observed a significantly higher correlation ($r = 0.12$, p-value < 2.2e-16) for RAD loci from restriction fragments shorter than 10 kb (71% of all unique covered RAD loci), while the correlation between read depth and the logarithm of restriction fragment length was not significant ($r = 0.01$, p-value = 0.1458) for RAD loci coming from restriction fragments above 10 kb in length (29% of all unique covered loci).

Variant calling and annotation

Variants on unique high quality alignments were called using UnifiedGenotyper module of Genome Analysis Toolkit (GATK) program [179]. We identified 37,748 SNPs that included 120 variants discovered on mtDNA sequence and 34 SNPs within the cpDNA genome. 93% of markers belonged to the 19 assembled chromosomes with an average of 1.8K SNPs per chromosome (Figure 4A). SNP density ranged from one SNP every 10 kb on chromosome 8 to one SNP every 16 kb on chromosome 19. Finally chromosome size and number of SNPs per chromosome were moderately correlated ($r = 0.68$). We split the reference genome in 985 bins of 500 kb and the number of SNPs per each bin was determined. Thirty five SNPs were present on average per bin. While 3 bins showed zero variants, 655 bins had 10 to 50 SNPs, 83 bins had < 10 SNPs and 244 bins had 51 to 104 SNPs.

According to the grape gene annotation v2.1 more than half of the SNPs fell in intergenic regions. 18,121 SNPs belonged to 6,634 grapevine predicted genes of which 1,680 presented 2,557 nonsynonymous polymorphisms (Figure 4B). We looked for which GO terms of biological process ontology were more represented among the annotated genes which showed sequence variation. An over-representation of metabolism-related functions, referring both to biosynthetic

and catabolic processes, as well as of regulation and transportation mechanisms were observed. Moreover, a small but significant amount of nonsynonymous variants fell in genes related with the detection and response to stimuli such as oxidative and water stresses.

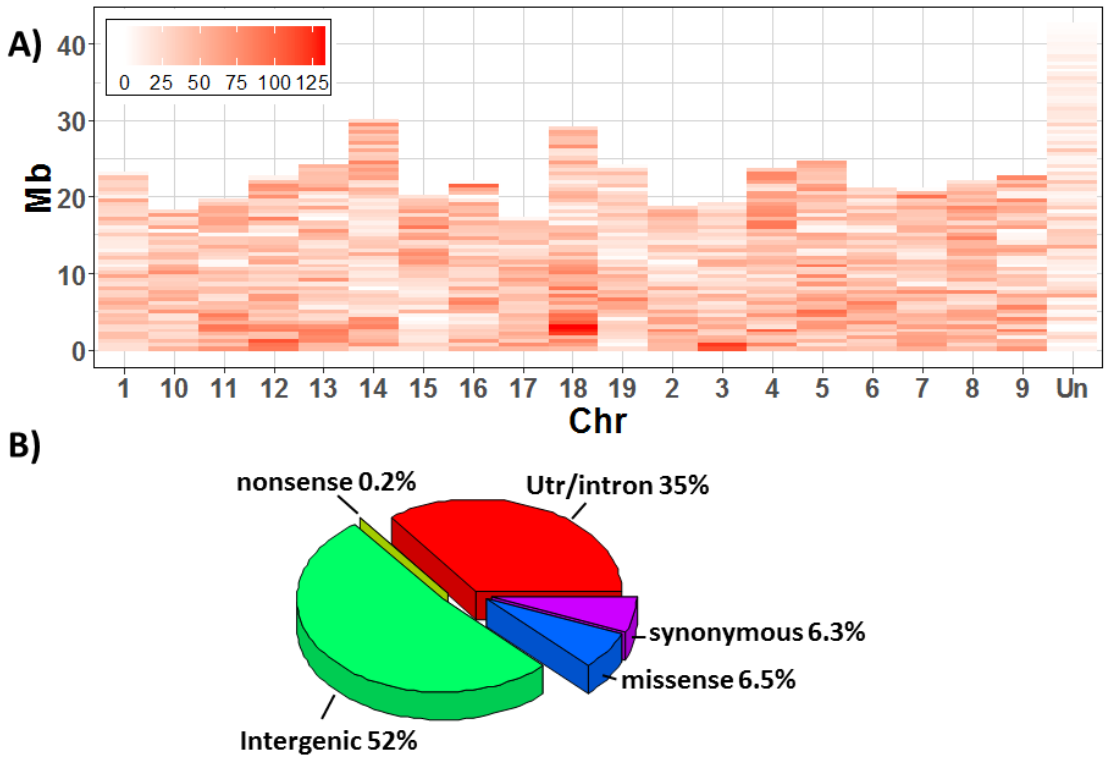

Figure 4: A) SNP density across the 12X grapevine reference genome PN40024. Each block represents a bin of 500 kb. The bar "Un" shows SNPs found on unassembled genomic sequences. **B)** summary of SNPs annotation according to the grape gene annotation v2.1.

SNP validation

Fifty PCR fragments ranging from 266 to 1002 bp were Sanger sequenced on eight grapevine genomic DNA samples in order to validate 183 SNPs discovered by RAD-seq. The validation panel included 4 *V. sativa* and 3 *V. sylvestris* accessions already used to construct the RAD-seq libraries, and one outer *V. sativa* variety. Targeted SNPs included 123 transitions and 60 transversions which were found at 10X coverage in at least 50 libraries. Out of 148 confirmed SNPs, 43.9% perfectly agreed with the RAD-seq data in all the resequenced samples, while 51.3% showed from 1 to 3 different genotypes. The overall rate of fitted genotypes was 86% which may indicate the ability of RAD-seq to accurately determine genotypes in a highly heterozygous species such as grapevine. Moreover, the exceptionally high level of homozygosity of the outer cultivated accession, that was homozygous for 49% of the 312 microsatellite markers tested [186], was proved by 78% of the confirmed SNPs. Nonetheless, a heterozygous profile was still observed for 33 SNPs, highlighting how RAD-seq is able to reveal unknown genetic variability. Our RAD-seq assay sampled 115 SNPs of those included in the commercial GrapeReseq 20K chip. The last had produced a final panel of 16,563 SNPs when applied to our germplasm population. 23% of the common SNPs showed identical genotypes in all 94 samples both using the Illumina chip and

the RAD-seq assays, while 72% differed in 1 to 15 cases bringing the overall rate of fitted genetic profiles among the two different genotyping approaches to 96%.

Discussion

Nowadays several genomic approaches, such as genome-wide association studies (GWAS) and genomic selection (GS), are of potential interest for gene mapping and phenotype prediction of agricultural traits. The application of these methods is still limited in perennial species with high levels of genetic diversity such as grapevine [187]. Indeed, grapevine plants are highly heterozygous (Ho = 0.80; [49]), despite being hermaphroditic self-fertile, likely as a result of selection for fruit production [25, 188]. High heterozygosity is thought to result from the dioecy of wild grapevines and has been maintained in cultivated plants through vegetative propagation from the earliest time of viticulture to preserve favorite genotypes [28]. This high polymorphisms rate and the resulting low LD in grapevine make the increase of marker density throughout the genome fundamental in order to improve the power and resolution of genetic mapping studies to identify significant marker-trait correlations [21]. Here, we applied a novel protocol of RAD-seq to a germplasm collection of wild and cultivated grapevine individuals in order to generate a tool enabling further association mapping and population genetic studies. We obtained 36.8 Gb of sequences, of which over 40% did not align successfully or were mapped in multiple locations on the 12X *V. vinifera* reference genome (Figure 3). This may be due to incomplete assembly of the reference genome or to high levels of genetic variation between the PN40024 and the investigated grapevine accessions. Similar findings have also emerged from the comparison of both “Tannat” and “Sultanina” de-novo assembled grapevine genomes with the reference genome [94, 95]. This can be even more evident in our study since half of the population belongs to the wild Eurasian vine *V. sylvestris* whose genome has not yet been thoroughly investigated. By now it is well accepted that plant genomes contain core sequences that are common to all individuals, as well as dispensable sequences comprising partially shared and non-shared genes that contribute to intraspecific variation [189]. Moreover, the heterozygous cultivar Pinot Noir showed a relevant portion of hemizygous DNA that confirms how the grape genome exists in a dynamic state mediated in part by transposable elements [190]. More than two thousands BamHI restriction sites where identified in our sequences which are absent in the reference genome. The absence/presence of a restriction site could be related to loss/gain of the RE site because of mutations occurring during grapevine evolution and propagation. The predicted restriction sites not recovered by RAD-seq assay could also be explained by imperfect digestion or poor quality reads as well as the presence of RE sites within repetitive sequences, as proved by the moderate percentage of reads discarded during the pre-processing and alignment analysis (Figure 2-3). A considerable level of genetic diversity within the investigated population has been proved by the 37K SNPs discovered, given that half of the investigated population is composed of wild grapevine genotypes which have shown less genetic variability compared to cultivated grapes [25]. This panel exhibited a uniform marker density among chromosomes and significantly higher than those reported for SNPs identified in grapevine through RRL methods [119, 191, 43]. Our SNP set also included a remarkable number of chloroplast SNPs, that can be extremely useful to investigate genetic relatedness among wild and cultivated grapevines and to clarify the process of domestication in grapevine [28]. The RAD-seq survey identifies and scores markers simultaneously in the investigated population, surpassing one of the major limitation of SNP array technologies,

that are often based on genetic diversity discovered in a few resequenced individuals. For instance, the Vitis20K chip comprises 18,071 SNPs discovered within 47 *V. vinifera* genotypes and other 18 *Vitis* species [160]. Out of the *V. vinifera* genotypes just four accessions are *V. sylvestris*, which likely leads to an underestimation of genetic diversity in wild grape populations. The simultaneous discovery and genotyping of SNPs can also increase the number of high-quality markers useful in further analysis. Array-based technologies often fail in SNP genotype call, especially when the discovery panel is evolutionary divergent from the studied accessions [161]. For instance, Myles et al. [21] genotyped 146 grapevine individuals with the Vitis9KSNP chip but just 5,840 SNPs overcomed the SNP genotype quality threshold and were used for the population genetic analysis.

The high number of variants found in less than 1 Mb is further evidence of the high level of heterozygosity in grapevine plants [62, 190]. This high genetic variability can be challenging for genome-wide polymorphisms discovery and genotyping [192]. In RRL approaches restriction site heterozygosity can skew read depth, leading to discarding low coverage RE sites, and it can cause null alleles at flanking SNP loci [193]. Since this bias depends on the size of the sample assayed and on the level of restriction site conservation across the sample, more individuals are sequenced, a larger fraction of variants will be identified. Indeed, sequencing many individuals at low depth has a higher rate of polymorphisms discovery and fair accuracy in genotype inference compared to high coverage sequencing for a few individuals [194]. Our effective sequencing coverage - 1.1% of the genome in 95 wild and cultivated genotypes - has permitted finding about 2% of the expected polymorphisms based on the SNP frequency in whole-sequenced grapevine varieties [94, 95, 96]. Low coverage sequencing may soften the bias of restriction fragment length on RAD loci read depth. Indeed, Davey et al. [195] reported a correlation between restriction fragments length and read depth of RAD loci, which could be related to the shearing step during RAD library preparation, regardless of the shearing technique applied. We found that the bias was significantly lower compared to Davey et al. [195] for RAD loci from restriction fragments below 10 kb. Therefore, a lower distortion of RAD loci read depth, with special regard to those up- and downstream of a heterozygous restriction site, may be expected in our RAD-seq assay. The application of a posteriori filters concerning missing data rate and minor allele frequency per each SNP can handle the implications of restriction site heterozygosity on RAD-seq genotyping.

Given that the coding regions are about 46% of the grapevine genome [100], an interesting result of our study is that 48% of the identified SNPs fell in genic regions, of which the annotated ones are mostly assigned to the functional categories of metabolism and regulation. Actually, plant metabolism is the most represented functional category among the unique set of predicted genes in the grapevine genome [196]. On the other hand, the polymorphisms observed in genes related to both biosynthetic and catabolic processes as well as regulatory or transport functions may reflect different adaptation mechanism among wild and cultivated grapevines.

Conclusions

In this study we underlined the ability of RAD-seq to discover high quality SNPs and add new insights on the level of sequence variation between grapevine genomes. Being the first application of RAD-seq to a germplasm population of grapevine, our findings supply a genome-wide comparison within grapevine species, economically the most important fruit plant in the world [197]. We provided a novel panel of 37K SNPs evenly distributed across the genome that may be useful in future genomic survey regarding the level of differentiation between wild and cultivated grapevines, in order to better explore their genetic relationship. This high-quality SNP data set enables the application of population genetics methods to capture the signals of selection left during the weak domestication process of grapevine and to access the genetic diversity of several *sylvestris* individuals [198]. Moreover, the identification of sequence polymorphisms within genomic regions associated to metabolism and regulation pathways makes our SNP panel rather informative for discovering the genetic mechanisms that contribute to the phenotypic variation associated with domestication traits.

Supplementary Data

Table S1: Number of restriction sites recovered with RAD-Seq on PN40024 genomic DNA using three different restiction enzymes.

	sites covered on both up- and downstream ends, each end having at least x reads					sites covered with at least x reads without considering up- and downstream ends				
	x ≥ 2	x ≥ 4	x ≥ 8	x ≥ 16	x ≥ 24	x ≥ 2	x ≥ 4	x ≥ 8	x ≥ 16	x ≥ 24
HindIII	116,230	102,839	81,696	43,724	19,019	202,522	182,045	165,777	135,674	107,640
BamHI	42,712	38,894	36,375	30,091	23,930	80,861	57,013	53,013	49,462	45,904
NcoI	77,105	68,868	55,893	31,899	14,878	124,160	108,386	108,386	85,190	69,945

Chapter 3

GENOMIC SIGNATURES OF DIFFERENT ADAPTATIONS TO ENVIRONMENTAL STIMULI BETWEEN WILD AND CULTIVATED *V. vinifera*

Abstract

Background: The selective pressure applied by humans to domesticate plants is thought to have reduced the genetic diversity of genes contributing to elected traits. This selection process left genomic signs known as "signatures of selection". While domestication produced crops with high yield and rapid growth, it arguably led to a decrease of plants resilience. Today wild ancestors are considered valuable sources of resilience factors, whose re-discovery can be fundamental for future sustainable agriculture. During domestication, changes in berry size and a transition from dioecious to hermaphrodite plants occured in cultivated grapevines (*V. vinifera* subsp. *sativa*) from its wild form (*V. vinifera* subsp. *sylvestris*). Population genetic analysis can help to clarify how these changes happened and to map genes contributing to adaptive traits in grapevine.

Results: We investigated the genetic diversity of a grapevine germplasm collection composed by 44 *V. sylvestris* and 48 *V. sativa* accessions. We genotyped the whole population using the commercial GrapeReSeq Illumina 20K SNP chip and a novel RAD-seq procedure, obtaining a high density panel of 26K solid polymorphisms. Population genetic structure highlighted a clear separation among wild and cultivated accessions with a low level of admixure. The evaluation of LD extent in the two subgroups showed how LD decayed more slowly in wild grapevines (~20 kb) than within the domesticated subgroup (~10 kb). The F_{ST} metric was evaluated between cultivated and wild accessions along the whole genome. Over two thousand of SNPs showed a significant high value of F_{ST}, validated empirically with permutation test. These loci fall within putative "signatures of selection" that contain genes presumably involved in adaptation during domestication in grapevine. In addition, an overall reduction of nucleotide diversity was observed along the whole genome within *V. sylvestris* accessions, highlighting the small effective population size of wild grapevine. Positive values of Tajima's D were detected in both wild (D ~0.89) and cultivated (D ~1.35) subgroups, probably indicating an ongoing balancing selection.

Conclusions: The application of population genetic methods enabled the discovery of numerous signals of selection, including genes mainly related with the plant response to environmental stimuli. Future studies of functional genomics and/or candidate-gene association mapping will provide additional information about how the two forms of *V. vinifera* react to biotic and abiotic stresses. Finally, this study is further evidence of the broad genetic diversity still present within wild grapevines, which needs to be explored in future breeding programs in view of a sustainable viticulture.

Background

The Eurasian grape (*Vitis vinifera* L.) is one of the most important crop worldwide due to its global distribution and economic value [197]. Nowadays *V. vinifera* L. exists as the cultivated form *V. vinifera* subsp. *sativa* (or *vinifera*) and the wild form *V. vinifera* subsp *sylvestris* which are sometimes referred as two separated subspecies based on morphological differences. However, it can be argued that those differences are likely the results of domestication by humans instead of geographic isolation [25]. Indeed, archeological and historical findings suggest that cultivated grapevines have been domesticated from wild populations of *V. sylvestris* circa 5,500-5,000 BC in the Near East [199], in the region known as Transcaucasia, which still presents a large genetic diversity of grapes [52]. From the primo-domestication sites, there was gradual spread to adjacent regions such as Egypt and Lower Mesopotamia and then further dispersal around the Mediterranean [27]. However, successive genetic analyses raised the outstanding question whether multiple domestication events occurred along the Mediterranean basin [33, 66]. For instance, Arroyo-Garcia et al. [28] suggest the existence of at least two important origins for the cultivated grapevine, one in the Near East and another in the western Mediterranean region. On the other hand, recent studies carried out using 5K SNPs provided further genetic evidence of the Eastern origin of most cultivars as well as the existence of introgressions from wild individuals in Western regions [34]. During domestication, genotypes producing bigger fruits with higher sugar content were selected to ensure greater and more regular yields as well as a better fermentation. In this process, the changes in seed morphology and flower sex were crucial [25]. In particular, *V. vinifera* cultivars generally exhibit hermaphroditic flowers while almost all wild grapevines are dioecious with separate male and female individuals [200].

Many surveys of genetic diversity in grapevine collections have outlined a low but clear differentiation among cultivated and wild accessions by using plastid markers [28, 33], nuclear microsatellites [49] and SNPs [34, 54]. The cultivated grapevine is very diverse, with 6,000-10,000 different varieties believed to exist in the world [69]. This large diversity is mostly the result of sexual reproduction, vegetative propagation and somatic mutations which have been crucial during the long history of grapevine cultivation [25]. On the other hand, *V. sylvestris* is less diverse than the domesticated grapevine. Nowadays relict populations of wild *vinifera* are present with very few individuals. Indeed, the distribution of wild grapevine has drastically been reduced over the last two centuries because of the introduction of pathogens (phylloxera, oidium, mildew) from North America and a fragmentation of wild grapevine habitats by humans [69]. In addition, the level of genetic flow detected between wild and cultivated grapevines may have consequences on the genetic diversity of the small wild populations as introgression, pollution of the gene pool and genetic loss [70]. However, the wild forms still conserve an overall genetic diversity that need to be explored as a putative valuable resource for breeding [85]. Indeed, as in other crops, genetic erosion or loss of variability is occurring in grapevine due to the low number of grown cultivars worldwide that had rapidly displaced old local varieties or landraces [69]. This loss of agrobiodiversity can increase the vulnerability of different cultivars to new environmental changes or the appearance of new pests and diseases [71]. Accordingly, several efforts have been recently devoted to explore the responses to biotic and abiotic stresses in wild *V. vinifera*, revealing tolerant accessions to salt stress [201, 202] and lime-induced chlorosis [203]. Furthermore, *V. sylvestris* was screened for genotypic differences in stilbene accumulation and susceptibility to downy mildew of grapevine (*Plasmopara viticola*), showing how wild accessions with high stilbene inducibility are also less susceptible to infection by *P. viticola* [83]. Whole-

genome comparison of the level of genetic diversity between wild and cultivated individuals is an alternative approach for discovering the genes and genetic mechanisms involved in the domestication process and in the local adaptation to different environmental changes [126]. Indeed, the selection pressure may have shaped the pattern of variation across the genome, leading to a population-wide reduction in genetic diversity of genes contributing to selected traits [114]. These reductions are well defined 'signatures of selection' and persist until recombination and mutation restore diversity at the selected loci in the population [204]. Genomic insights of selection have been reported for several crops, such as tomato [150], maize [205], rice [206] and barrel medic [207].

In this regard, we evaluated the genetic diversity of a grapevine germplasm collection composed of cultivated and wild *V. vinifera* by using a panel of 26K SNPs. The F_{ST} analysis disclosed a significant high level of differentiation between the two subspecies at several genomic regions which include genes mainly involved in primary metabolism and in the response to environmental stimuli. We provide further evidence that wild grapevines represent a valuable source of resilience factors whose re-discovery and re-introduction in cultivars can be fundamental for future sustainable agriculture.

Materials

Plant material

A germplasm collection of 48 cultivated (*Vitis vinifera* spp. *sativa*) and 44 wild (*Vitis vinifera* spp. *sylvestris*) grapevines (Appendix B, page 125) was sorted at the FEM grape repository (ITA362), located in San Michele all'Adige, Italy (46°18' N, 11°13' E). The sativa cultivars were chosen in order to maximize the genetic diversity based on a set of 22 SSR and 384 SNPs markers in the source collection [49]. The wild individuals, mostly from the Italian Peninsula, were selected within 110 different *sylvestris* genotypes, according to a cluster analysis performed with a model based approach as implemented in STRUCTURE [49]. Young leaf tissue of one field grown plant per accession was harvested and stored immediately in sterile tubes at -80°C for DNA extraction and successive analysis. The total genomic DNA was isolated from freeze-dried tissue after grinding with the MM 300 Mixer Mill system (Retsch., Germany). DNA extraction was performed using the DNeasy 96 plant mini kit (QIAGEN, Germany). DNA concentration and purity were inspected using both the Synergy HT Multi-Mode Microplate Reader (BioTek) and the NanoDrop 8000 UV-Vis Spectrophotometers (Thermo Scientific). DNA samples were also checked for quality with gel electrophoresis.

Genotyping with the GrapeReseq 20K SNPs array

DNA samples were adjusted to a minimum concentration of 100 ng/µL in 10 µL aliquots. The commercial GrapeReseq 20K SNPs array [184] was used to genotype the whole population with the Infinium technology according to the Illumina protocol (Illumina, Inc., San Diego, CA, USA). The genomic DNA of the Pinot Noir cultivar was used as control. SNPs genotypes were scored using the Genotyping Module v1.9 of the Illumina GenomeStudio Data Analysis software. SNPs with a Call Freq score 0 and a GenTrain score < 0.6 were filtered out. Markers with a Cluster Sep score < 0.4 were visually inspected for accuracy of the SNP calling. SNPs with R mean score > 0.3 and with clusters not overlapped were retained.

RAD-Seq assay

Restriction associated DNA sequencing (RAD-Seq) libraries were constructed using the method described in Chapter 2. Briefly, DNA for each sample were digested with BamHI enzyme and ligated to a P1 biotynilated adapter. After random shearing, biotynilated fragments were captured using Dynabeads® MyOne™ Streptavidin C1, and end-repaired. Standard barcoded P2 adapters of 5500 SOLiD Fragment libraries were then ligated to the biotinylated samples. Afterwards each library was amplified and purified before fragments sequencing on 5500 SOLiD™ System (Applied Biosystems, Life Technologies). The raw data produced were filtered to remove low quality reads (mapping quality <10). The clean data were analyzed with the UnifiedGenotyper tool of the Genome Analysis Toolkit (GATK) v3.2-2 [179], and SNPs genotypes for each sample were inferred through the Bayesian genotyper implemented in GATK.

SNP filtering

The two SNP data sets obtained with the 20K Illumina chip and the RAD-seq assays were merged in a unique panel. For the SNPs in common between the RAD-seq and the 20K Illumina chip we retained only the SNP profiles of the latter. Samples and SNPs with a missing rate > 0.2 were filtered out. Genotype imputation was performed to fill in the missing data using LinkImpute v1.1.1 software, which is based on a k-nearest neighbour genotype imputation method (LD-kNNi) designed to work with unordered markers [169]. Finally SNPs with a minor allele frequency (MAF) lower than 0.05 were removed using Plink v1.9 software [208, 209].

Analysis of population structure

The genetic structure of the germplasm population was analyzed with fastSTRUCTURE software v1.0 [210], which uses a variational Bayesian framework for approximate inference of subpopulations [211]. A number of ancestral genetic groups (K), ranging from 1 to 10, was tested by 10 independent iterations for each K. The most likely K value was chosen running the algorithm for multiple choices of K and by plotting the marginal likelihood of the data. The software CLUMPP v1.1.2 [212] was used to find optimal alignments of the independent runs and the output was used directly as input into the program for cluster visualization DISTRUCT v1.1 [213]. Moreover, a Principal Component Analysis (PCA) was performed as implemented in 'adegenet' [214] R package for the multivariate analysis of genetic markers.

LD decay

Linkage disequilibrium (LD) was estimated between all SNPs with a MAF > 5% in the whole germplasm population and within *sativa* and *sylvestris* subgroups separately by using Plink v1.9 software [209]. The classical r^2 estimate of correlation between genotypes was used [109]. LD decay was explored by plotting the median r^2 in sequential bins of 10 kb against physical position. Moreover, LD landscape of each chromosome was also inspected through heat-map visualization with the software Haploview v4.1 [215].

Genomic differentiation between sativa and sylvestris genotypes

Since F_{ST} is often applied to evaluate the degree of population differentiation [147, 216], F_{ST} was measured between *sativa* and *sylvestris* accessions with VCFtools v0.1.13 [180], by using sliding windows of 100 kb with a step size of 10 kb. Genomic windows with the top 5% of F_{ST} values were selected as candidate regions for further analysis. In order to verify the empirical cutoff with low false discovery rate, we performed whole-genome permutation tests to ascertain the thresholds for identifying genomic regions highly differentiated between the two grapevine subgroups. In particular, all the accession genotypes of *sativa* and *sylvestris* were shuffled and then F_{ST} analysis was performed with the same parameters 1,000 times. To better interpret the results gained with the F_{ST} analysis and to clarify how *sativa* and *sylvestris* genotypes are

differentiated, nucleotide diversity (π) and Tajima's D [217] were estimated along the whole genome in 100-kb windows with a step size of 10 kb using VCFtools.

Functional Genes Annotation

The grape gene annotation v2.1 hosted on http://genomes.cribi.unipd.it/grape/ [100] was used to investigate the putative functions of genes present in the genomic regions with the top 5% of F_{ST} values. In particular, the distribution of the identified genes into different biological processes was evaluated using the weight01 method provided by the R package topGO [218]. The Kolmogorov-Smirnov like test was performed to assess the significance of over-representation of GO categories compared with all genes in the grapevine gene prediction. In addition, the differentiation in the genomic regions reported in the literature as associated to flower and fruit traits was checked.

Results and discussion

Analysis of population structure

A total of 92 grapevine *sylvestris* and *sativa* accessions were genotyped using the custom Vitis20K SNP array and a novel RAD-seq approach (see Chapter 2). We marged the two SNPs matrices in a unique panel, since they showed the same distribution of allele frequency and linkage disequilibrium (LD) pattern. After removing low quality loci, the filtered merged data set counted 54,157 SNPs (Table 1). Six samples (Appendix B, page 125) and 22,258 markers were removed because of a missing rate > 0.2. As showed in Table 1, the higher percentage of missing data produced by the RAD-seq assay could be related with several technical factors that led all sequenced regions to not be evenly covered in all individuals of the population [156]. After imputing the missing genotypes, SNPs with a minor allele frequency (MAF) < 0.05 were removed gaining a final panel of 26,893 SNPs with an average of 1.3K SNPs per chromosome. 70% of the SNPs with a MAF < 0.05 came from the genotyping assay with the Vitis20K array. These SNPs probably resulted from some errors in genotype calling and represent an underestimation of the real genetic diversity within the investigated populations, which is a well known bias of array-based technologies [161]. SNP density ranged from one SNP every 15 kb on chr8 to one SNP every 21 kb on chr19. Chromosome size and number of SNPs per chromosome were strongly correlated (r = 0.87).

Table 1: Summary of SNPs filtering after population genotyping assays with the Vitis20K Illumina chip and RAD-seq approaches.

Genotyping technology	Initial N° of SNPs	N° of SNPs with missing rate > 0.2	N°of SNPs with MAF < 0.05	Final Number of SNPs
Vitis20K	16,563	338	3,600	12,625
RAD-seq	37,594	21,920	1,330	14,268
Total	54,157	22,258	4,930	26,893

We used this SNPs panel to investigate the population structure and visualize the relationships among individual accessions applying two different approaches. Principal Component Analysis (PCA) was first performed and Figure 1 shows the first two principal components (PCs) which accounted for the 21% of the total variation. PC1 clearly differentiates sylvestris genotypes from cultivated varieties, whereas PC2 reflects the variability within *sativa* accessions.

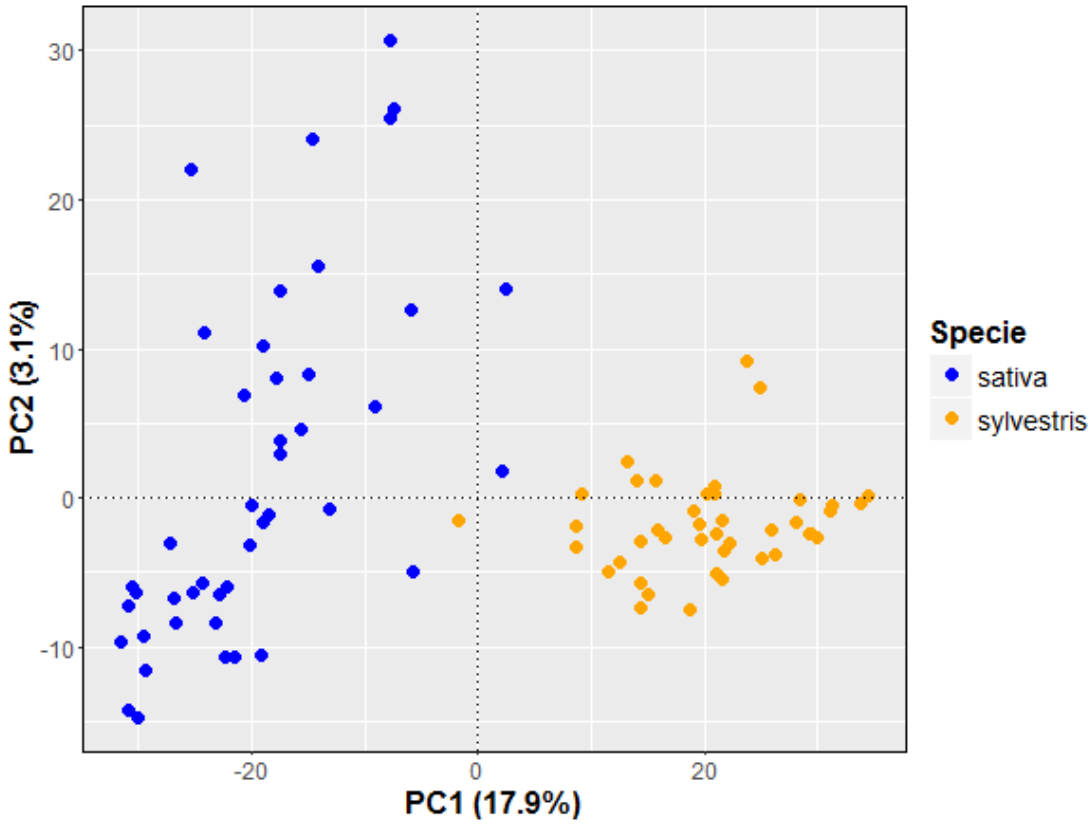

Figure 1: Visualization of the genetic relationships among wild and cultivated *vinifera* by their projection onto the first two PC axes. Along each axis the proportion of the total variance accounted by each PC is shown in parentheses.

To better understand the genetic structure of the analyzed germplasm collection, the clustering algorithm implemented in fastSTRUCTURE software [210] was used by exploring different possible numbers of subpopulations (Figure 2). The optimal number of subgroups was three: 81% of the individuals showed a clear assignment (membership likelihood > 0.75 %) to a cluster (Supplementary Table S1). Two major groups included 28 *sativa* accessions and 36 *sylvestris* individuals respectively, while Pinot Noir, Gewurtztraminer and Mornan Noir cultivars clustered together in a third separated group. A first-degree relationship of Pinot Noir and Traminer has already been suggested by previous studies with microsatellite markers (SSRs) [87]. Moreover, Pinot Noir and Traminer have presumably ancient origins and many moderns cultivars are their first-degree relatives [219]. Probably these two cultivars could have arisen from hybridization between Roman grapes and local wild populations or from secondary domestication of the latter. Indeed, many of the 19 genotypes (13 *sativa* and 6 *sylvestris*) not clearly assigned to a defined group by fastSTRUCTURE exhibited admixture with this small cluster (K2, Supplementary Table S2). However, the analysis of population structure highlighted how the *sativa* and *sylvestris*

individuals were well distinguished as two separated groups with a low level of admixture. This result is consistent with previous reports based SSR and SNP genetic profiles, that have shown clear distinctions between wild and cultivated individuals [49, 28, 54]. Moreover, the low complex pattern of admixture observed between *sativa* and *sylvestris* accessions may agree with the hypothesis in which grapevine domestication took place in a single location from a restricted pool of wild genotypes, followed by the spread of cultivars to other regions where likely introgressions from local *sylvestris* occurred. However, we used sylvestris individuals mainly from the Italian Peninsula and already clustered through a hierarchical STRUCTURE analysis [49]. At the same time, *sativa* accessions were selected from a core collection that maximize the genetic diversity present in the whole germplasm collection. Therefore, biases in allele frequencies may have been introduced, leading to an underestimation of the real level of admixture between the two subspecies.

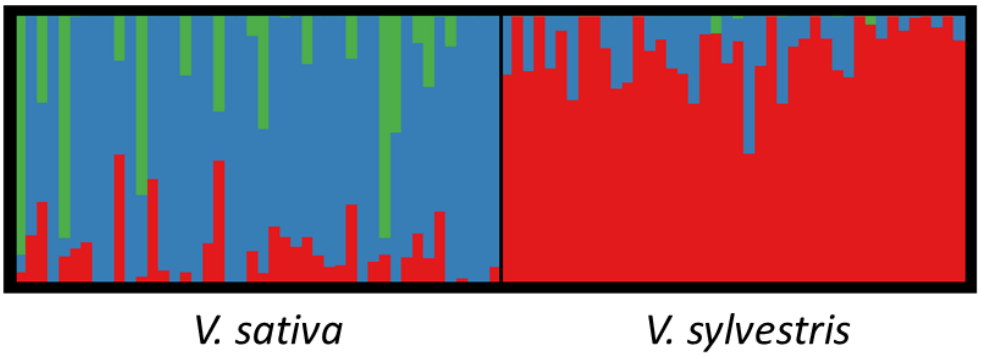

Figure 2: Barplot of admixture proportions of wild and cultivated subpopulations, as measured by fastSTRUCTURE at K = 3. Each individual is represented as a vertical bar, reflecting assignment probabilities to each of the three groups. K1: red bars; K2: green bars; K3: blue bars.

Estimation of Linkage Disequilibrium

To estimate the level of LD along the whole genome, pairwise analysis between all SNPs with a MAF > 5% was used. LD, as measured by the classical r^2 correlation coefficient [109], decayed below 0.2 within 10 kb (Figure 3a). Such rapid LD decay is consistent with the results of Myles et al. [34], which detected low level of LD (r^2 < 0.2) at short physical distances using the Vitis9K SNP array. An even lower level of LD was observed by Lijavetzky et al. [62], which found in more than 200 gene sequences a decay of r^2 within 100-200 bp. On the other hand, Nicolas et al. [113] observed that the decay of LD down to 0.2 ranged from 9 to 458 kb. These discrepancies may be related to the low number of genomic regions investigated in both LD surveys [62, 113] compared to our genome-wide analysis of LD. However, we confirmed the evidence of a rapid LD decay in grapevine, which is in agreement with the high polymorphic rate of the grapevine genome [89]. When analyzed separately in the two subpopulations, the decay of LD appears quite different between wild and domesticated grapes (Figure 3b). In particular, a slower LD decay was observed within the *sylvestris* group, where r^2 reached values below 0.2 within 20 kb. This results is in contrast with previous reports on LD decay between *sativa* and *sylvestris*, where it appeared unchanged among the two subspecies [34, 113] or slower in the cultivated data set [220]. This discrepancy is not surprising since LD extent can vary according to different factors, such as

mating system, natural and artificial selection, the population under investigation and its mating history [107]. The longer extent of LD in wild grapevine can be related with an elevated level of inbreeding linked to a small effective population size and the absence of gene flow between wild populations [70, 66]. Indeed, no structure was identified within the *sylvestris* group by the above analysis of population structure, confirming the close genetic relationship between wild individuals. Furthermore, the mainly Italian origin of our wild accessions is limited compared to the large geographic area of the wild grapevine form. The differences of LD extent between *sativa* and *sylvestris* accessions were more evident when LD patterns per each chromosomes were compared. In particular, long-range LD (LRLD) between loci that are widely separated on chromosome (distance > 1Mb) was observed for almost all the chromosomes of the *sylvestris* group, especially on chromosomes 2, 4, 8, 13, 15 and 18 (Supplementary Figure S2). Findings of LRLD suggest that some forces are acting, such as population admixture, genetic drift, epistatic selection, hitchhiking with positive-selected mutation or structural variation in chromosomes [221]. Blocks of short-range LD were also observed within the *sativa* on chromosomes 2, 6, 17 and 18 (Supplementary Figure S1). QTLs associated with important traits in grapevine have been detected on these chromosomes, such as those for flower sex and berry skin color on chr2 [222, 223, 224, 200, 123], and berry weight on chr17 and chr18 [42].

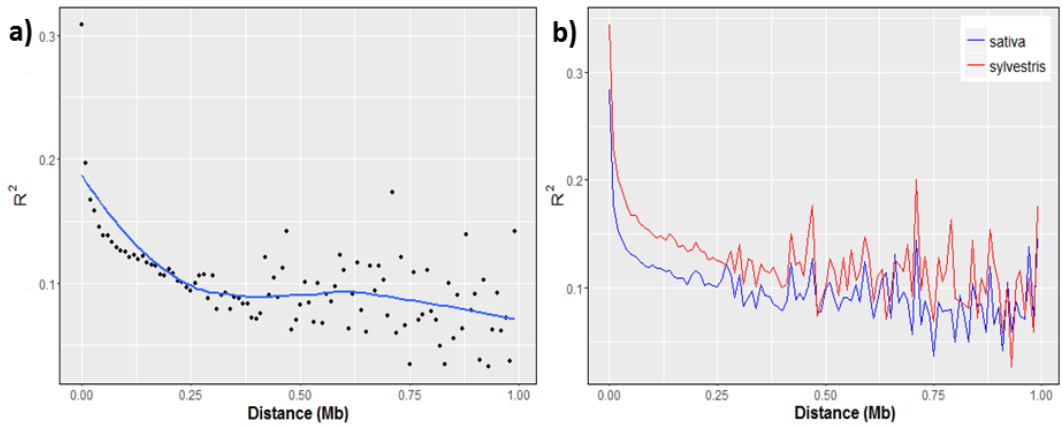

Figure 3: Decay of LD **(a)** in the whole population and **(b)** in sativa and sylvestris separately. Each point represents the median r^2 value in sequential bins of 10 kb against physical position.

Genomic differentiation between sativa and sylvestris genotypes

Since the analysis of population structure underlined a clear separation between sativa and sylvestris accessions, population differentiation statistic (F_{ST}) was computed across the grapevine genome in order to identify genomic regions with altered allele frequency among the two *V. vinifera* subspecies. The overall level of genetic differentiation between cultivated and wild grapes was moderate (F_{ST} =0.12). A similar genetic divergence was identified between Western European cultivars and wild genotypes [34] as well as among grapevine accessions of *sativa* and *sylvestris* from Spain [66] and Morocco [225]. This low level of genetic differentiation suggests the existence of genetic exchange between cultivated and wild individuals, supporting the hypothesis that the introgression from local wild *sylvestris* has played an important role during grapevine domestication. However, a non-random distribution of divergent sites was observed along the whole genome: the top 5% had a F_{ST} > 0.27 and no positive signals was found to pass this

empirical cutoffs after permutation test (Supplementary Figure S3). 2,461 SNPs were included in 2,001 windows identified as significantly differentiated between *sativa* and *sylvestris* individuals. All 19 chromosomes of the grapevine genome showed divergent sites, ranging from chr12 with 14 windows to chr4 with 382 bins (Figure 4a). In particular, the genomic region spanning for 7.5 Mb at the beginning of chr17 has already been identified as a putative candidate domestication locus in previous studies [34, 113]. A shift in the distribution of alleles in populations may result from a sweep toward fixation of a selected locus and its nearby hitchhikers [114]. This sweep causes a population-wide reduction in the genetic diversity around the selected locus. Therefore, nucleotide diversity [151] was evaluated across the grapevine genome in *sativa* and *sylvestris* groups separately. Nucleotide diversity measured by the π value was slightly higher for the *sativa* group (2.34×10^{-5}) than that for the *sylvestris* (2.00×10^{-5}) group. Several surveys in grapevine germplasm collection consisting of both cultivated and wild *V. vinifera* accessions underlined this overall lower genetic diversity in the wild gene pool compared to the cultivated panel [28, 226, 227]. Indeed cultivated grape has a big effective population size planted over multiple locations, where sexual crossing and somatic mutations coupled with a massive vegetative propagation have been the main driving forces during grapevine evolution, accumulating and increasing genetic variability in cultivated grapevine. This high level of diversity in cultivated *V. vinifera* may also arose from multiple domestication events [28, 33] through hybridizations with wild individuals [70]. Furthermore, our selection of *sativa* accessions from a core collection may overestimate the real level of nucleotide diversity in cultivated grapevines. On the other hand, the wild relatives are nowadays present in low number in isolated populations [228]. In addition, the anthropogenic pressure on natural habitats and disease-causing agents introduced in Europe from North America at the end of the 19th century may also explain the progressive decrease of nucleotide diversity in wild populations [25]. As showed in Figure 4b, the average value of the ratio $\pi_{sylvestris}/ \pi_{sativa}$ was 0.89, confirming that π is higher in cultivated grapevine in most of the investigated genomic regions. In particular, a drastic reduction in nucleotide diversity of *sylvestris* individuals ($\pi_{sylvestris}/ \pi_{sativa} = 0$) was observed on chromosomes 5, 14 and 15 at genomic regions with a total of 6 SNPs monomorphic in the *sylvestris*. At the same time, a reduction in nucleotide diversity of the *sativa* was observed on chromosomes 5, 12, and 19, where $\pi_{sylvestris}/ \pi_{sativa}$ had values higher than 10. However, while the reduction of genetic diversity in cultivated grapevine on chr19 was associated with a significant differentiation ($F_{ST} = 0.32$) between *sativa* and *sylvestris* group, no divergence in allele frequencies was observed for the other genomic regions with extreme values of $\pi_{sylvestris}/ \pi_{sativa}$. Indeed, both cultivated and wild individuals showed low minor allele frequency at those loci (MAF < 0.1). Therefore, this common reduction in nucleotide diversity in both subspecies may suggest reciprocal introgressions between wild and cultivated grapes [70] or could reflect local conditions affecting diversity in both populations [154]. Another common test used to detect signals of selection as distortion of allele frequency and nucleotide diversity is the Tajima's D, which compares the number of pairwise differences between individuals with the total number of segregating polymorphisms [151, 217]. We observed mostly positive values of Tajima's D in both wild (D_{syl} ~0.89) and cultivated (D_{sat} ~1.35) subgroups. As reported by Riahi et al. [227], a positive value of Tajima's D, especially for cultivated accessions, may indicate an excess of intermediate frequency alleles in these populations. Such configuration of allele frequencies may arose by a balancing selection, which maintains both alleles at the selected loci [229]. This may happen as the result of an heterozygote advantage as well as frequency-dependent selection or spatial and temporal habitat heterogeneity [143]. A balancing

selection is in line with the high heterozygosity of grapevine genome and with the heterogeneity of uses and habitats to which *V. vinifera* is adapted.

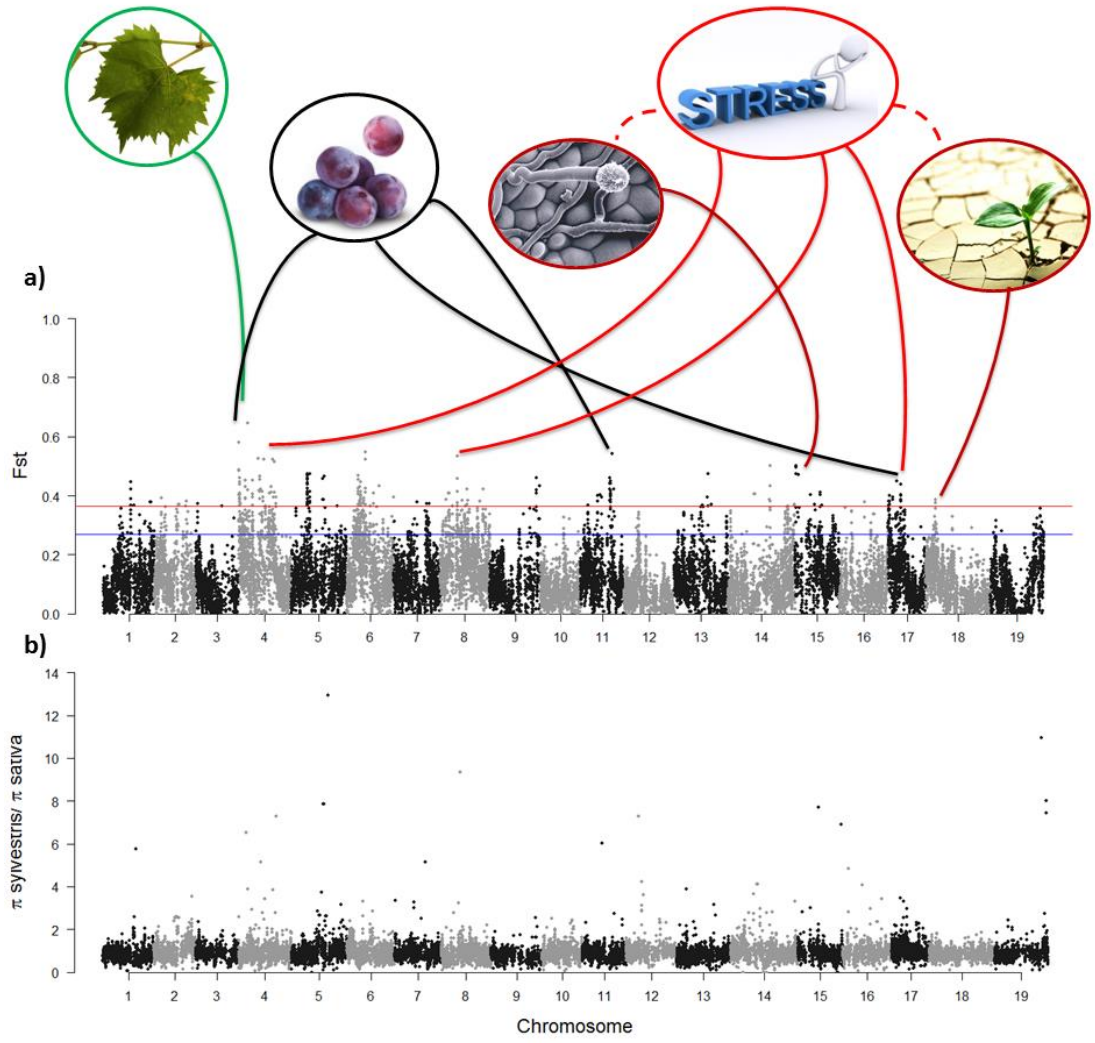

Figure 4: a) Manhattan plot of F_{ST} values for all SNP sites between cultivars and wild grapevines. The horizontal blue and red lines indicate respectively the 95th (F_{ST} = 0.27) and the 99th (F_{ST} = 0.37) percentiles of the F_{ST} empirical distribution. The circles reported the putative functions and the related metabolic processes of the genes with the highest F_{ST} values in the enriched functional classes. **b)** Reduction in nucleotide diversity in the comparison of *sylvestris* and *sativa* accessions ($\pi_{sylvestris}/\pi_{sativa}$) across the genome.

Identification of biological functions underlying sweep

We looked at the new gene prediction v2.1 of the grapevine genome within windows of 20 kb around the SNPs detected as putatively under selection. Out of the 2,032 predicted genes found in LD with the most significant SNPs 1,714 were annotated. Twelve functional classes were significantly enriched in the list of differentiated genes (Table 2), accounting for 109 of them (Supplementary Table S3). 69% of these genes had a predicted function related to organic compound metabolic processes, especially those of nitrogen and carbohydrate, while the 24%

was assigned to functional classes involved in perception, response and/or adaptation to environmental stimuli.

Table 2: Functional Classes significantly differentiated between *sativa* and *sylvestris* accessions.

GO ID	Term	Annotated genes	Significant genes	P-value
GO:0071704	organic substance metabolic process	1516	33	0.01596
GO:0006807	nitrogen compound metabolic process	604	32	0.01372
GO:0005975	carbohydrate metabolic process	148	10	0.00019
GO:0055114	oxidation-reduction process	143	9	0.00262
GO:0009737	response to abscisic acid	114	8	0.00232
GO:0006952	defense response	446	3	0.03388
GO:0032259	methylation	72	5	0.01715
GO:0009607	response to biotic stimulus	124	3	0.00045
GO:0009651	response to salt stress	50	2	0.01213
GO:0010363	regulation of plant-type hypersensitive response	20	2	0.0378
GO:0010118	stomatal movement	9	1	0.00899
GO:0090305	nucleic acid phosphodiester bond hydrolysis	11	1	0.03897

Out of the 109 genes in the enriched classes 14 showed a F_{ST} values > 0.37 (99th percentile of the F_{ST} empirical distribution; Supplementary Table S3). Therefore, understanding the putative functions and the related methabolic processes of these genes has a particular relevance in the genomic comparison between *sativa* and *sylvestris* (Figure 4). At the top of the genes list showing highest value of F_{ST} we identified the 'RPL5B' gene (VIT_204s0008g00050; Table S3), which codifies the 60S ribosomal protein L5-2 [230]. This gene could illustrate differences in organ development and expansion between the two subspecies. Indeed, the angusta3 (ang3) mutant of *A. thaliana* for RPL5b gene displayed altered growth and development of several organs, notably of leaves [230, 231]. Therefore, it is likely that balancing selection (D_{sat} = 1.13; D_{syl} = 1.37) has acted to promote the strong morphological variation observable today about leaf shape and size within and between cultivated and wild grapevines. Indeed, the former has palmate-lobed leaves with a huge variability regarding size, shape and hirsuteness [45], while the latter presents hairy leaves with small to medium size [77, 47]. Different climatic conditions, such as radiation and precipitation of certain geographic regions, could have caused the current variation in leaf shape and size [232]. A particular enrichment in genes with a role in the carbohydrate metabolic process has been observed in the list of genes with a significant differentiation between wild and cultivated populations (Table 2). The identification of the soluble starch synthase IV-1 gene (SS4; VIT_211s0065g00150; F_{ST} = 0.4) highlighted differences between the two subspecies in starch and

sucrose metabolism which is relevant for berry development [233]. Indeed, starch concentrations decline significantly during the ripening and maturation phase of berry growth for the production of hexose sugars, essential for flesh berry sweetness and fermentation [234]. In addition, we identified a nuclear transport factor 2 (NTF2) gene (VIT_217s0000g05240) which is involved in organelle-nucleus communication and has a predicted role in the response to abscissic acid (ABA). ABA is the plant hormone that promotes the ripening of non-climateric fruits and is associated with the plant's response to different kinds of abiotic stresses such as drought, high temperature, chilling and salinity [235]. During grape ripening an increase in free ABA levels around véraison accompanies sugar accumulation, softening and anthocyanin synthesis [236]. NTF2 gene is located within the significant signature of selection on chr17, which included candidate domestication-loci for berry size and development [34]. A reduction of nucleotide diversity in *sativa* accessions ($\pi_{sylvestris}/ \pi_{sativa} = 1.23$) was observed at this locus, supporting the evidence of a putative selection for berry composition and ripening traits in cultivated grapevines [34]. Another diversified gene involved in the carbohydrate metabolic process is the NADP-isocitrate dehydrogenase gene (cICDH; VIT_204s0079g00530), which catalyzes the oxidative decarboxylation of isocitrate. An up-regulation of the genes encoding isocitrate dehydrogenases in tobacco (*Nicotiana tabacum* cv Xanthi) and grape (*V. vinifera* cv Sultanina) accompanied the increased aminating activity of glutamate dehydrogenase (GDH) under stress conditions, such as salinity, thanks to the signaling function of reactive oxygen species (ROS) [237]. Other loci involved in the response to different environmental stimuli were identified among the most differentiated genes between *sativa* and *sylvestris*. Indeed, the 10 kda chaperonin gene (CPN10; VIT_208s0040g01150) encodes the plant mitochondrial homologue of GroES or chaperonin 10 (CPN10) in *E.coli* [238]. It is well known that the essential function of molecular chaperonins is to prevent the formation of 'improper' protein structures, which may occur during the exposure to stresses such as heat shock [239]. The 'LPA66' gene (VIT_204s0008g00480) encodes for a pentatricopeptide repeat (PPR)-containing protein, which is RNA binding and is involved in post-transcriptional processes [240], including RNA editing. The *Arabidopsis* high chlorophyll fluorescence mutant low psII accumulation66 (lpa66) had impaired PSII functions resulting in the high chlorophyll fluorescence phenotype [241]. In the top 1% of the Fst empirical distribution we found also the gene VIT_204s0008g01360, which encodes a U-box domain-containing protein 35-like. The Plant U-box (PUB) proteins have a ubiquitin protein ligase activity during protein ubiquitination [242]. The ubiquitin machinery is involved in responses to changes in abiotic or biotic environment by chromatin modification and transcription factor modulation, cell surface receptor localization and/or stability, and by controlling key enzymes in metabolic pathways [243]. In addition, the FATB genes (VIT_217s0000g01100) encodes the myristoyl-acyl carrier protein thioesterase which plays an essential role in chain termination during de novo fatty acid synthesis. Analyses on Arabidopsis mutants of FATB thioesterase revealed the crucial role of this enzyme in seed development and viability [244] as well as in the promotion of the hypersensitive response (HR) to pathogen attack [245]. Furthermore, the rhomboid-like protein 11 gene (RBL11; VIT_204s0008g03830) encodes a transmembrane serine protease which modulates several cellular processes in different biological contexts, such as cell signaling [246], through the regulated intramembrane proteolysis (RIP). The presence of the desacetoxyvindoline 4-hydroxylase gene (VIT_204s0008g01360) among the genomic regions with the highest F_{ST} values supports the evidence of different adaptive response to environment changes between *sativa* and *sylvestris* genotypes. The desacetoxyvindoline 4-hydroxylase is involved in the biosynthesis and regulation of terpenoid indole alkaloids [247], secondary metabolites which provide protection

against microbial infection, herbivores consumption, and abiotic environmental stresses [248, 249]. Differences in allele frequencies were also observed at the ERF2 transcription factor (VIT_215s0021g01590) and 'RAP2' (VIT_218s0001g05250) genes, which encode respectively the plant transcription factors ERF2 and RAP2, two members of the APETALA 2/ethylene-responsive element binding factor (AP2/ERF) family [250]. ERF proteins have been identified as ehylene-responsive element (GCC box)-binding protein [251]. In tobacco, the GCC box has been found in the promoter of various defense genes and has been shown to function as a cis-acting element responsive to ethylene and elicitors [252]. RAP2 is a dehydration-responsive element-binding protein (DREB) with a role in plant abiotic stress responses, such as high-salt stress, water deficit and extreme temperatures [250]. Furthermore, the gene VIT_204s0008g03840 codifies for an ankyrin repeat-containing protein. Several studies have elucidated the regulatory function of ankyrin repeat proteins during plant growth and development stages as well as during stress conditions, such as drought stress and pathogen attack [253]. Finally, the identification of a splicing factor 3b subunit 1-like gene (VIT_208s0040g00270) supports that alternative splicing may contributes to the evolutionary adaptation. Indeed, the assortment of different protein isoforms can be quickly modified as a response to a sudden and strong selective pressure [100].

In almost all the identified genes involved in stress responses a slightly reduction in nucleotide diversity was observed in the *sylvestris* ($\pi_{sylvestris}$/ π_{sativa} ~0.95), associated with a positive value of the Tajima's D (D_{syl} = 1.41). These results imply that a balancing selection is likely acting in wild populations for adaptation to several environmental changes which may occur in their natural habitat along river banks, also as a consequence of human action that has disrupted the original environment of wild grapevine populations [25]. Our results are in line with recent studies on the tolerance of *sylvestris* genotypes to different stress conditions such as pathogen attack [83] or calcareous soils [203]. Therefore, *sylvestris* grapevines represent valuable resources to mine for resilience genes or alleles which may have been lost during the domestication process, making cultivated grapevine dependent to agricultural means such as fertilization, irrigation, weeding, and chemical plant protection. The CPN10 and RAP2 genes represent an exception of this trend. Lower genetic diversity was observed at these loci in *sativa* accessions, suggesting a putative ongoing selection for adaptive mechanisms to salt stress.

In addition to the GO enrichment analysis we looked for genes identified in previous QTL mapping studies as associated to main agronomic traits in grapevine, such as berry weight, berry skin color and flower sex (Table 3). Indeed, a large difference in berry size can be observed between wild and cultivated genotypes [25]. The wild *V. sylvestris* produces mature berries weighting less than 1 g, while berries of some table grape varieties can weight 10 g or more [228]. We found several genes of those reported in literature under berry weight QTLs [42], such as the genes for the xyloglucan endotransglycosylase (XTH;VIT_201s0150g00460) [254], the histone deacetylase 2C (VvHD2C; VIT_206s0061g01240) [255] and the cytochrome p450 78a3-like (CYP78A10; VIT_217s0000g05110), which has been found to regulate fruit size during tomato domestication [256, 42, 113]. One of the main changes likely occurred during grapevine domestication affected flower sex [25]. Common microsatellite (SSR) loci relatively close to the sex locus have been reported on chr2 [120, 224, 223, 200]. We found a genomic region spanning from 4.7 to 5.0 Mb on chr2 with a high level of differentiation between cultivated and wild accessions (F_{ST} ~ 0.31). This genomic region included 4 SNPs in LD with the APT [200], SNP4AC and Vvib23 [222] markers of flower sex.

Table 3: Genes reported in literature under QTLs for berry weight, flower sex and berry skin color, and identified in this study as significant differentiated between wild and cultivated grapevines.

Gene ID	Chr	Position	Gene Annotation	Trait	F_{ST}	Reference
VIT_201s0150g00460	1	22826079: 22829099	xyloglucan endotransglycosylase (XTH5)	Berry weight	0.28	[254]
VIT_206s0061g01240	6	19041829: 19047502	histone deacetylase 2C (VvHD2C)	Berry weight	0.26	[255]
VIT_217s0000g05110	17	5600225: 5602640	cytochrome p450 78a3-like (CYP78A10)	Berry weight	0.32	[256][42][113]
VIT_211s0016g04630	11	3959481: 3961177	DELLA protein SLR1-like (GAI)	Berry weight	0.27	[42]
VIT_218s0001g14000	18	12002927: 12003389	auxin-induced protein X10A-like	Berry weight	0.29	[42]
VIT_218s0001g14030	18	12073128: 12076336	probable cytokinin riboside 5 -monophosphate phosphoribohydrolase logl6-like	Berry weight	0.29	[42]
VIT_202s0241g00050	2	4698823: 4704204	uncharacterized protein	Flower sex	0.29	[200][222]
VIT_202s0241g00060	2	4715393: 4718698	uncharacterized protein	Flower sex	0.29	[200][222]
VIT_202s0241g00060	2	4715393: 4718698	uncharacterized protein	Flower sex	0.29	[200][222]
VIT_202s0154g00230	2	5036984: 5037952	pinus taeda anonymous locus 0_16347_01 genomic sequence	Flower sex	0.36	[200][222]
VIT_202s0109g00370	2	13050602: 13056119	RNA recognition motif-containing protein	Berry Skin color	0.28	[44]
VIT_202s0109g00380	2	13057949: 13076992	dead-box atp-dependent rna helicase 5 (STRS1)	Berry Skin color	0.28	[44]
VIT_202s0033g00450	2	14308288: 14309480	transcription factor MYBA3 (MYB113)	Berry Skin color	0.28	[44]
VIT_202s0033g00460	2	14313417: 14314479	transcription factor MYBA4 (MYB113)	Berry Skin color	0.36	[44]
VIT_207s0005g04890	7	8141027: 8142187	Glutathione S-transferase 25 (GSTU7)	Berry Skin color	0.28	[44]
VIT_208s0040g01040	8	12066763: 12073699	serine carboxypeptidase-like 45-like (scpl46)	Berry Skin color	0.28	[44]

A cluster of MYB-type transcription factor genes, which control the anthocyanin content in berry skin, is also located on chr2 [257][258][44]. We observed differences in allele frequency (F_{ST} = 0.36) at the transcription factor MYBA3 gene (MYB113; VIT_202s0033g00460) between wild individuals, which presented only colored fruits, and cultivated genotypes, composed by both colored and white varieties. This gene is located within the 5 Mb region on chr2 identified as associated with berry color by Myles et al. [34]. Moreover, the candidate genes for the glutathione S-transferase 25 (GSTU7; VIT_207s0005g04890) and the serine carboxypeptidase-like 45-like (scpl46; VIT_208s0040g01040), identified under berry skin color QTLs [44], revealed a high level of differentiation (F_{ST} ~0.28).

Conclusions

In the present research, we displayed the first whole-genome survey of the genetic differentiation between wild and cultivated grapevines by using population genetics approaches. An overall reduction of genetic diversity has been observed within the wild panel, supporting the occurence of an ongoing progressive decline of natural wild grapevine populations, and the necessity of developing new programs for the characterization and conservation of *V. sylvestris*. Moreover, we identified several genomic regions with divergent allele frequencies between grapevine cultivars and their wild relatives. These genomic regions showed a significant enrichment of gene functional classes related with responses to biotic and abiotic stresses, unraveling putative different mechanisms of adaptation to environmental changes between the two *V. vinifera* subspecies. Indeed, while grapevine cultivars are almost completely addicted to human agricultural practices, wild grapes keep likely facing the constant environmental alterations that still occur in natural habitats. In this regard, our findings pave the way for future studies of functional genomics and/or candidate-gene association mapping, which will provide additional information about how the two forms of *V. vinifera* react to environmental stimuli and stresses, such as water deficit and pathogen attacks. Finally, our results support the broad potential of *V. sylvestris* as spring of resilience factors in future breeding programs to deal with the ongoing climate changes and the increasing demand of a sustainable viticulture.

Supplementary Data

Table S1: Ancestry values inferred by fastSTRUCTURE for 44 grapevine cultivars and 42 wild individuals genotyped at 26,893 SNP loci. The three subgroups inferred based on a membership cutoff of 0.75 are highlighted in grey.

Sample ID	Accession Name	Population	Cluster membership		
			K1	K2	K3
GRAPE_55		sylvestris	1	0	0
GRAPE_56		sylvestris	1	0	0
GRAPE_57		sylvestris	1	0	0
GRAPE_58		sylvestris	1	0	0
GRAPE_62		sylvestris	1	0	0
GRAPE_63		sylvestris	1	0	0
GRAPE_64		sylvestris	1	0	0
GRAPE_65		sylvestris	1	0	0
GRAPE_68		sylvestris	1	0	0
GRAPE_75		sylvestris	1	0	0
GRAPE_77		sylvestris	1	0	0
GRAPE_84		sylvestris	1	0	0
GRAPE_81		sylvestris	0.97	0.03	0
GRAPE_83		sylvestris	0.95	0	0.05
GRAPE_70		sylvestris	0.94	0	0.06
GRAPE_78		sylvestris	0.94	0	0.06
GRAPE_54		sylvestris	0.93	0	0.07
GRAPE_79		sylvestris	0.93	0.07	0
GRAPE_60		sylvestris	0.91	0	0.09
GRAPE_69		sylvestris	0.91	0	0.09
GRAPE_76		sylvestris	0.91	0	0.09
GRAPE_82		sylvestris	0.91	0	0.09
GRAPE_91		sylvestris	0.91	0	0.09
GRAPE_80		sylvestris	0.9	0	0.1
GRAPE_73		sylvestris	0.89	0	0.11
GRAPE_66		sylvestris	0.88	0	0.12
GRAPE_93		sylvestris	0.86	0	0.14
GRAPE_53		sylvestris	0.82	0	0.18
GRAPE_67		sylvestris	0.81	0	0.19
GRAPE_87		sylvestris	0.8	0	0.2
GRAPE_89		sylvestris	0.8	0	0.2
GRAPE_85		sylvestris	0.79	0	0.21
GRAPE_72		sylvestris	0.78	0	0.22
GRAPE_92		sylvestris	0.78	0	0.22
GRAPE_94		sylvestris	0.78	0	0.22
GRAPE_52		sylvestris	0.77	0	0.23
GRAPE_29	Pinot Noir	sativa	0	1	0
GRAPE_06	Gewuerztraminer	sativa	0.08	0.92	0
GRAPE_10	Mornan Noir	sativa	0.08	0.89	0.03
GRAPE_20	Zilavka	sativa	0	0	1
GRAPE_02	Alarije	sativa	0	0	1
GRAPE_36	Rossola	sativa	0	0	1
GRAPE_38	Armenia chi 10	sativa	0	0	1
GRAPE_39	Trollinger Rot	sativa	0	0	1

GRAPE_24	Ak chekerek	sativa	0	0	1
GRAPE_01	Alba aganyn isyoum	sativa	0	0	1
GRAPE_26	Limnio	sativa	0	0	1
GRAPE_27	Canorroio	sativa	0	0	1
GRAPE_46	Ak ouzioum tagapskii	sativa	0	0	1
GRAPE_47	Ahmed	sativa	0	0	1
GRAPE_04	Brustiano	sativa	0.02	0	0.98
GRAPE_21	Vernaccia di S.Gimignano	sativa	0.04	0	0.96
GRAPE_17	Saperavi	sativa	0.06	0	0.94
GRAPE_08	Beli Medenac	sativa	0.06	0	0.94
GRAPE_09	Macabeu	sativa	0.06	0	0.94
GRAPE_18	Malvasia Istriana	sativa	0.08	0	0.92
GRAPE_35	Piè di Palombo	sativa	0.09	0	0.91
GRAPE_45	Buffalo	sativa	0.1	0	0.9
GRAPE_41	Muscat Bleu	sativa	0	0.12	0.88
GRAPE_34	Moscato	sativa	0.13	0	0.87
GRAPE_40	Espadeiro blanco	sativa	0.13	0	0.87
GRAPE_31	Pignoletto	sativa	0.15	0	0.85
GRAPE_48	V.berlandieri Colombard	sativa	0.15	0	0.85
GRAPE_30	Verdelet	sativa	0.17	0	0.83
GRAPE_19	Jacquere	sativa	0.18	0	0.82
GRAPE_37	Castor	sativa	0.12	0.07	0.81
GRAPE_32	Aris	sativa	0.21	0	0.79
GRAPE_42	Bracciola nera	sativa	0.05	0.21	0.73
GRAPE_14	Roussanne	sativa	0.27	0	0.73
GRAPE_25	Ortrugo	sativa	0.18	0.1	0.71
GRAPE_07	Leon Millot	sativa	0.17	0.19	0.64
GRAPE_12	Corbera	sativa	0.09	0.28	0.63
GRAPE_50	V,silvestris cl, Guemuld 103-64	sativa	0.39	0	0.61
GRAPE_05	Forsellina	sativa	0.29	0.16	0.55
GRAPE_15	Csaba gyongye	sativa	0	0.47	0.53
GRAPE_43	Semidano	sativa	0.03	0.45	0.52
GRAPE_86		sylvestris	0.48	0	0.52
GRAPE_22	Shiraz	sativa	0.29	0.35	0.36
GRAPE_11	Lambrusco casetta	sativa	0.48	0.18	0.35
GRAPE_59		sylvestris	0.67	0	0.33
GRAPE_90		sylvestris	0.67	0	0.33
GRAPE_95		sylvestris	0.68	0	0.32
GRAPE_03	Arnsburger	sativa	0	0.71	0.29
GRAPE_71		sylvestris	0.73	0	0.27
GRAPE_88		sylvestris	0.75	0	0.25
GRAPE_23	Claverie coulard	sativa	0.45	0.39	0.16

Figure S1: LD plots (GOLD heatmap) base on r^2 values obtained with Haploview v4.1 for each chromosome within the *sativa* subgroup (red = high r^2; blue = low r^2).

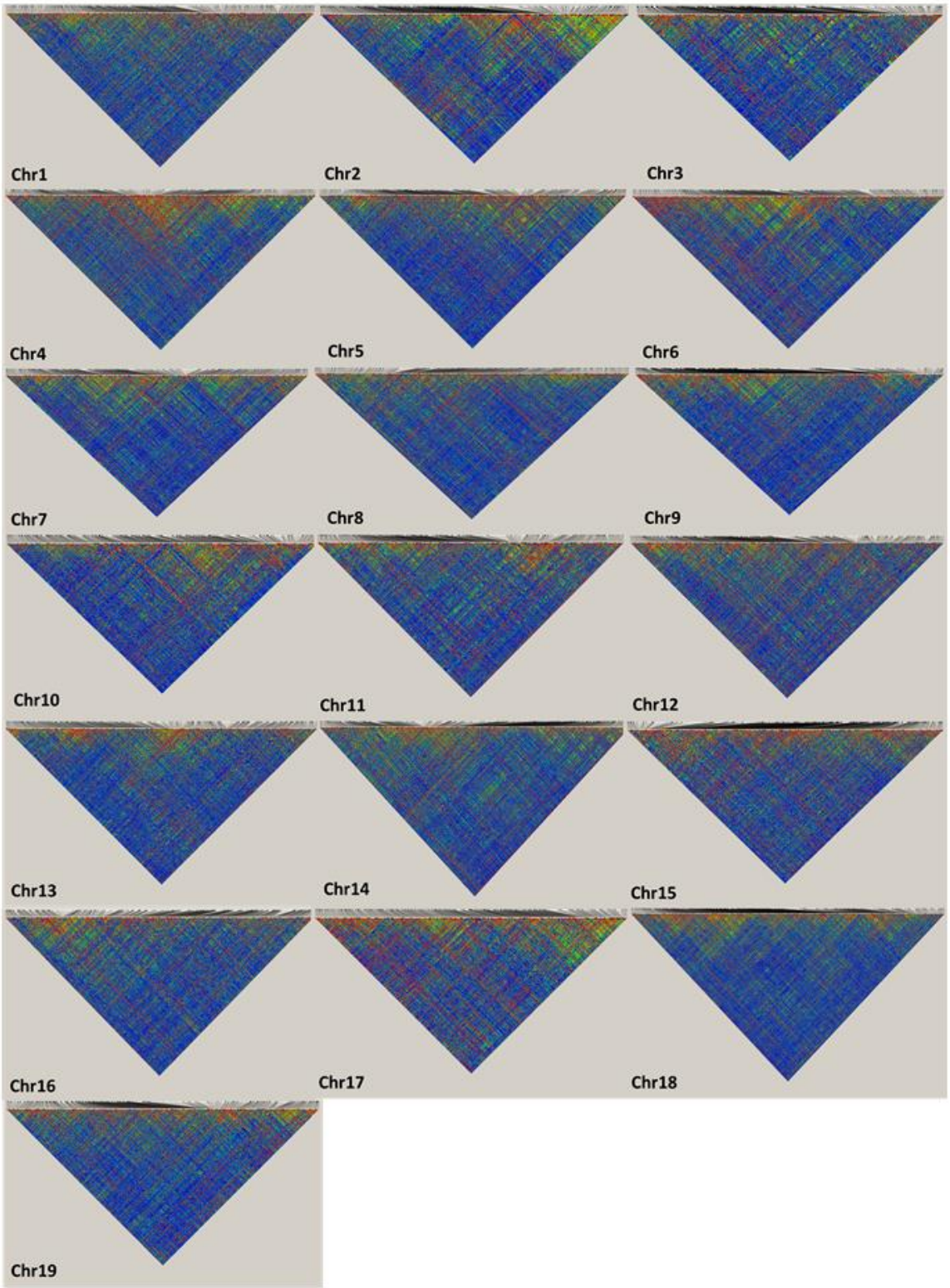

Figure S2: LD plots (GOLD heatmap) base on r^2 values obtained with Haploview v4.1 for each chromosome within the *sylvestris* subgroup (red = high r^2; blue = low r^2).

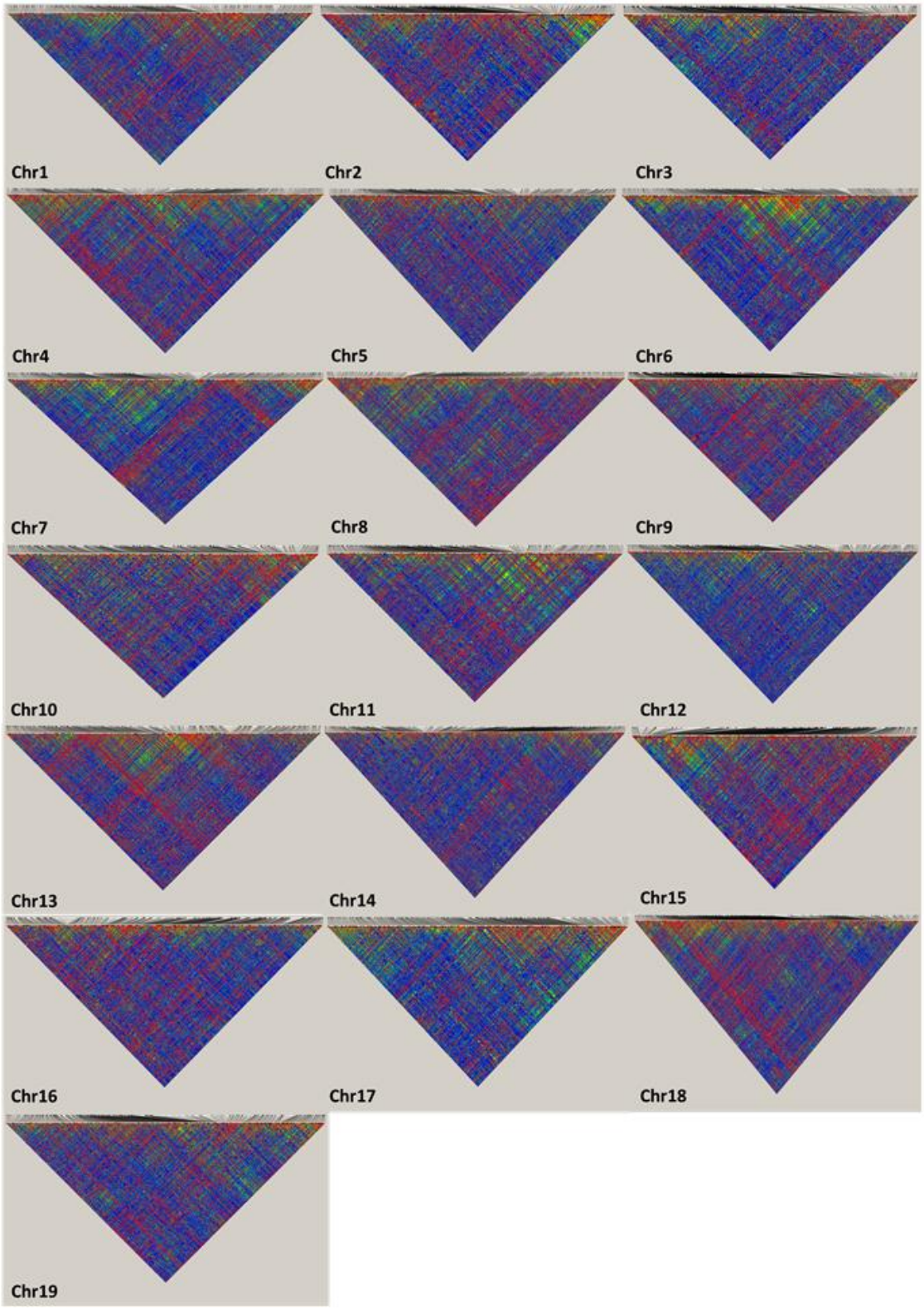

Figure S3: empirical distribution of F_{ST} values across the whole genome between wild and cultivated grapevines. The area shaded in red indicated the top 5% of F_{ST} values. The average and standard deviation (sd) of the 95th percentiles of F_{ST} values gained over 1,000 permutations are also reported.

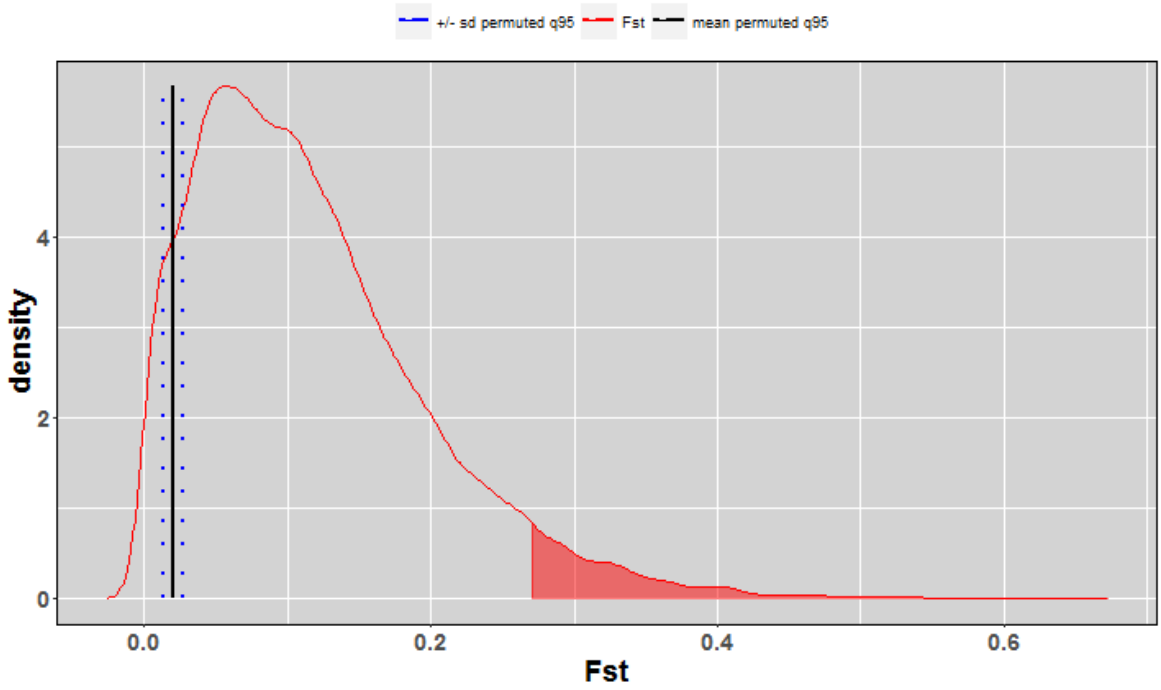

Table S2: Grapevine genes included in the enriched functional classes significantly differentiated between *sativa* and *sylvestris* accessions (significance cutoffs: **99th percentile; *95th percentile).

Gene ID	Chr	Position	Fst	GO Term	Gene name	Gene Annotation (v2.1)
VIT_204s0008g00050	4	16202:16771	**0.49****	nitrogen compound metabolic process	RPL5B	ribosomal protein
VIT_204s0008g03840	4	3182895:3185997	**0.45****	response to biotic stimulus	-	ankyrin repeat-containing protein
VIT_204s0008g03830	4	3178460:3182280	**0.45****	organic substance metabolic process	RBL11	rhomboid family protein
VIT_208s0040g01150	8	12159018:12162600	**0.42****	carbohydrate metabolic process	CPN10	10 kda chaperonin
VIT_217s0000g05240	17	5737662:5753467	**0.4****	response to abscisic acid	-	nuclear transport factor 2 and rna recognition motif domain-containing protein
VIT_211s0065g00150	11	13509591:13529415	**0.4****	carbohydrate metabolic process	SS4	soluble starch synthase iv-1
VIT_204s0008g00480	4	415618:417970	**0.39****	nitrogen compound metabolic process	LPA66	pentatricopeptide repeat-containing protein chloroplastic-like
VIT_205s0049g00250	5	7334595:7335474	**0.39****	oxidation-reduction process	-	Desacetoxyvindoline 4-hydroxylase
VIT_204s0079g00530	4	11130821:11137310	**0.38****	carbohydrate metabolic process	cICDH	nadp-isocitrate dehydrogenase
VIT_208s0040g00270	8	11213199:11217147	**0.37****	nitrogen compound metabolic process	-	splicing factor 3b subunit 1-like
VIT_204s0008g01360	4	1114709:1118921	**0.37****	organic substance metabolic process	-	u-box domain-containing protein 35-like
VIT_217s0000g01100	17	769342:770298	**0.37****	organic substance metabolic process	FATB	myristoyl-acyl carrier protein chloroplastic-like
VIT_218s0001g05250	18	4220268:4222313	**0.37****	response to salt stress	RAP2	ap2 erf domain-containing transcription factor
VIT_215s0021g01590	15	12223557:12224200	**0.37****	defense response	ERF2	erf2 transcription factor
VIT_206s0004g06420	6	7163946:7166617	0.35*	nitrogen compound metabolic process	-	probable lrr receptor-like serine threonine-protein kinase at1g56140-like
VIT_206s0004g04180	6	5152464:5153142	0.35*	response to salt stress	-	nucleic acid binding
VIT_206s0004g06890	6	7615724:7620455	0.35*	stomatal movement	KT1	potassium transporter 1-like
VIT_217s0000g05270	17	5761030:5764757	0.34*	organic substance metabolic process	-	uncharacterized protein

VIT_206s0004g07820	6	8601569:8603477	0.33*	nitrogen compound metabolic process	OTP82	pentatricopeptide repeat-containing protein at1g08070-like
VIT_201s0127g00190	1	7536963:7537993	0.33*	organic substance metabolic process	CRK2	cysteine-rich receptor-like protein kinase 2
VIT_214s0171g00140	14	26008741:26011003	0.33*	organic substance metabolic process	-	type receptor kinase
VIT_206s0004g06980	6	7687431:7690514	0.33*	organic substance metabolic process	-	probable phytol kinase chloroplastic-like
VIT_211s0103g00110	11	15613800:15614090	0.33*	oxidation-reduction process	-	photosystem II protein D2
VIT_217s0000g01040	17	744117:746348	0.33*	response to abscisic acid	HSD7	protein
VIT_213s0101g00050	13	11498528:11499084	0.32*	nitrogen compound metabolic process	RPS1	ribosomal protein s1
VIT_217s0000g06390	17	6970923:6972402	0.32*	nitrogen compound metabolic process	-	uncharacterized protein
VIT_208s0056g01650	8	2648151:2649186	0.32*	response to biotic stimulus	LBD20	protein
VIT_208s0105g00480	8	7811452:7821404	0.32*	carbohydrate metabolic process	SAC8	transmembrane protein g5p
VIT_210s0042g00290	10	13123742:13128460	0.32*	organic substance metabolic process	SMO1-3	protein
VIT_212s0059g01590	12	6491575:6511602	0.32*	organic substance metabolic process	-	gdsl esterase lipase
VIT_205s0049g00410	5	7455371:7457433	0.32*	oxidation-reduction process	-	1-aminocyclopropane-1-carboxylate oxidase homolog 1
VIT_208s0007g05410	8	19354822:19363364	0.31*	nitrogen compound metabolic process	CBL	cystathionine beta-lyase
VIT_207s0141g00580	7	324514:331958	0.31*	carbohydrate metabolic process	GAUT6	alpha- -galacturonosyltransferase
VIT_214s0066g00600	14	27087956:27088713	0.31*	organic substance metabolic process	-	uncharacterized protein
VIT_217s0000g00170	17	88282:89690	0.31*	organic substance metabolic process	VIM1	zinc finger
VIT_207s0129g00680	7	15918217:15921201	0.31*	organic substance metabolic process	-	pentatricopeptide repeat-containing protein chloroplastic-like
VIT_213s0156g00150	13	23889012:23889722	0.31*	oxidation-reduction process	-	protein
VIT_211s0016g02340	11	1886162:1888997	0.3*	nitrogen compound metabolic process	CDA1	cytidine deaminase
VIT_217s0000g04710	17	5104509:5107597	0.3*	nitrogen compound metabolic process	-	pentatricopeptide repeat-containing protein
VIT_206s0004g05500	6	6342667:6344806	0.3*	nitrogen compound metabolic process	-	myosin heavy chain-related protein
VIT_205s0102g00773	5	22709504:22709776	0.3*	defense response	-	probable disease resistance protein rdl6 rf9-like

VIT_208s0040g03200	8	14139778:14141968	0.29*	methylation	-	60s ribosomal protein l4-1
VIT_205s0020g03060	5	4794345:4798018	0.29*	nitrogen compound metabolic process	CYCT1-4	cyclin t1
VIT_205s0020g03070	5	4800109:4802663	0.29*	nitrogen compound metabolic process	-	cyclin family protein
VIT_208s0040g03290	8	14236327:14250749	0.29*	nitrogen compound metabolic process	MCM8	dna replication licensing factor mcm8-like
VIT_206s0009g03385	6	16617200:16618027	0.29*	response to biotic stimulus	-	protein
VIT_201s0026g00090	1	8711856:8728580	0.29*	carbohydrate metabolic process	ULP1D	ubiquitin-like-specific protease 1c
VIT_201s0026g00100	1	8738934:8750115	0.29*	carbohydrate metabolic process	ULP1D	ulp1 protease
VIT_206s0004g08080	6	8843648:8845413	0.29*	carbohydrate metabolic process	XLG1	-
VIT_208s0007g03030	8	17078610:17080600	0.29*	methylation	UBQ1	ubiquitin fusion protein
VIT_211s0065g00640	11	14530232:14586262	0.29*	organic substance metabolic process	CAS1	cycloartenol synthase
VIT_217s0000g00070	17	35355:36590	0.29*	organic substance metabolic process	-	protein
VIT_219s0085g00190	19	22516711:22517184	0.29*	organic substance metabolic process	SK4	skp1-like protein
VIT_219s0085g00195	19	22520035:22520493	0.29*	organic substance metabolic process	SK4	skp1-like protein
VIT_205s0029g00180	5	14365888:14385337	0.29*	organic substance metabolic process	ERD2B	ER lumen protein retaining receptor-like
VIT_206s0004g05610	6	6420422:6423513	0.29*	organic substance metabolic process	-	subtilisin-like serine endopeptidase family protein
VIT_208s0007g08780	8	22158694:22159606	0.29*	organic substance metabolic process	MIZ1	uncharacterized protein
VIT_214s0066g02170	14	28394896:28398366	0.29*	oxidation-reduction process	-	prolyl 4-hydroxylase
VIT_208s0040g03180	8	14133347:14136582	0.29*	oxidation-reduction process	RAP2	ap2 domain-containing transcription factor
VIT_214s0066g02040	14	28307901:28309789	0.29*	response to abscisic acid	AATP1	atp binding
VIT_214s0066g02050	14	28314798:28316327	0.29*	response to abscisic acid	-	protein
VIT_214s0066g02060	14	28318004:28319797	0.29*	response to abscisic acid	AATP1	atp binding
VIT_214s0066g02100	14	28353929:28355637	0.29*	response to abscisic acid	AATP1	mitochondrial chaperone bcs1
VIT_214s0066g02110	14	28358493:28360078	0.29*	response to abscisic acid	AATP1	atp binding
VIT_213s0067g03350	13	1837463:1838505	0.28*	methylation	-	60s ribosomal protein l4-1
VIT_206s0004g03730	6	4680671:4687232	0.28*	nitrogen compound metabolic process	NRPC1	dna-directed rna polymerase iii subunit rpc1-like

VIT_206s0004g03740	6	4692715:4737715	0.28*	nitrogen compound metabolic process	NRPC1	dna-directed rna polymerase iii subunit rpc1-like
VIT_206s0004g03780	6	4759942:4760340	0.28*	nitrogen compound metabolic process	PRS	wuschel-related homeobox 3
VIT_206s0004g04040	6	5021170:5031333	0.28*	nitrogen compound metabolic process	-	pentatricopeptide repeat-containing protein
VIT_206s0004g05930	6	6666460:6669570	0.28*	nitrogen compound metabolic process	PCNA2	proliferating cell nuclear antigen
VIT_208s0007g03340	8	17307190:17311533	0.28*	nitrogen compound metabolic process	-	ribosomal protein l1
VIT_215s0046g01190	15	18217371:18218971	0.28*	regulation of plant-type hypersensitive response	GT72B1	hydroquinone glucosyltransferase
VIT_215s0046g01210	15	18226980:18228562	0.28*	regulation of plant-type hypersensitive response	GT72B1	hydroquinone glucosyltransferase
VIT_204s0023g00110	4	16084017:16085284	0.28*	carbohydrate metabolic process	-	alpha- -glucan-protein synthase
VIT_206s0004g05040	6	5967109:5968595	0.28*	methylation	-	isoprenylcysteine carboxyl methyltransferase
VIT_215s0021g01110	15	11136019:11137148	0.28*	organic substance metabolic process	CYP714A1	cytochrome p450
VIT_215s0021g01380	15	11651026:11665356	0.28*	organic substance metabolic process	-	kinase like protein
VIT_215s0046g01150	15	18197371:18198267	0.28*	organic substance metabolic process	-	anthocyanidin reductase-like
VIT_215s0046g01320	15	18339430:18341953	0.28*	organic substance metabolic process	-	protein kinase-like protein
VIT_216s0098g01780	16	21844781:21851539	0.28*	organic substance metabolic process	SSI1	soluble starch synthase I
VIT_217s0000g05592	17	6117344:6117909	0.28*	organic substance metabolic process	-	momilactone a synthase
VIT_217s0000g05600	17	6124001:6125145	0.28*	organic substance metabolic process	-	short-chain alcohol dehydrogenase
VIT_205s0049g01050	5	8090670:8091361	0.28*	organic substance metabolic process	-	protein
VIT_205s0094g01270	5	24541244:24543738	0.28*	organic substance metabolic process	BIR1	probably inactive leucine-rich repeat receptor-like protein kinase at5g48380-like
VIT_207s0005g04840	7	8105826:8106863	0.28*	organic substance metabolic process	MAPKKK21	mitogen-activated protein kinase kinase kinase anp1-like
VIT_207s0005g03750	7	6715566:6716963	0.28*	oxidation-reduction process	RIC7	protein
VIT_207s0005g04060	7	7161837:7164031	0.28*	oxidation-reduction process	-	protein

VIT_215s0046g00440	15	17407010:17409649	0.28*	defense response	PI4K	phosphoinositide 4-kinase gamma 4
VIT_211s0016g04700	11	3988735:3992060	0.27055*	organic substance metabolic process	KCS11	beta-ketoacyl-coa synthase family protein
VIT_206s0004g06310	6	7093095:7104132	0.27*	methylation	-	60s acidic ribosomal protein p0
VIT_216s0022g01860	16	14100069:14219500	0.27*	nucleic acid phosphodiester bond hydrolysis	CPSF160	protein
VIT_204s0008g02230	4	1834035:1834800	0.27*	nitrogen compound metabolic process	-	ap2 erf domain-containing transcription factor
VIT_204s0008g03960	4	3298715:3326185	0.27*	nitrogen compound metabolic process	-	protein
VIT_211s0016g04580	11	3888012:3891011	0.27*	nitrogen compound metabolic process	CRR21	chlororespiratory reduction partial
VIT_211s0016g04630	11	3959481:3961177	0.27*	nitrogen compound metabolic process	GAI	della protein
VIT_211s0016g04640	11	3966363:3968501	0.27*	nitrogen compound metabolic process	GONST4	gdp-mannose transporter
VIT_217s0000g01760	17	1307825:1308955	0.27*	nitrogen compound metabolic process	-	duf246 domain-containing protein at1g04910-like
VIT_218s0001g06980	18	5220072:5221115	0.27*	nitrogen compound metabolic process	-	pentatricopeptide repeat-containing protein
VIT_205s0020g04520	5	6332992:6339892	0.27*	nitrogen compound metabolic process	LFR	leaf and flower related protein
VIT_208s0032g00010	8	2798984:2799422	0.27*	nitrogen compound metabolic process	-	maturase
VIT_208s0007g03150	8	17167523:17169413	0.27*	nitrogen compound metabolic process	-	pentatricopeptide repeat-containing protein
VIT_209s0070g00360	9	13521416:13522967	0.27*	nitrogen compound metabolic process	-	aryl-alcohol dehydrogenase - like
VIT_209s0054g01000	9	21857946:21860542	0.27*	nitrogen compound metabolic process	-	uncharacterized protein
VIT_211s0206g00140	11	7470447:7473588	0.27*	carbohydrate metabolic process	SVL4	glycerophosphoryl diester phosphodiesterase family protein
VIT_212s0059g01660	12	6564094:6572507	0.27*	organic substance metabolic process	-	-
VIT_214s0066g00170	14	26743582:26745706	0.27*	organic substance metabolic process	CYP724A1	cytochrome p450 724b1
VIT_209s0070g00320	9	13450437:13452476	0.27*	organic substance metabolic process	-	cyclin-dependent kinase f-4-like
VIT_207s0031g00100	7	16332009:16337077	0.27*	oxidation-reduction process	-	2-oxoglutarate-fe - dependent oxygenase domain-containing protein

VIT_204s0008g03950	4	3287257:3289094	0.27*	response to abscisic acid	RD22	dehydration-responsive protein rd22

Chapter 4

A GENOME-WIDE ASSOCIATION STUDY TO REVEAL CANDIDATE GENES FOR DOMESTICATION-RELATED TRAITS IN GRAPEVINE

Abstract

Background: Domestication involved strong novel selection operating on suites of traits, which underwent phenotypic evolution from wild plants to crops. Association mapping is one of the methods currently employed in identifying the genes and mutations that have been targets of selection during crop domestication, and to explore the considerable genetic variation still maintained in natural populations.

Results: An association panel consisting of 42 wild and 46 cultivated accessions of *V. vinifera* was phenotyped for up to ten traits, including berry and bunch weight, yield and berry composition. A huge phenotypic variation was observed within and between the two grapevine subspecies, notably for berry size, pH, acid contents and titratable acidity. By using a panel of 26K SNPs, association analysis for each trait was carried out testing three different models which account for either population structure (GLM (Q)), familial relatedness (MLM (K)) or both (MLM (Q+K)). Significant genotype-phenotype associations were identified for all traits, except for single berry weight. In addition, cross associations were detected between yield and single bunch weight, and among malate concentrations and titratable acidity. 20 kb genomic regions surrounding the SNPs significantly associated to traits were scanned to search for candidate genes, yielding a total of 127 genes. In particular, genes encoding proteins related to Ca^{2+} sequestration and signalling, transcription factors and enzymes involved in the metabolism of polyamines were identified in linkage with the SNPs significantly associated to yield and bunch weight. At the same time, genes with a central role in the control of berry flesh pH and acidity were detected, such as the isocitrate lyase and V-type proton ATPase subunit a3 genes.

Conclusions: our findings support the feasibility of association mapping to identify the genes and mutations underlying the phenotyping changes occurred during grapevine domestication and improvement. However, in order to increase the power and resolution of GWA studies in grapevine, further progresses are required towards the high-throughput acquisition of genome-wide markers in grapevine and the accurate collection of phenotypic data in bigger association panels.

Background

Modern crops resulted from the long process of selection, breeding and adaptation, which have started with the beginning of agriculture around 10,000 years ago [144]. Although the domestication process deeply influenced the genetic diversity of modern crops, the majority of their current genetic variation has arisen from spontaneous mutations in their wild progenitors [112]. Therefore, understanding the phenotypic variation associated with the domestication process in crops will help to identify the genetic bases of domestication-related traits, and to better utilize the genetic resources for crop improvement (Huang and Han 2014). Advances in plant genomics during the last 10 years have introduced new tools for breeding strategies, such as genome-wide association studies (GWAS) and genomic selection (GS) [259]. Unlike the traditional quantitative trait loci (QTL) mapping, which uses bi-parental populations to study the co-segragation of traits with markers, GWAS and GS are applied to populations of unrelated individuals, designed to capture a substantial portion of specie-wide variation [112]. The main difference between GS and GWAS strategies is that the former is used to predict phenotypes from marker profiles alone, reducing the time and costs involved in phenotyping breeding lines [260], while the latter aims at the identification of novel genotype-phenotype correlations that can be implemented in cultivar improvement through marker-assisted selection (MAS) [112]. GWAS takes full advantage of ancient recombination events occurred during the history of the association panel [145]. This provides higher mapping resolution than traditional gene mapping, which instead relies on the limited recombination history of a bi-parental population [261]. Moreover, while QTL mapping strategies use only the allelic diversity that segregates between the parents of a particular F2 population or within a Recombinant Inbred Lines (RIL) family, GWAS studies explore a broader genetic variation which depends on the size, geographic origin and genetic history of the population [262]. This increases the power to detect significant genotype-phenotype correlations for traits with a polygenic nature [263]. However, the trait genetic architecture has a huge influence on the GWAS performance. If the trait is controlled either by many rare variants with a large effect on the phenotype or by many common variants with a small phenotypic effect, the power of GWAS to identify a true marker-trait association is deeply compromised [264]. Rare variants can only be detected with adeguate local sampling and may create synthetic genome-wide associations because they are usually linked with many other non-causative variants within the genome, regardless of the extent of Linkage Disequilibrium (LD) [265]. Allelic and genetic heterogeneities are two other common drawbacks of GWAS [262]. Allelic heterogeneity occurs when multiple functional alleles of the same gene contribute to different phenotypes [266], while genetic heterogeneity consists in the control of phenotype by multiple major genes in LD [267]. Moreover, the genetic interaction between loci (epistasis) as well as the interaction between genes and environment (GxE) and the epigenetic variation represent other important GWAS issues [268, 269]. All these factors may account for the “missing heritability”, defined as a portion of genetic variance that cannot be explained by all significant marker-trait associations detected by GWAS [270]. The influence of each factor on GWAS performance depends strongly on the population sampled [262]. A matter requiring attention in crop GWAS is the need to account for confounding factors, that is population structure and criptic relatedness among studied individuals [271]. Population stratification results from the inclusion of individuals from different populations (i.e. diverse geographic origin), while cryptic relatedness refers to the degree of genetic relationship between individuals of the association panel. Indeed, samples with

a common genetic background share both casual and non-casual alleles and the LD between these sites can lead to spurious phenotype-genotype associations [272]. Accordingly GWAS methods based on the unified mixed linear model (MLM) have been developed [133] to account for confounding factors. In addition, more efficient alghorithms have been implemented to make MLM less computationally intensive [273, 132].

GWAS has been widely applied in human genetics to identify major genes involved in diseases [274]. Recently GWAS approaches have also been carried out successfully in many crops, including maize [275], rice [206], sorghum [135] and barley [276]. Few applications of GWAS have been reported so far for perennial species, notably fruit trees. Kumar et al. [139] carried out a GWA analysis in apple for various fruit quality traits by applying a family-based design with controlled structure. On the other hand, Sardos et al. [140] applied GWAS to reveal the genetic bases of seedless phenotype in banana. The low number of GWA studies in fruit species may be ascribed to the difficulties in building up an ideal association panel. Indeed, extended juvenile phases, large plant size and the difficulties to collect information on commercially relevant traits (i.e. fruit quality) discourage breeding programs, which usually rely on only a small number of elite varieties [51]. This leads to have less unique genotypes in germplasm collection of perennial species than for annual crops, and a large part of these genotypes shares a high degree of genetic relationship [113]. Therefore, the design of a broad association panel composed by numerous unique individuals without introducing complex pattern of population stratification or familial relatedness is quite difficult in fruit trees species. Chitwood et al. [45] performed GWAS to map the genetic basis of leaf morphology in grapevine by using a population of 961 accessions genotyped with the Vitis9KSNP array [34]. Out of the 13 phenotyped traits only 4 resulted significantly associated with a handful of SNPs on chromosomes 1 and 6 after multiple testing p-value adjustment. This GWA study underlined the limited power of association mapping studies in grapevine because of the rapid LD decay [21]. Fodor et al. [187] simulated GWAS for traits of different complexity on a population of 3,000 grapevine accessions, structured into three groups, using approx 90K SNPs. This simulation revealed how GWAS in grapevine was more efficient to detect a few loci with a large effect (characteristic of simple traits) than to identify multiple loci with small additive effects. Moreover, they postulated how sample size and the level of genetic diversity can compromise the power of GWAS in grapevine.

In the present research GWAS has been applied as an alternative approach to dissect the genomic bases of domestication-related traits in grapevine. A germplasm collection of cultivated and wild grapevines has been evaluated for single berry and single bunch weight, number of bunches per plant, yield and berry composition (sugar, organic acid and K^+ concentrations, titratable acidity and pH). The use of wild relatives aimed to explore novel genetic diversity potentially interesting for crop improvement [142]. In addition, this study contributes to add novel biological information about the changes occurred during the domestication process in grapevine.

Methods

Plant material and phenotypes

The association population consisted of 88 grapevine (*V. vinifera* L.) accessions, grafted on the rootstock Kober 5BB at the FEM experimental field ''Giaroni" in San Michele all'Adige (Trentino, Italy), and uniformly pruned and trained according to the Guyot system (Appendix B, page 125). This population included 42 *V. sylvestris* and 46 cultivars of the G-110 core collection of *V. sativa*, which includes the overall genetic diversity at 22 SSR loci and 384 SNPs found within the source collection [49]. Phenotypic evaluation of 2 to 5 replicates per genotype was performed in 2012 and 2013 for all traits as follows: clusters of each fruit-bearing plant were harvested six weeks after véraison for the evaluation of single bunch weight (OIV code number 502; SBCW), single berry weight (OIV code number 503; SBW), yield (OIV code number 504) and number of bunches per plant (NBCs). Juice samples (50 ml) from berries were measured with FTIR (Fourier transform infrared) using a FOSS instrument (FOSS NIRSystems, Oatley, Australia) for standard maturity analyses. Thus, total soluble solids (Brix°), titratable acidity, pH, malic and tartaric acid concentrations (g l^{-1}), and potassium (g l^{-1}; K^{+}) content were assayed.

Statistical analysis

All statistical analyses were performed using R packages 'stats' v3.4.0 [177] and 'ggplot2' v2.1.0 [178]. Average values for replicates were used to evaluate correlation between the two year measurements. Moreover, Pearson correlation value (R) between each pair of variables was estimated in the whole population and the two subspecies separately with the 'Hmisc' v 3.17-3 R package [277]. One to six aberrant values were discarded according to traits. Different mixed models were fitted with lme4 package [278] in order to identify the best fit model for each trait. Model comparison was performed from the simplest model, based only on general mean and random genotypic effect (G), to the most complete one, based on general mean, random genotypic effect, fixed year effect (Y) and random genotype x year effect (GxY). Model selection was based on the Bayesian information criterion (BIC). Moreover, the mixed model assumption of normality of residual and BLUPs was checked after model fitting by quantile-quantile plot comparing the distribution of residual and random effect predictors to a theoretical normal distribution. No data transformation of phenotypes was performed. Based on the best fitted model, genotypic best linear unbiased predictor (BLUP) and broad-sense heritability were extracted [279].

SNP genotyping and LD estimation

Details of genotyping protocols for the studied population are reported in previous chapters. Briefly, SNPs genotypes were obtained by applying both the commercial GrapeReseq 20K chip (see Chapter 3) and a novel protocol of RAD-seq (see Chapter 2). SNP loci with a missing rate > 0.2 were filtered out and genotype imputation was performed to fill in the remaining missing data using LinkImpute v1.1.1 software [169]. SNPs with Minor Allele Frequency (MAF) > 0.05 were used to analyze the genetic structure of the population. Both a Bayesian approach, as

implemented in fastSTRUCTURE software v1.0 [210], and a Principal Component Analysis (PCA) [214] were perfomed, revealing a clear distinction between *sylvestris* and *sativa* genotypes. In addition, pairwise LD between SNPs was calculated with Plink v1.9 software [208] using the classical correlation coefficient r^2 [109]. A degree of LD below 0.2 was observed within 10 kb.

Marker-trait association analysis

Genotype-phenotype associations were tested using both BLUPs and the average performance of each sample in each year separately. In addition, genome-wide association study (GWAS) was run for the trait "species" by giving to *sativa* accessions a score of 1 and to *sylvestris* samples a score of 0. GWAS was carried out applying three models which account for different confounding factors to avoid spurious marker-trait associations. The first model applied was the General Linear Model (GLM), which takes into account the population structure inferred by fastSTRUCTURE. The GLM equation can be expressed as

$$y_i = \mu + x_i\beta + Q\nu + \varepsilon \quad (1)$$

where y_i is the phenotype of i^{th} sample, μ is the model intercept, β is a vector of SNP effects, ν is a vector of population effect and ε is a vector of residual effects. Q is the matrix from fastSTRUCTURE which presents the individual probabilities to belong to a subpopulation. The second model applied was the Mixed Linear model, which extends equation (1) by incorporating a kinship matrix (K) to define the degree of genetic covariance between pairs of individuals [133]. A centered identical-by-state K matrix was estimated in TASSEL v5.0 [280] by using the method of Endelman and Jannink [281]. As both population structure and kinship were incorporated, this full model was called MLM (Q + K). Meanwhile, K only model, called MLM (K), which omits the population structure Q from the full model, was also used. All three models are implemented in TASSEL v5.0 software [280]. A quantile-quantile (Q-Q) plot was used to choose the model which better accounts for population structure and familial relatedness in the marker-trait association. Indeed, in this plot the negative logarithms of the p-values from each model were plotted against their expected values under the null hypothesis of no association with the trait. P-values adjustment for multiple testing was performed, and the Bonferroni-corrected critical p-values and False Discovery Rate (FDR) were used to identify significant marker-trait associations. Manhattan plots were displayed accordingly by using the 'qqman' v0.1.3 R package [282].

Identification of candidate genes

The positions of markers significantly associated to phenotypes were used to investigate the grapevine gene annotation v2.1 [100]. With regard to the extent of LD, windows of 10 kb upstream and downstream the SNPs of interest were used to identify candidate genes. In particular, the pattern of LD was inspected through heatmap visualization with Haploview v4.1 [215] to ensure the extent of LD around the SNPs associated with phenotypes. Indeed, if the markers fell within long LD blocks, the entire genomic region located between the extreme SNPs was explored.

Results

Phenotypic data

The grapevine population of wild and cultivated accessions was phenotyped six weeks after véraison in two years for up to ten traits. Differences were observed between the *sativa* and the *sylvestris* for all variables in both years (Figure 1A-E). For most traits cultivated varieties exhibited higher variation than wild genotypes as indicated by standard deviation (SD; Table 1), except for tartaric acid whose concentration varied more in the *sylvestris*. In addition, six wild genotypes didn't produce any bunch in both years and for other two wild accessions bunches couldn't be harvested in 2013.

The number of bunches per plant (NBCs) ranged from 1.6 (accession "Ahmed") to 38.8 (cv "Pinot Meunier") in *sativa* group with an average of 14.6 bunches per plant (Table 1). Instead, the *sylvestris* had an average of 7.8 NBCs, ranging from 1 to 25 bunches. The differences between cultivars and wild grapevines were more evident in yield (kg), single berry weight (g; SBW) and single bunch weight (g; SBCW). Indeed, grapevine varieties produced on average 1.9 kg of grapes per year with a maximum of 6.7 kg (cv "Zilavka"), while sylvestris genotypes had a yield 91.7% smaller (Table 1). The single berry weight (SBW) as well as the single bunch weight (SBCW) varied, respectively, by a four- and ten- fold factor (Table 1) between *sativa* and *sylvestris* genotypes. The former presented SBW from 5.9 g (accession "Ak ouzioum tagapskii") to 0.5 g (accession "Aris") and SBCW from 456.3 g (accession "Rossola") to 9.9 g (accession "Aris"), while the latter showed SBW from 1.3 g to 0.3 g and SBCW from 47.8 g to 1.7 g (Figure 2).

While Brix° and Potassium contents (g l^{-1} ;K^+) showed less variability between cultivated and wild grapevines (Figure 1C-D), significant differences were observed between the two subspecies for pH, titratable acidity (as tartaric acid g l^{-1}), malic and tartaric acid concentrations (g l^{-1}). In particular, the *sylvestris* presented on average lower pH with higher acid concentrations (Table 1; Table 2B-C) than the grapevine cultivars. However, the missing rate was higher for malic and tartaric concentrations, titratable acidity and K^+ content, because not enough juice was produced for 14 *sylvestris* in 2012 and 20 *sylvestris* in 2013. Nevertheless, phenotypic data of the two years were strongly correlated for all traits, notably for SBW, SBCW, titratable acidity, tartaric and malic acid concentrations, and K^+ content (Figure 3).

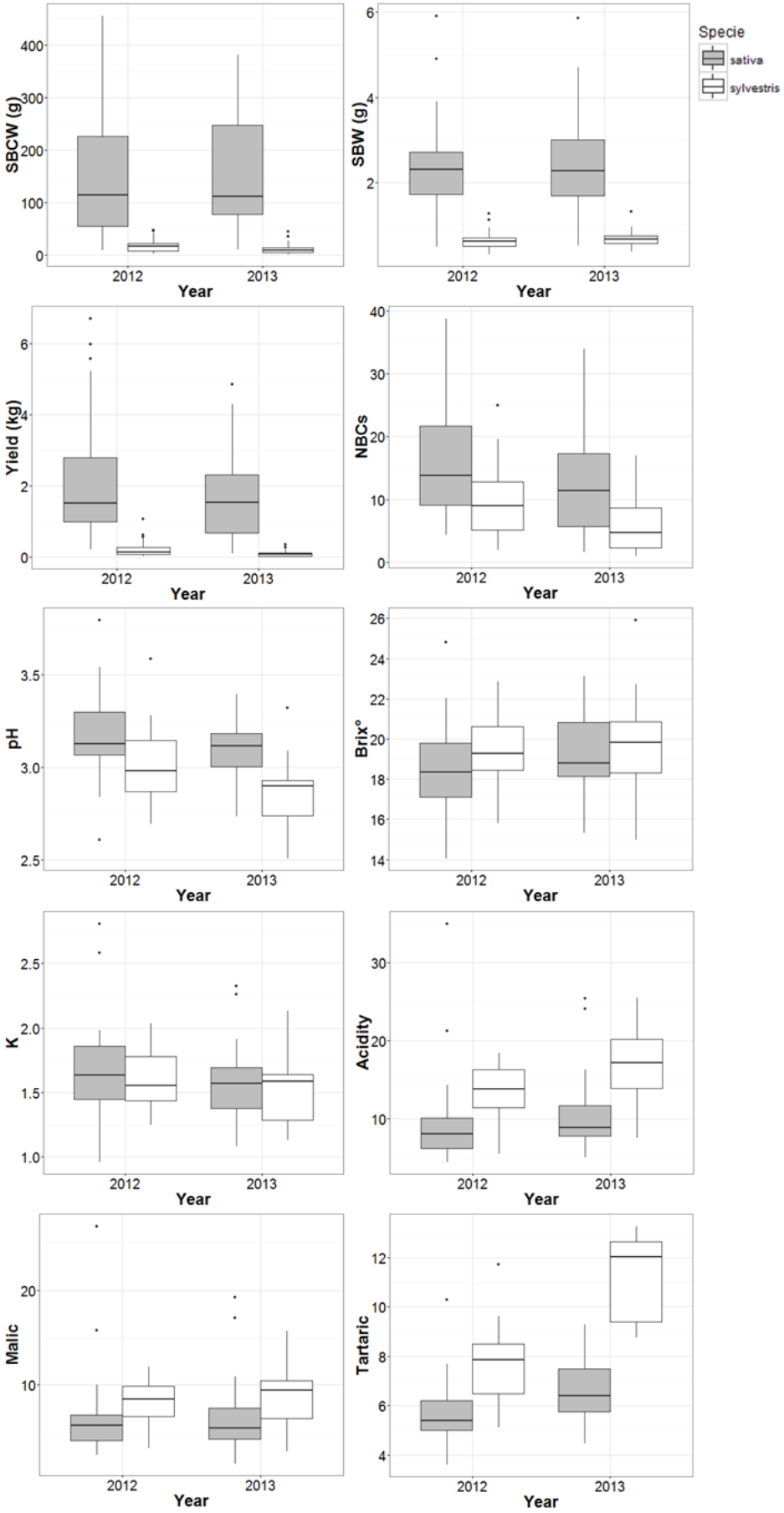

Figure 1: comparison of phenotypic data between cultivated (in grey) and wild (in white) individuals in the two years of measurements (2012, 2013).

Table 1: Descriptive statistics and comparison of the phenotypic data from *sativa* and *sylvestris* accessions.

Specie	NBCs		Yield		SBW		SBCW		Brix°		pH		Acidity		Tartaric		Malic		K	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
sativa	14.62	8.92	1.93	1.45	2.40	1.05	154.28	110.40	18.93	2.05	3.14	0.19	9.54	4.58	6.15	1.34	6.28	3.58	1.61	0.32
sylvestris	7.77	5.30	0.16	0.18	0.65	0.20	15.07	11.58	19.65	1.93	2.93	0.20	14.80	4.48	8.96	2.44	8.40	2.95	1.58	0.27

Table 2-A: Descriptive statistics of the phenotypic data from *sativa* and *sylvestris* accessions in each year of phenotyping.

Specie	Year	NBCs				Yield				SBW				SBCW			
		Mean	Max	Min	SD	Mean	Max	Min	SD	Mean	Max	Min	SD	Mean	Max	Min	SD
sativa	*2012*	15.70	38.80	4.33	8.17	2.14	6.71	0.22	1.65	2.34	5.92	0.48	1.02	149.43	456.27	9.90	112.97
	2013	13.49	34.00	1.67	9.62	1.72	4.85	0.08	1.21	2.47	5.86	0.51	1.08	159.13	381.52	10.07	108.85
sylvestris	*2012*	9.71	25.00	2.00	5.51	0.22	1.07	0.02	0.23	0.63	1.26	0.31	0.21	18.31	47.81	2.47	12.58
	2013	5.77	17.00	1.00	4.30	0.09	0.35	0.00	0.10	0.67	1.32	0.37	0.18	11.74	45.39	1.75	9.53

Table 2-B

Specie	Year	Brix°				pH				Acidity				Tartaric			
		Mean	Max	Min	SD	Mean	Max	Min	SD	Mean	Max	Min	SD	Mean	Max	Min	SD
sativa	*2012*	18.56	24.81	14.07	2.19	3.18	3.80	2.61	0.21	9.02	34.98	4.40	4.99	5.63	10.28	3.61	1.22
	2013	19.31	23.14	15.33	1.86	3.10	3.40	2.74	0.16	10.06	25.43	5.00	4.11	6.67	9.28	4.46	1.25
sylvestris	*2012*	19.46	22.85	15.82	1.71	3.00	3.59	2.70	0.20	13.44	18.40	5.45	3.54	7.77	11.71	5.10	1.78
	2013	19.83	25.92	14.98	2.14	2.85	3.32	2.51	0.16	16.97	25.50	7.45	5.14	11.16	13.27	8.74	1.93

Table 2-C

Specie	Year	Malic				K			
		Mean	Max	Min	SD	Mean	Max	Min	SD
sativa	*2012*	6.28	26.76	2.63	3.91	1.65	2.81	0.96	0.36
	2013	6.28	19.24	1.70	3.26	1.56	2.33	1.08	0.27
sylvestris	*2012*	8.23	11.89	3.37	2.42	1.60	2.04	1.25	0.25
	2013	8.69	15.69	2.96	3.77	1.54	2.13	1.13	0.30

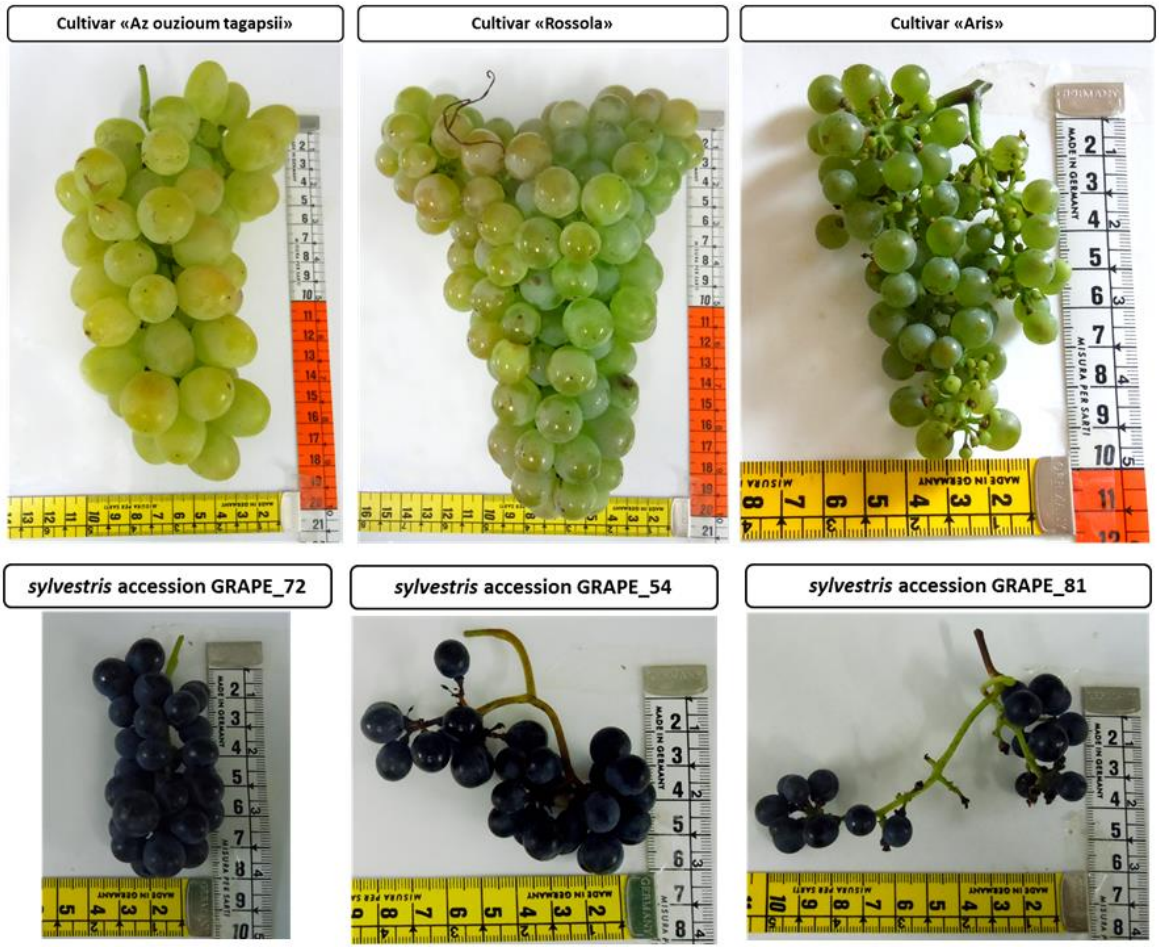

Figure 2: bunches from grapevine cultivars and *sylvestris* accessions showing the highest or lowest value of SBW or SBCW.

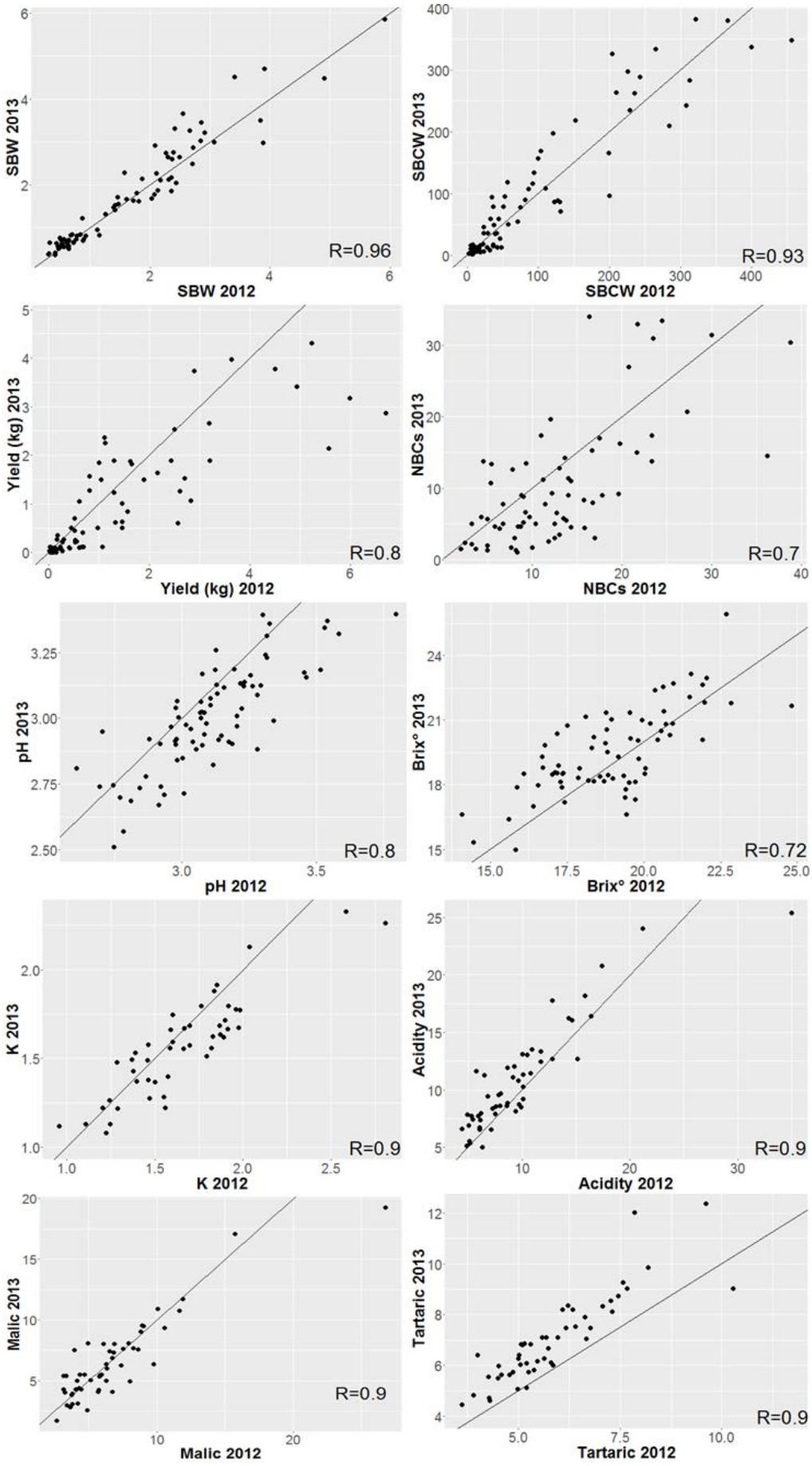

Figure 3: correlation analysis between phenotypic data collected in 2012 and 2013 for each trait.

Pearson's correlation coefficient (R) was estimated between each pair of variables in the whole population and in the two subgroups separately (Figure 4).

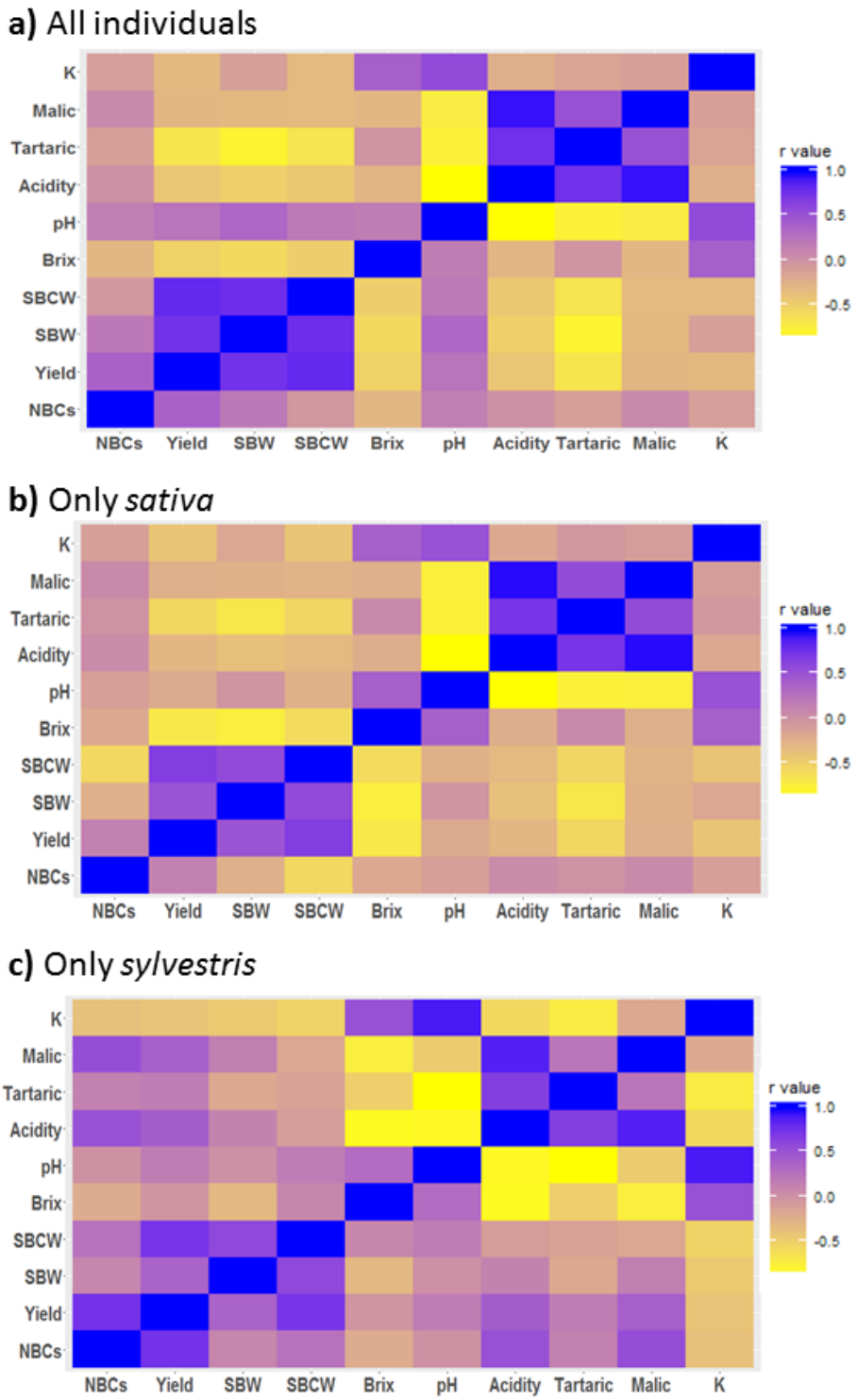

Figure 4: Pearson'correlation analysis beween each pair of traits within the whole population **(a)**, the *sativa* **(b)** and the *sylvestris* **(c)** subgroups.

The yield was more correlated (R ~0,8) with both SBW and SBCW than with NBCs (R = 0.4) in the whole population (Table 3). Instead, in the *sylvestris* the yield was highly correlated with both NBCs and SBCW rather than with SBW (Table 4; Figure 4c). This suggests that the productivity of wild grapevine may be more related with the number of clusters and the number of berries per bunch produced, since the berry weight barely reached values higher than 1.5 g. In addition, a significant inverse correlation (Table 3-5) was recorded for total soluble solids (Brix°) with SBW and yield in both the whole population and the cultivated grapevines. This result can be explained by the shrinkage of berries which occurs during véraison due to the loss of water by transpiration [283]. On the other hand, Brix° was correlated (R=0.5) with pH in the wild grapes. As expected, pH, malic and tartaric acid concentrations, and titratable acidity were highly correlated. However, in the *sylvestris* the pH was strongly correlated with tartaric acid concentration, while titratable acidity was mainly related with malic acid content (Table 4). Moreover, in the *sativa* the concentration of tartaric acid was negatively correlated with yield, SBW and SBCW (Table 5; Figure 4b). Finally, K^+ concentration was correlated with pH in both *sylvestris* and *sativa* groups. In the cultivated subgroup, a correlation between K^+ and Brix° was also found(Figure 4).

Table 3: Pearson's correlation analysis between traits within the whole population.

	NBCs	**Yield**	**SBW**	**SBCW**	**Brix°**	**pH**	**Acidity**	**Tartaric**	**Malic**	**K**
NBCs	-	0.43*	0.25*	0.03	-0.21*	0.21	0.09	-0.03	0.14	-0.03
Yield	0.43**	-	0.77**	0.83**	-0.41**	0.29*	-0.31*	-0.55*	-0.20	-0.22
SBW	0.25*	0.77**	-	0.79**	-0.46**	0.4**	-0.39**	-0.66**	-0.23	-0.03
SBCW	0.03	0.83**	0.79**	-	-0.38**	0.25*	-0.33*	-0.54*	-0.23	-0.23
Brix°	-0.21	-0.41**	-0.46**	-0.38*	-	0.23*	-0.19	0.05	-0.21	0.45**
pH	0.21	0.29*	0.4**	0.25*	0.23*	-	-0.74**	-0.64**	-0.6**	0.6**
Acidity	0.09	-0.31*	-0.39**	-0.33*	-0.19	-0.74**	-	0.77**	0.95**	-0.14
Tartaric	-0.03	-0.55**	-0.66**	-0.54**	0.05	-0.64**	0.77**	-	0.56**	-0.07
Malic	0.14	-0.20	-0.23	-0.23	-0.21	-0.6**	0.95**	0.56**	-	-0.03
K	-0.03	-0.22	-0.03	-0.23	0.45**	0.6**	-0.14	-0.07	-0.03	-

Significance levels: * 0.05; ** 0.001

Table 4: Pearson's correlation analysis between traits within the *sylvestris.*

	NBCs	Yield	SBW	SBCW	Brix°	pH	Acidity	Tartaric	Malic	K
NBCs	-	0.75**	0.10	0.26	-0.21	0.01	0.52	0.14	0.55	-0.38
Yield	0.75**	-	0.38*	0.73**	-0.02	0.17	0.43	0.17	0.41	-0.40
SBW	0.10	0.38*	-	0.59**	-0.31	0.01	0.12	-0.19	0.15	-0.45
SBCW	0.26	0.73**	0.59**	-	0.09	0.18	-0.09	-0.14	-0.18	-0.53
Brix°	-0.21	-0.02	-0.31	0.09	-	0.30	-0.84*	-0.50	-0.74	0.52
pH	0.01	0.17	0.01	0.18	0.30	-	-0.83*	-0.88**	-0.46	0.9**
Acidity	0.52	0.43	0.12	-0.09	-0.84	-0.83*	-	0.66	0.87*	-0.57
Tartaric	0.14	0.17	-0.19	-0.14	-0.50	-0.88**	0.66	-	0.23	-0.73
Malic	0.55	0.41	0.15	-0.18	-0.74	-0.46	0.87*	0.23	-	-0.19
K	-0.38	-0.40	-0.45	-0.53	0.52	0.90	-0.57	-0.73	-0.19	-

Significance levels: * 0.05; ** 0.001

Table 5: Pearson's correlation analysis between traits within the *sativa.*

	NBCs	Yield	SBW	SBCW	Brix°	pH	Acidity	Tartaric	Malic	K
NBCs	-	0.21	-0.14	-0.43**	-0.08	-0.02	0.13	0.09	0.14	-0.02
Yield	0.21	-	0.54**	0.69**	-0.55**	-0.09	-0.19	-0.42**	-0.14	-0.29
SBW	-0.14	0.54**	-	0.61**	-0.6**	0.06	-0.26	-0.54**	-0.15	-0.07
SBCW	-0.43*	0.69**	0.61**	-	-0.45**	-0.15	-0.20	-0.41**	-0.17	-0.29
Brix°	-0.08	-0.55**	-0.6**	-0.45**	-	0.46**	-0.12	0.14	-0.14	0.46**
pH	-0.02	-0.09	0.06	-0.15	0.46**	-	-0.72**	-0.61**	-0.6**	0.56**
Acidity	0.13	-0.19	-0.26	-0.20	-0.12	-0.72**	-	0.75**	0.97**	-0.08
Tartaric	0.09	-0.42**	-0.54**	-0.41**	0.14	-0.61**	0.75**	-	0.6**	0.04
Malic	0.14	-0.14	-0.15	-0.17	-0.14	-0.6**	0.97**	0.6**	-	0.00
K	-0.02	-0.29	-0.07	-0.29	0.46**	0.56**	-0.08	0.04	0.00	-

Significance levels: * 0.05; ** 0.001

The distributions of phenotypic data in the whole population and in the two subspecies for each year are shown in the Figures 5-6. Most traits displayed a continuous variation within the subspecies. However, in the whole population NBCs, SBW, SBCW, tartaric acid and yield were clearly bimodal since cultivars and wild genotypes displayed divergent values.

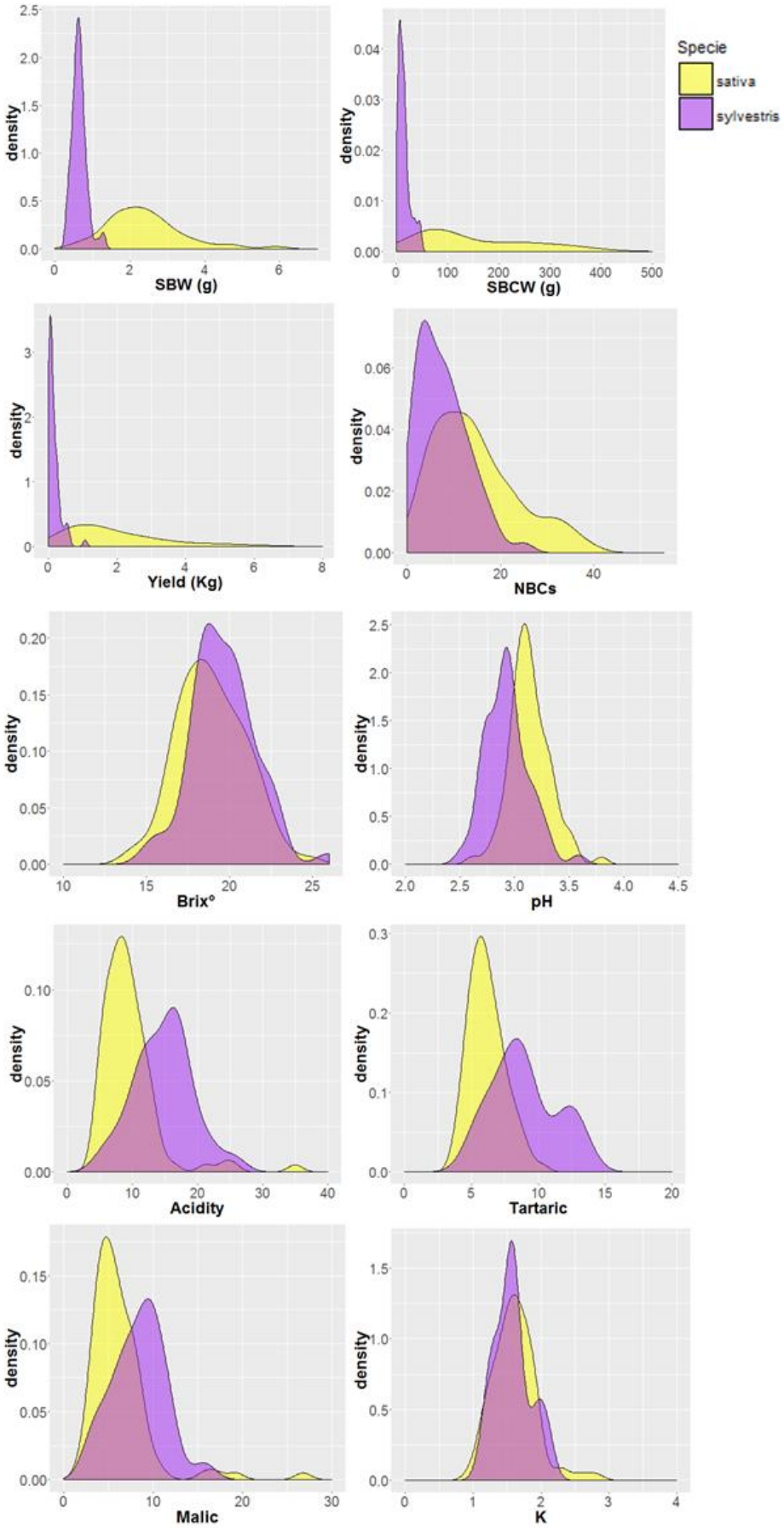

Figure 5: distribution of the average values per each trait in cultivated and wild accessions separately.

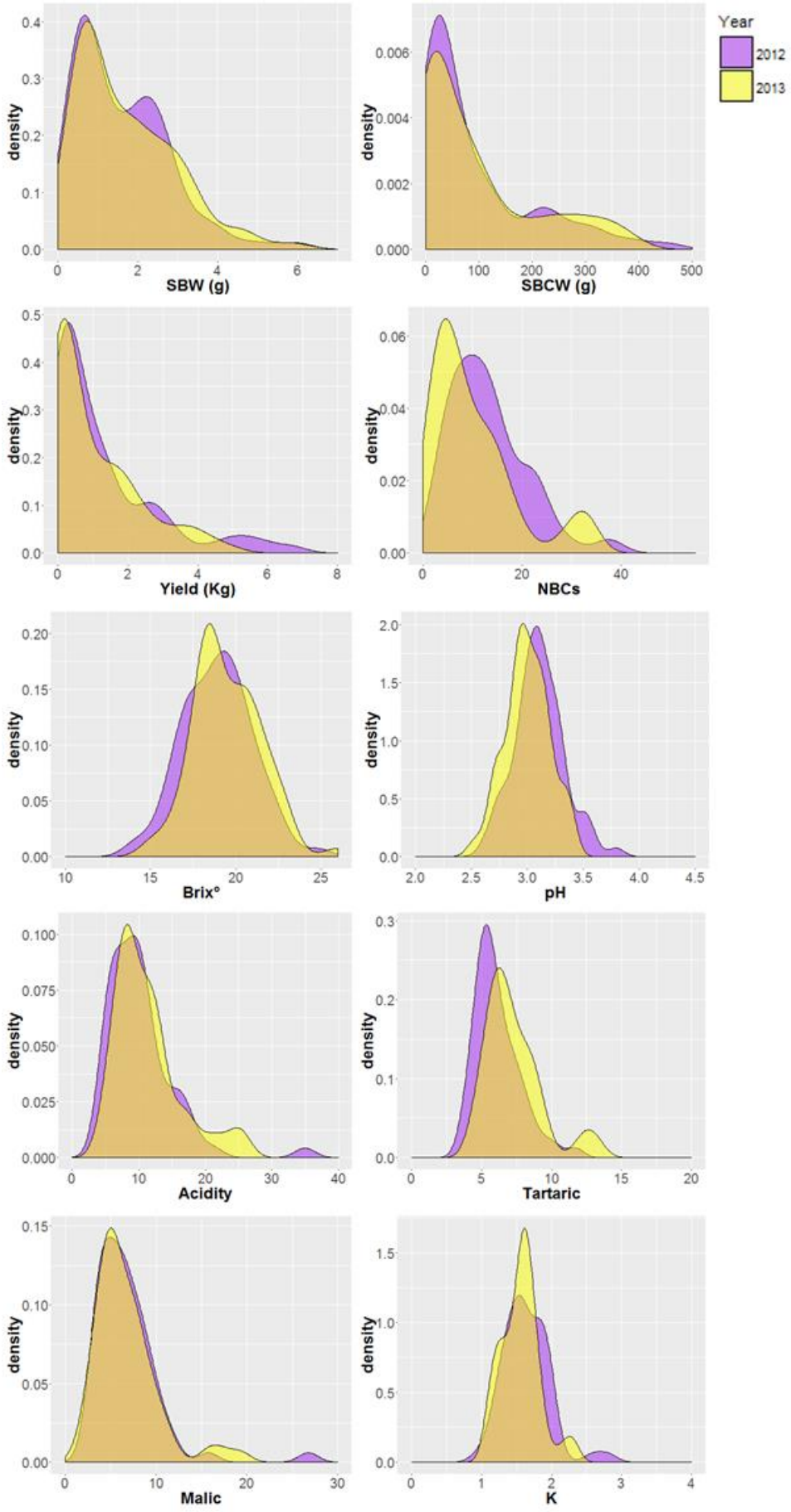

Figure 6: distribution of the average values per each trait in the two measurements year separately.

Most of the models selected to estimate heritability and BLUP included both genotypic and year effects, except for SBW, SBCW and malic concentration for which the year effect was not significant. Broad-sense heritability (H^2) was higher than 0.86 for all traits, especially for SBW and SBCW which showed the highest heritabilities (0.98).

Genome-wide associations

Association analysis for each trait was carried out testing three different models which account for either population structure (GLM (Q)), familial relatedness (MLM (K)) or both (MLM (Q+K)). MLM results with or without incorporating Q (population structure) were not materially different, suggesting how kinship matrix was sufficient to account for population stratification. For all traits GLM (with Q-matrix for K = 3 from the analysis with fastSTRUCTURE) was chosen as the best fitted model, except for SBW where MLM (K) greatly reduced false-positives compared to GLM. Indeed, Quantile- Quantile plots comparisons showed how MLM (K) produced overfitting or false-negatives for most of phenotypic variables (Supplementary Figures S1-2). The profiles of p-values (in terms of $-\log_{10}(p)$) for all tested SNPs for each trait are illustrated in Figures 7-8. Marker-trait significant associations were identified for all phenotypic variables, except for SBW where no SNPs exhibited significant p-values after multiple testing corrections (Table 6). However, 2 SNPs located on chr6 were strongly associated with single berry weight (SBW) before p-value correction, regardless of the model applied for GWAS. Moreover, different values of SBW were observed between the individuals with AA (0), BB (2) and AB (1) genotypes at both SNPs (Figure 9). The average value of SBW for genotypes AA (0) at both SNPs was 1.3 g in 2012 and 1.4 g in 2013, while the heterozygotes AB (1) showed greater values in both years ranging from 2.0 to 2.5 g. The genotypes homozygous (2) for the minor allele at both SNPs (cv ‘Alba aganin isioum’, cv ‘Ak chekerek’ and the sativa accession ‘Ak ouzioum tagapskii’) exhibited the highest values of SBW, which was on average 4.7 g in 2012 and 5.0 g in 2013 (Figure 2).

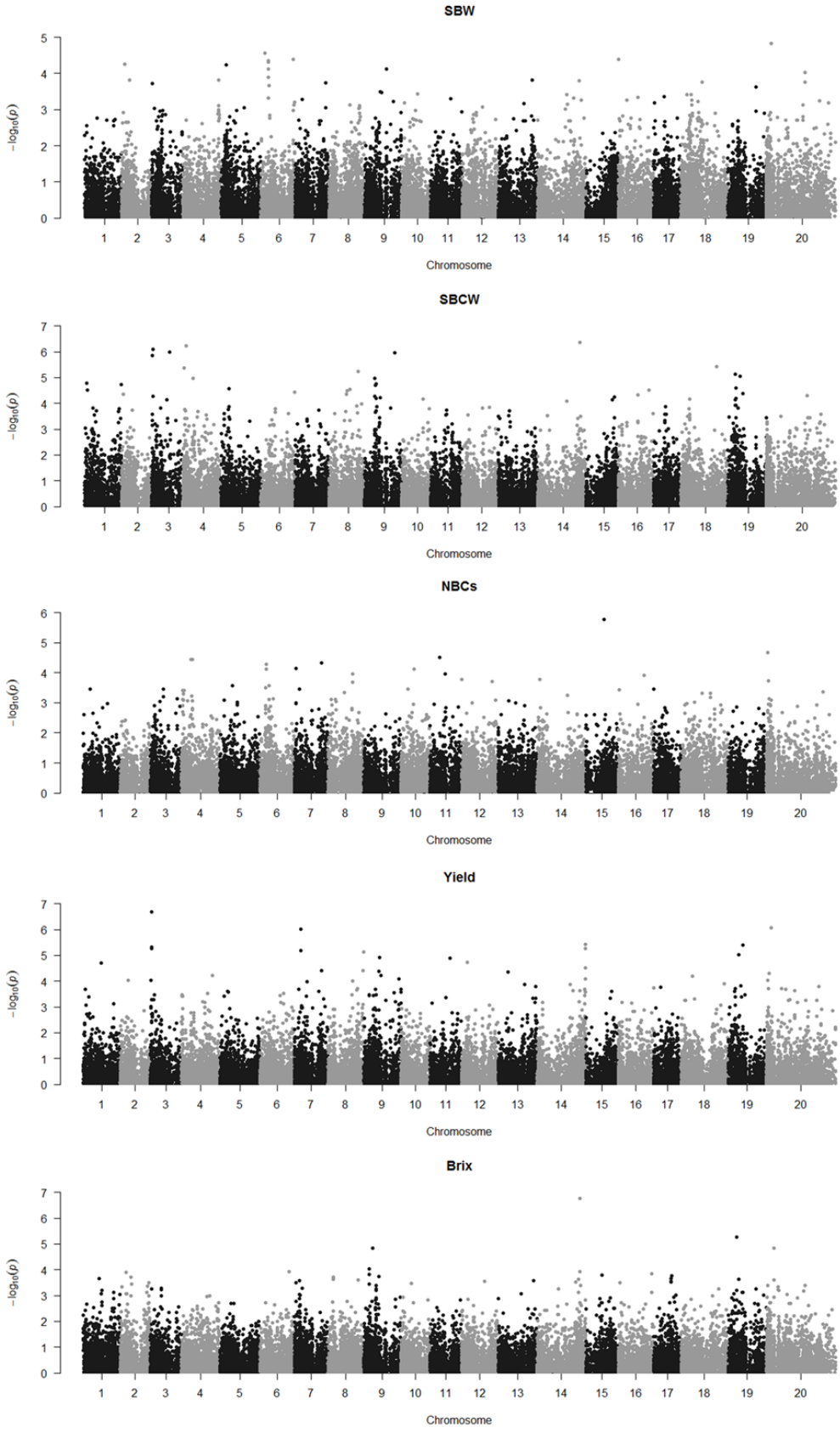

Figure 7: Manhattan plots of GWA analysis for SBW, SBCW, NBCs, yield, Brix ° traits.

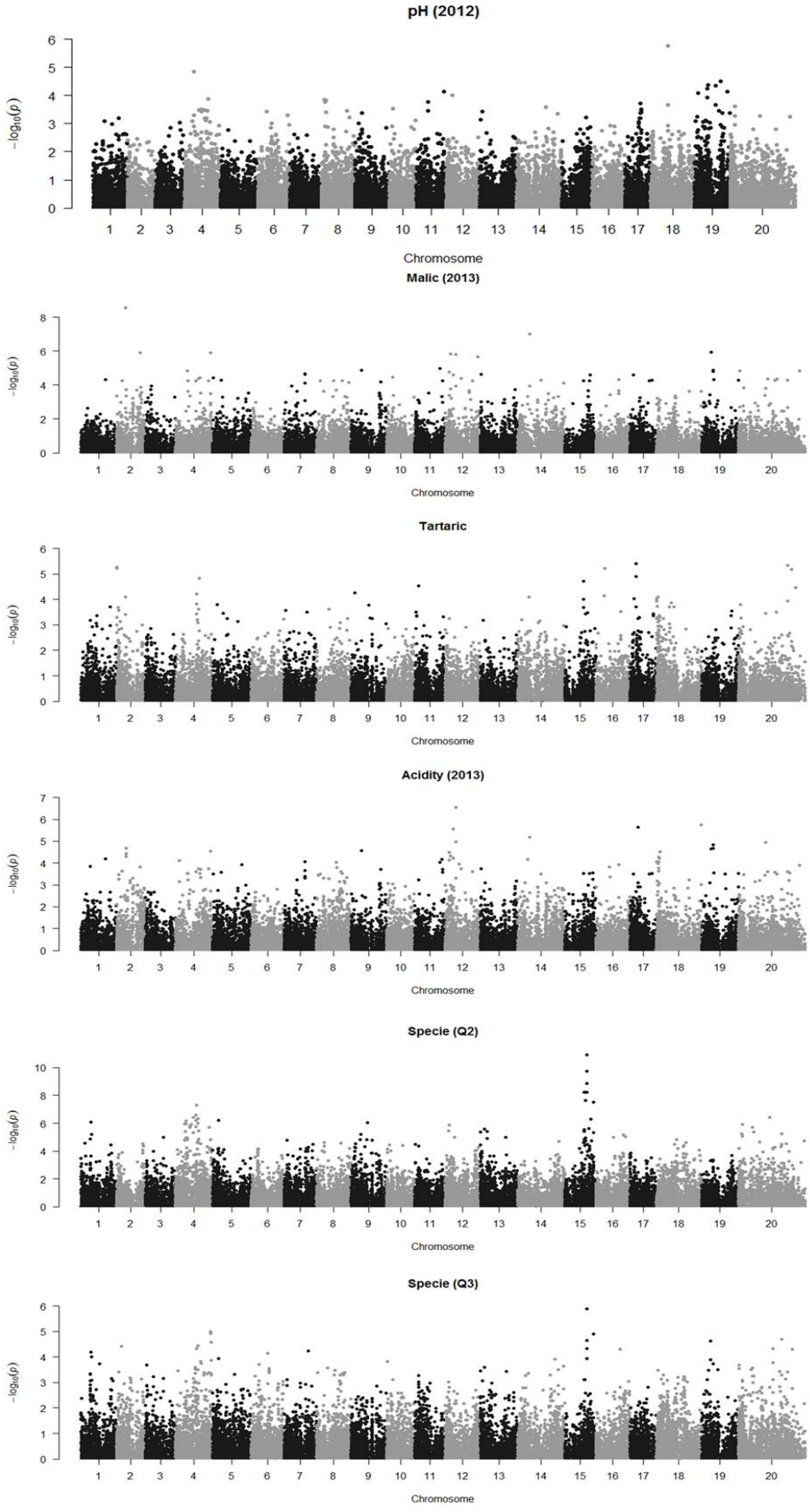

Figure 8: Manhattan plots of GWA analysis for pH (2012), malic (2013) and tartaric acids, titratable acidity (2012), and 'species' traits.

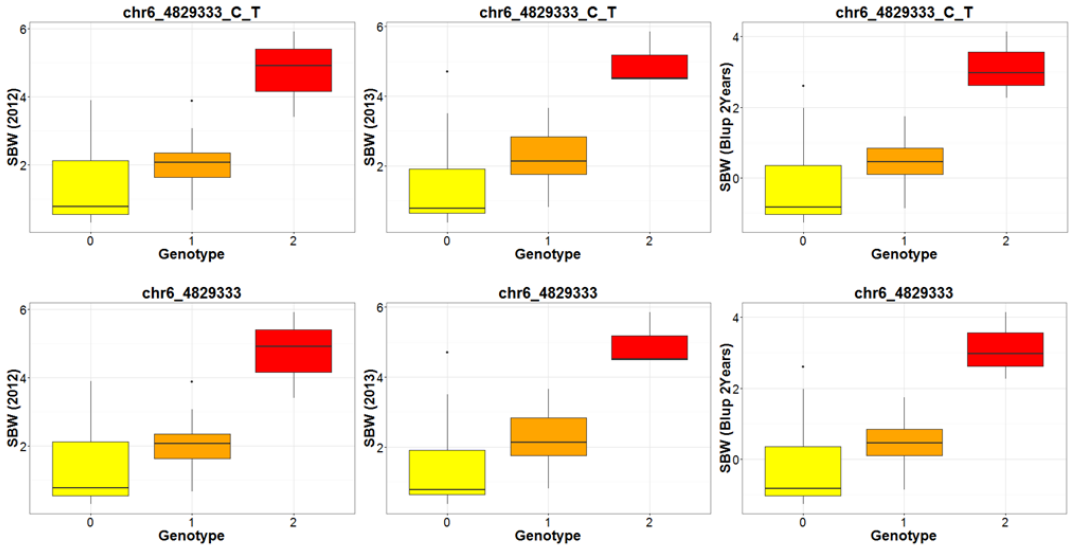

Figure 9: differences in berry size (2012, 2013, BLUP 2 years) between the three genotypes AA (0), AB (1) and BB (2) of the two most associated SNPs with SBW on chr6.

The GWAS for single bunch weight (SBCW) identified six markers associated at 5% after Bonferroni correction (Table 6). Out of these 6 SNPs, five markers located on chromosomes 14, 4, 3 and 9 were significantly correlated with SBCW in both years, while one SNP on chr19 showed a p-value < 0.05 only in 2013. Totally they explained a high proportion of observed phenotypic variance (R^2, Table 6).

Table 6: SNPs significantly associated to the ten traits analyzed, with the corresponding Bonferroni-corrected p-values. When marker-trait associations were identified in one year, the latter is indicated in brackets. Differently, SNPs without any year specification were significantly associated to traits in both years. SNPs associated to more traits are underlined. MAF: minor allele frequency. R^2: the proportion of phenotypic variance explained by the marker.

Trait	Chr	SNP	Position	Alleles	MAF	p-value	R^2
SBW	6	chr6_4829333_C_T	4829333	G\A	0.14	0.00	0.14
SBW	6	chr6_4822590	4822590	T\A	0.15	0.00	0.14
SBCW	14	chr14_26447823	26447823	C\T	0.25	0.01	0.16
SBCW	4	chr4_2286974	2286974	G\A	0.38	0.02	0.16
SBCW	3	chr3_724399_C_T	724399	G\A	0.28	0.02	0.16
SBCW	3	chr3_11296490_A_C	11296490	A\C	0.06	0.03	0.15
SBCW	9	chr9_18755332	18755332	T\C	0.30	0.03	0.15
SBCW	3	chr3_621609_C_T	621609	A\G	0.27	0.04	0.15
SBCW (2013)	19	chr19_9279384	9279384	C\A	0.28	0.00	0.17
Yield	3	chr3_621609_C_T	621609	A\G	0.27	0.01	0.19
Yield	13_random	chr13_random_2675668	2675668	A\G	0.21	0.02	0.17
Yield	7	chr7_4151125_C_T	4151125	G\A	0.23	0.03	0.17
NBCs	15	chr15_11573065_C_T	11573065	A\G	0.41	0.05	0.20
Brix°	14	chr14_26697249	26697249	C\T	0.49	0.00	0.36
pH (2012)	18	chr18_11437074	11437074	A\G	0.32	0.05	0.25
Species	1	chr1_6322315_A_G	6322315	A\G	0.23	0.02	0.03
Species	1	chr1_6366603_A_G	6366603	A\G	0.23	0.02	0.03

Species	4	chr4_6119158_A_G	6119158	G\A	0.09	0.03	0.03
Species	4	chr4_6801276_C_T	6801276	A\G	0.07	0.02	0.03
Species	4	chr4_6962355	6962355	A\G	0.09	0.03	0.03
Species	4	chr4_7097309_C_T	7097309	A\G	0.09	0.03	0.03
Species	4	chr4_9423465_A_G	9423465	G\A	0.10	0.04	0.03
Species	4	chr4_9539079	9539079	C\A	0.10	0.04	0.03
Species	4	chr4_10214943	10214943	A\G	0.11	0.03	0.03
Species	4	chr4_11771908	11771908	A\T	0.10	0.01	0.04
Species	4	chr4_11779492	11779492	T\C	0.10	0.01	0.04
Species	4	chr4_13331268	13331268	C\G	0.10	0.01	0.03
Species	4	chr4_13542485	13542485	A\G	0.11	0.02	0.03
Species	4	chr4_13633810	13633810	C\T	0.13	0.00	0.04
Species	4	chr4_14607996	14607996	T\C	0.13	0.03	0.03
Species	4	chr4_14622644_A_G	14622644	G\A	0.13	0.03	0.03
Species	4	chr4_14637406	14637406	A\G	0.14	0.01	0.04
Species	4	chr4_14651154_A_G	14651154	A\G	0.14	0.01	0.04
Species	5	chr5_3968213_G_T	3968213	C\A	0.42	0.02	0.03
Species	9	chr9_10609663	10609663	T\C	0.06	0.02	0.03
Species	12	chr12_2806062_A_G	2806062	A\G	0.13	0.03	0.03
Species	15	chr15_12863124	12863124	T\G	0.05	0.00	0.04
Species	15	chr15_12988021	12988021	C\T	0.05	0.00	0.04
Species	15	chr15_13584268	13584268	C\G	0.06	0.00	0.04
Species	15	chr15_14467891	14467891	G\C	0.06	0.00	0.05
Species	15	chr15_14532929	14532929	T\C	0.06	0.00	0.05
Species	15	chr15_14532954	14532954	G\A	0.06	0.00	0.05
Species	15	chr15_14532983	14532983	T\C	0.06	0.00	0.05
Species	15	chr15_14547396	14547396	A\T	0.07	0.00	0.04
Species	15	chr15_14547453	14547453	A\G	0.08	0.00	0.04
Species	15	chr15_16809941_A_G	16809941	A\G	0.07	0.01	0.03
Species	15	chr15_18786403	18786403	T\C	0.09	0.00	0.04
Species	18_random	chr18_random_2214072	2214072	T\A	0.22	0.03	0.03
Species	UN	chrUn_19893727	19893727	T\C	0.10	0.01	0.04
Malic (2013)	2	chr2_6004521	6004521	A\G	0.07	0.00	0.44
Malic (2013)	14	chr14_7669507	7669507	C\T	0.07	0.00	0.35
Malic (2013)	19	chr19_6331908_C_T	6331908	A\G	0.06	0.03	0.34
Malic (2013)	2	chr2_15460662	15460662	C\T	0.23	0.03	0.30
Malic (2013)	4	chr4_22974764_A_G	22974764	A\G	0.07	0.03	0.34
Malic (2013)	12	chr12_3449410	3449410	C\T	0.18	0.04	0.34
Malic (2013)	12	chr12_7131824	7131824	G\C	0.11	0.04	0.30
Tartaric	2	chr2_80304*	80304	G\C	0.23	0.03	0.18
Tartaric	2	chr2_62051*	62051	T\C	0.27	0.03	0.18
Tartaric	4	chr4_15696818*	15696818	G\T	0.41	0.04	0.17
Tartaric	16	chr16_5721952*	5721952	T\C	0.32	0.03	0.18
Tartaric	17	chr17_4154180*	4154180	C\T	0.10	0.03	0.18
Tartaric	17	chr17_4061210*	4061210	G\C	0.09	0.04	0.17
Tartaric	17	chr17_4061215*	4061215	C\T	0.09	0.04	0.17
Tartaric	UN	chrUn_31463774*	31463774	T\C	0.25	0.03	0.18
Tartaric	UN	chrUn_34044935_A_C*	34044935	C\A	0.29	0.03	0.18
Acidity (2013)	12	chr12_7131824	7131824	G\C	0.11	0.01	0.27
K	4	chr4_13542485	13542485	A\G	0.11	0.02	0.41

However, differences in bunch weight were observed between the three genotype AA, BB and AB of SNPs on chr4 and chr19, where the homozygotes for the minor allele showed highest values of SBCW (Supplementary Figure S3a). Notably the marker chr3_621609_C_T showed a significant association (p-value < 0.05) also with plant yield, which was very high for the homozygotes of the minor allele (Supplementary Figure S3b). Other two markers on chr3 (chr3_724399_C_T; chr3_754845) were associated with yield in 2012 (False-Discovery-Rate < 0.05). The high correlation between yield and SBCW (R = 0.8; Table 6) supports this cross association. In addition, two more SNPs located on chromosomes 7 and 13_random were significantly associated with yield.

Association analysis identified one SNP on chr15 and one SNP on chr14 significantly correlated with number of bunches per plant (NBCs) and total soluble solids (Brix°), respectively. In particular, the three genotypes AA, AB and BB of SNP chr14_26697249 showed divergent values of Brix° at harvest (Supplementary Figure S3c). Associations for just one year were identified for pH (2012), malic acid concentration (2013) and titratable acidity (2013). pH was correlated with a single marker on chr 18, where a long LD of 89 kb was revealed (Supplementary Figure S4). On the other hand, malic acid content exhibited significant associations with 7 SNPs located on chromosomes 2, 4, 12, 14 and 19. The SNP chr12_7131824 revealed a high association also with titratable acidity in 2013, accounting for 27% of its phenotypic variation (Table 6). This result is a further evidence of how the berry flesh acidity is strongly correlated with the berry content of malic acid, one of the most studied acids for wine production. Moreover, 9 markers, located on chromosomes 2, 4, 16, 17 and Unknown, exhibited a significant FDR-corrected association with tartaric acid concentrations. GWAS was carried out on 'species' trait codifying the *sativa* and *sylvestris* subspecies as 1 and 0 respectively. Since the analysis of population structure with fastSTRUCTURE (see chapter 3) showed two or three main groups within the association panel, GLM was applied using Q-matrix for either K = 2 (GLM-Q2) or K = 3 (GLM-Q3). 34 SNPs resulted associated to the subspecies membership, out of which 3 SNPs on chr15 exhibited significant Bonferroni-corrected associations also with GLM-Q3. In particular, 2 SNPs were located on chr1, 16 on chr4, 1 on chr5, 1 on chr9, 1 on chr12, 11 on chr15, 1 on chr18_random and 1 on chrUn (Table 6). The marker chr4_13542485 showed a singnificant association also with the potassium (K^+) concentration of the berry flesh, which had different values between the three genotypes of this SNP, notably the homozygous for the minor allele (Supplementary Figure S3c).

Candidate genes

Since LD decays below 0.2 within 10 kb (see Chapter 3), 20 kb genomic regions surrounding the SNPs significantly associated to traits were scanned to search for candidate genes. In addition, when specific LD patterns were observed around the associated markers, the full genomic regions in LD was explored. For instance, the two SNPs associated with SBW belong to a LD block of 81kb on chr6 (Supplementary Figure S5), and LD blocks were observed around the markers correlated with 'species' variable on chromosomes 1 (44 kb; Supplementary Figure S6) and 4 (62.5 kb, 129 kb, 85 kb and 85 kb; Supplementary Figures S7-10). Moreover, the association on chr3 for both SBCW and yield fell within a LD block which extended for circa 250 kb (Supplementary Figure S11), while a long LD pattern of 150 kb was observed around the SNP chr14_26697249 significantly correlated with Brix° (Supplementary Figure S12). Out of the 64 SNP loci associated with

phenotypic traits 41% were located within genes, while the remaining 38 SNPs were mainly intergenic. The genome scan for candidate genes within the regions identified by GWAS yielded 127 genes, of which 67% were in strong LD with the significant SNPs (distance < 10 kb). In particular, the number of genes ranged from 2 for NBCs to 10 for SBCW to 41 for 'Species'. The detailed list of candidate genes is shown in Table 7.

Table 7: List of candidate genes functionally annotated. Candidate genes for more traits are underlined.

Trait	Candidate	Description	Ch	Start	Stop
SBW	VIT_206s0004g03	chitinase 1	6	474255	474348
SBW	VIT_206s0004g03	protein	6	474422	474725
SBW	VIT_206s0004g03		6	475035	475195
SBW	VIT_206s0004g03	wuschel-related homeobox 3	6	475994	476034
SBW	VIT_206s0004g03	respiratory burst oxidase-like protein	6	476219	476552
SBW	VIT_206s0004g03	nuclear factor related to kappa-b-binding protein	6	477157	477938
SBW	VIT_206s0004g03	125 kda kinesin-related	6	477979	478758
SBW	VIT_206s0004g03	chitinase 2-like	6	479462	479619
SBW	VIT_206s0004g03	chitinase 2-like	6	479689	479781
SBW	VIT_206s0004g03	chitinase 2-like	6	480114	480235
SBW	VIT_206s0004g03		6	480343	480881
SBW	VIT_206s0004g03	chitinase 2-like	6	481081	481182
SBW	VIT_206s0004g03	cct motif family protein	6	481903	482056
SBW	VIT_206s0004g03	ribonuclease p subunit rpp30	6	482247	482637
SBW	VIT_206s0004g03	histone -like	6	482664	482765
SBW	VIT_206s0004g03	calcium-transporting atpase endoplasmic reticulum-	6	482798	483386
SBW	VIT_206s0004g03		6	483828	483895
SBCW	VIT_203s0038g00		3	614468	614930
SBCW	VIT_203s0038g00	nadh ubiquinone oxidoreductase b22-like subunit	3	615693	621041
SBCW	VIT_203s0038g00	30s ribosomal protein mitochondrial	3	623867	624348
SBCW	VIT_203s0038g00	epimerase family protein slr1223-like	3	624568	632364
SBCW	VIT_203s0038g00	ubiquitin fusion degradation 1	3	631573	639607
SBCW	VIT_203s0097g00	glutamyl-trna reductase	3	112918	112969
SBCW	VIT_203s0038g00		3	722303	732914
SBCW	VIT_204s0008g02	transcription factor bzip	4	228412	228711
SBCW	VIT_204s0008g02	uncharacterized protein	4	229394	229539
SBCW	VIT_214s0219g00	pentatricopeptide repeat-containing protein	14	264462	264488
SBCW	VIT_203s0038g00	arginine decarboxylase	3	644491	647420
SBCW	VIT_219s0015g01	myb-like protein h-like	19	926823	926863
SBCW	VIT_219s0015g01	myb-like protein h-like	19	930751	930811
SBCW	VIT_219s0015g01		19	934661	934683
SBCW	VIT_219s0015g01	ubiquitin-conjugating enzyme e2-17 kda	19	934910	935195
SBCW	VIT_219s0015g01	Ca^{2+} binding protein	19	937245	937673
SBCW	VIT_219s0015g01	kh domain-containing protein	19	937576	938411
Yield	VIT_203s0038g00		3	614468	614930
Yield	VIT_203s0038g00	nadh ubiquinone oxidoreductase b22-like subunit	3	615693	621041
Yield	VIT_203s0038g00	30s ribosomal protein mitochondrial	3	623867	624348
Yield	VIT_203s0038g00	epimerase family protein slr1223-like	3	624568	632364
Yield	VIT_203s0038g00	ubiquitin fusion degradation 1	3	631573	639607

Yield	VIT_203s0038g00	arginine decarboxylase	3	644491	647420
Yield	VIT_207s0005g01	pentatricopeptide repeat-containing protein	7	415225	415745
NBCs	VIT_215s0021g01	nucleoside diphosphate kinase	15	115808	115847
NBCs	VIT_215s0021g01	elmo domain-containing protein a-like	15	115829	115974
Brix°	VIT_214s0066g00		14	266966	267000
Brix°	VIT_214s0066g00	rna-binding protein cp31	14	267101	267149
Brix°	VIT_214s0066g00	cytochrome p450 724b1	14	267435	267457
Brix°	VIT_214s0066g00	gtp-binding protein gb2	14	267503	267550
Brix°	VIT_214s0066g00	pentatricopeptide repeat-containing protein	14	267576	267646
Brix°	VIT_214s0066g00	sgf29 tudor-like domain-containing protein	14	267679	267773
Brix°	VIT_214s0066g00	elongation factor chloroplastic-like	14	267780	267853
Brix°	VIT_214s0066g00	gdsl esterase lipase at5g14450-like	14	267852	267921
Brix°	VIT_214s0066g00	alpha-*l-fucosidase 2*	14	267925	267942
Brix°	VIT_214s0066g00	surfeit locus protein 2	14	267981	268020
Brix°	VIT_214s0066g00	methyltransferase pmt9	14	268013	268127
Brix°	VIT_214s0066g00	pseudouridylate synthase transporter	14	268469	268538
pH (2012)	VIT_218s0001g13	peptide transporter	18	113742	113816
pH (2012)	VIT_218s0001g13	auxin-induced protein 5ng4-like	18	113839	113874
pH (2012)	VIT_218s0001g13		18	113898	114066
pH (2012)	VIT_218s0001g13	cysteine proteinase rd19a-like	18	114096	114207
pH (2012)	VIT_218s0001g13	cysteine proteinase rd19a-like	18	114254	114276
pH (2012)	VIT_218s0001g13	V-type proton ATPase subunit a3	18	114286	114717
Species	VIT_201s0011g06	phagocytic receptor 1b-like	1	631940	632611
Species	VIT_201s0011g06	salt overly sensitive 1 (SOS1)	1	633765	639305
Species	VIT_204s0008g06	protein	4	679632	680194
Species	VIT_204s0008g06	enhancer of rudimentary	4	680506	681324
Species	VIT_204s0008g07	dynamin-related protein 3a	4	702737	710236
Species	VIT_204s0008g07	rieske iron-sulfur protein tic55	4	710644	711043
Species	VIT_204s0043g00	60s ribosomal export protein nmd3-like	4	133288	133328
Species	VIT_204s0043g00	cysteine-rich repeat secretory protein 3-like	4	133396	133416
Species	VIT_204s0043g00	tpx2 (targeting protein for xklp2) family protein	4	135327	135352
Species	VIT_204s0043g00	protein	4	135484	135525
Species	VIT_204s0043g00	transcription repressor kan1-like	4	136407	136505
Species	VIT_204s0043g00	two-component response regulator arr22	4	145986	145996
Species	VIT_204s0043g00	pentatricopeptide repeat-containing	4	146219	146831
Species	VIT_204s0043g00	hypoxia up-regulated protein 1-like	4	146921	147051
Species	VIT_204s0069g00	uncharacterized protein	4	941712	941993
Species	VIT_204s0069g01	uncharacterized transporter sll0355-like	4	942020	942612
Species	VIT_204s0079g00	gtp binding protein	4	116441	116589
Species	VIT_204s0079g00	unc93-like protein	4	117349	117354
Species	VIT_204s0079g00	acyl:coa ligase acetate-coa synthetase-like protein	4	117417	117443
Species	VIT_205s0020g02	at4g15540 dl3810w	5	396111	396530
Species	VIT_205s0020g02	sugar transporter erd6-like 16-like	5	396891	397224
Species	VIT_209s0002g09	low quality protein: patellin-3-like	9	106017	106023
Species	VIT_209s0002g09	protein	9	106023	106027
Species	VIT_209s0002g09	mitochondrial glycoprotein family protein	9	106065	106091
Species	VIT_212s0028g02	uncharacterized protein	12	280530	281021
Species	VIT_215s0021g02	uncharacterized protein	15	128630	128637
Species	VIT_215s0021g02	hypothetical protein VITISV_023274 [Vitis vinifera]	15	128674	128711
Species	VIT_215s0021g02	e3 ubiquitin-protein ligase bre1-like 1-like	15	129888	130256

Species	VIT_215s0046g01	udp-glycosyltransferase 91a1-like	15	187793	187866
Species	VIT_215s0046g01	udp-glycosyltransferase 91a1-like	15	187893	187908
Species	VIT_215s0048g00	udp-d-glucuronate 4-epimerase 2	15	144678	144692
Species	VIT_215s0048g00	udp-d-glucuronate 4-epimerase 2	15	144725	144756
Species	VIT_215s0048g00	nitrate transporter -like (NRT1)	15	145330	145335
Species	VIT_215s0048g00	peptide transporter ptr2	15	145335	145342
Species	VIT_215s0048g00	arginase	15	145346	145401
Species	VIT_215s0048g00	nitroreductase-like protein	15	145431	145472
Species	VIT_215s0048g00	uncharacterized protein	15	145473	145508
Species	VIT_215s0048g00	uncharacterized protein	15	145557	145598
Species	VIT_215s0048g02	uridylate kinase	15	168079	168096
Species	VIT_215s0048g02	protein	15	168139	168171
Species	VIT_215s0048g02	fad-binding domain-containing protein	15	168181	168215
Malic	VIT_202s0012g00	uncharacterized protein	2	599878	600106
Malic	VIT_202s0033g00	14-3-3 protein	2	154611	154615
Malic	VIT_204s0044g01	protein	4	229739	229787
Malic	VIT_204s0044g01	spotted leaf	4	229799	229813
Malic	VIT_212s0028g02	gtp-binding protein ras-like protein	12	344462	344774
Malic	VIT_212s0028g02	acyl-CoA oxidase acx3	12	344797	345561
Malic	VIT_212s0059g02	protein	12	712474	712634
Malic	VIT_212s0059g02		12	713170	713347
Malic	VIT_212s0059g02	syntaxin 1b 2 3	12	713395	713710
Malic	VIT_212s0059g02	isocitrate lyase	12	713931	714274
Malic	VIT_214s0081g00	pentatricopeptide repeat-containing protein	14	766531	766706
Malic	VIT_219s0090g00	taxane 13-alpha-hydroxylase	19	632231	632277
Malic	VIT_219s0090g00	cytochrome p450	19	632325	632705
Malic	VIT_219s0090g00	cytochrome p450	19	632738	632971
Malic	VIT_219s0090g00	transmembrane proteins 14c	19	633040	634138
Tartaric	VIT_202s0234g00	embryonic flower 2	2	59724	78128
Tartaric	VIT_202s0234g00	dna binding protein	2	79688	80318
Tartaric	VIT_204s0043g01	protein	4	156965	156976
Tartaric	VIT_204s0043g01	kinase family protein	4	156994	157038
Tartaric	VIT_216s0013g00		16	572593	573313
Tartaric	VIT_217s0000g04	atp-dependent clp protease adaptor protein	17	406268	406780
Tartaric	VIT_217s0000g04	carbon catabolite repressor protein 4-like 3	17	406891	407901
Tartaric	VIT_217s0000g04	ankyrin repeat-containing	17	408030	408154
Tartaric	VIT_217s0000g04	alpha beta-hydrolase domain-containing protein	17	412639	414220
Tartaric	VIT_217s0000g04	phytochrome and flowering time regulatory protein 1	17	414290	414445
Tartaric	VIT_217s0000g04	uncharacterized protein	17	414879	414929
Tartaric	VIT_217s0000g04		17	415208	415410
Tartaric	VIT_217s0000g04	zinc-finger domain of monoamine-oxidase a repressor	17	415540	415975
K	VIT_204s0043g00	pre-mrna-splicing factor 38b	4	134387	134558
K	VIT_204s0043g00	unnamed protein product [Vitis vinifera]	4	135030	135033
K	VIT_204s0043g00	tpx2 (targeting protein for xklp2) family protein	4	135115	135321
K	VIT_204s0043g00	tpx2 (targeting protein for xklp2) family protein	4	135327	135352
K	VIT_204s0043g00	protein	4	135484	135525
Acidity	VIT_212s0059g02	protein	12	712474	712634
Acidity	VIT_212s0059g02		12	713170	713347
Acidity	VIT_212s0059g02	syntaxin 1b 2 3	12	713395	713710
Acidity	VIT_212s0059g02	isocitrate lyase	12	713931	714274

Discussion

GWAS limitations in the association panel

GWA studies represent a new tool in agricultural genetics for revealing the genetic bases of phenotypic variation [262]. The GWAS approach achieves higher mapping resolution than traditional methods by taking full advantage of ancient recombination events occured during the successive generations separating common ancestors from individuals in the studied population [145]. However, both population stratification and familial relatedness among individuals can lead to spurious marker-trait associations [263]. According to Yu et al. [133], in order to yield the largest QTL power in GWAS, the ideal population should have the lowest structure and relatedness. In this sense, family-based population design with controlled parent crosses has been suggested in apple [139] as well as nested association mapping (NAM) or multiparent advanced generation inter-cross (MAGIC) populations have been constructed in maize [284, 285], *Arabidopsis* [286], barley [287] and wheat [288]. Creating such materials in grapevine could be time-consuming and expensive for the space required by the large size of the sample plants. Moreover, in perennial species such as grapevine a large part of the unique genotypes available in the germplasm collections are closely related since a small number of elite cultivars appears to have been used for breeding [51]. These difficulties may explain why a few GWA studies were attempted so far in fruit trees and how their association panels, usually consisting of 100-200 individuals [140], are smaller than those used for annual species. Recently, Nicolas et al. [113] designed an association panel of 279 grapevine genotypes by selecting key founder varieties of modern cultivars and removing their first-degree relatives, in order to performe future GWA studies. Even though our association panel is 3 times smaller than the population defined by Nicolas et al. [113], it comprises a good number of wild *vinifera*, which presents unexplored variation for quality and yield-related traits as well as for adaptation to environmental stresses. Moreover, for the first time such high number of *sylvestris* genotypes were phenotyped for traits of commercial interest, that is berry size and composition, whose genetic basis have been extensively investigated in previous works through classical bi-parental QTL mapping in cultivated varieties [228, 42, 43, 123]. The phenotypic variation observed for all traits in the whole population and separately in the subgroups of wild and cultivated grapevines makes our association panel suitable for applying association mapping in order to identify domestication-related genes [126]. In addition, we performed GWAS with 26K SNPs evenly distributed across the grapevine genome, which provide much higher resolution than that of Nicolas et al. [113], which used 501 SNPs, and of Myles et al. [34] and Chitwood et al. [45], that employed the same 5K SNPs matrix. Although our SNP panel represents just one fourth of the 90K SNPs simulated for GWAS in grapevine by Fodor et al. [187], a step towards increasing the power and the resolution of GWAS in grapevine has been done practically in our study. Indeed, as in maize [289], the rapid LD decay observed in the grapevine genome makes it a promising model species in GWAS [145] with single-gene resolution. Future availability of whole-genome sequences for numerous grapevine genotypes will satisfy the demand of tens of millions of SNPs for association mapping studies in grapevine. Furthermore, we accounted for confounding factors in GWAS by applying both GLM and MLM models, showing how the former was good enough for avoiding spurious associations due to the genetic structure of population. In addition, the stringent p-values corrections of Bonferroni or FDR should ensure a low rate of false-positive genotype-phenotype associations in our experiment.

Candidate genes controlling berry weight and yield

The application of association mapping is an alternative approach to identify and interpret the genetic basis of the phenotype shifts associated with domestication [126]. Indeed, studying domestication traits is important not only from an evolutionary point of view, but also in agricultural, economic and social contexts [290]. Despite the limited size of the association panel, we identified significant genotype-phenotype associations for almost all traits studied, which presumably include the selected characters during the transition from wild relatives to cultivated grapevines [25]. An exception was represented by SBW, for which no marker exhibited significant p-values after the multiple testing correction. This result may highlight the complex genetic architecture of berry weight, which is strictly correlated to berry size and is influenced by seed content [123] . Moreover, SBW may present genetic heterogeneity, where different variants may underlie a trait with a maximized genetic variance [291]. However, we identified two SNPs on chr6 associated with SBW in all applied GWAS models before the multiple testing correction of p-values. These SNPs, separated by 6.7 kb, fell in a LD block of 81 kb. In this genomic region, we found 5 genes encoding chitinases, known to be part of the Systemic Acquired Resistance strategy [292], which acts to prevent pathogen attack during berry development [234]. Furthermore, we identified a wuschel-related homeobox 3 gene, 74 kb apart from the most significant SNPs for berry weight. This candidate gene encode a member of the plant WOX family, whose genes have been shown to play a broad role in plant development, notably for meristem maintenance [293]. Another member of the WOX family is the WOX13 homeobox gene, which promotes replum formation in the *Arabidopsis thaliana* fruit [294]. A Ca^{2+} transporting ATPase endoplasmic reticulum-type-like gene (Tables 6-7) was also found among the candidate genes for SBW, supporting the role of calcium ion in the development of grape berries [295]. Indeed, Ca^{2+} has a central role in cell signalling, in maintenance of cell wall integrity [296] and in the vacuole as counter-cation for organic and inorganic anions. Low concentrations of cytosolic Ca^{2+} are required for normal cell function. Therefore, calcium homeostasis in the cytosol is tightly controlled by membrane transporters which work to keep Ca^{2+} at low concentrations in the cytosol. A large number of genes with functions related to calcium sequestration, transport and signalling have already been found to display developmentally regulated expression patterns [234]. The identification of the Ca^{2+} binding protein gene on chr19 as associated in 2013 with single bunch weight (SBCW) supports the central role of calcium in fruit development (Table 7). In addition, genes encoding for transcription factors MYB-H and bZIP22 were strongly associated to SBCW, highlighting how changes in developmentally and morphologically complex traits, including many domestication traits, occurred through selection on transcriptional regulators [142]. Moreover, previously studies showed how bZIP factors are involved in the ABA-dependent processes of response to abiotic stresses [297] and grape berry ripening [298], as well as in the regulation of flavonoid biosynthesis in grapevine [299]. A cross association between SBCW and yield traits was detected on chr3, where candidate genes involved in cellular respiration were identified (Tables 6-7). We also found the arginine decarboxylase gene (adc2) 10 kb apart from the marker chr3_621609_C_T significantly associated with both SBCW and yield. The ADC is involved in the biosynthesis of polyamines [295], growth regulators that have been implicated in several development processes and biotic responses [236]. In grape, a reduction in polyamines content was observed during berry development, reaching the lowest value at maturity [300, 301, 302]. It has been supposed that polyamines are important during early stages of fruit development,

notably promoting cell divisions [301]. However, gene expression studies during berry ripening have showed how genes coding for arginine decarboxylase increased their transcript abudance at the beginning of ripening and remained high in mature fruit [234, 303].

Candidate genes involved in flesh berry composition

Variation of berry composition was observed between *sativa* and *sylvestris* individuals. In particular, the former showed higher K^+ concentrations and pH, while the latter exhibited higher flesh berry acidity, notably for tartaric acid content (Figure 1). Berry composition undergoes several changes throughout the double sigmoidal growth cylcle of the fruit[304]. In particular, during the first phase organic acids, mainly malic, tartaric and citric acids, accumulate in the vacuoles which undergo intense enlargement. At the end of the lag phase, the véraison is charachterized by the onset of sugar and anthocyanins accumulation, which results in increasing of flesh berry swetness and pigmentation [236]. Tartaric and malic acids are the predominat organic acids in the grape berry, accounting for over 90% of its total acidity [295]. They also contribute greatly to the pH of the juice, must and wine during vinification and subsequent wine ageing [295]. Tartaric acid concentrations in ripe berries reflect the extent to which its synthesis occurred during the first stages of berry development [305]. Indeed, tartaric acid can be found in grapevine flowers and its levels increase in the berry during the four weeks after anthesis [306]. We identified two genes involved in the control of flowering in plants as candidate genes for tartaric acid content in grape berries: the embryonic flower 2 gene, whose role as repressor of reproductive development in phase transitions has been shown in *A. thaliana* [307], and the phytochrome and flowering time regulatory protein 1 gene, which encodes for a nuclear protein involved in the regulation of flowering time by light quality [308]. Unlike tartaric acid, the levels of malic acid in grape berries change during fruit development. In particular, malic acid formed in the berry pre-véraison is broken down during ripening, when malate becomes a substrate for the TCA cycle, the gluconeogenesis and the aerobic fermentation [309]. We identified the isocitrate lyase gene among the candidate genes for malic acid concentration in flesh berries. The isocitrate lyase is one of the five enzymes involved in the glyoxylate cycle, which converts acetyl-CoA into succinate via a series of reactions concerning malate and citrate [309]. The glyoxylate cycle may contribute to malate accumulation in young berries [310]. On the other end, the glyoxylate cycle may fuel the gluconeogenesis pathway for the synthesis of glucose by supplying malate during berry ripening. In this way, the glyoxylate cycle also contributes to the reduction of fruit acidity through the consumption of malic acid [309]. The identification of a second association between the SNP in LD with isocitrate lyase gene and the titratable acidity trait supports the central role of glyoxylate cycle in fleshy berry acidity. The vacuole, which can occupy more than 99% of the total intracellular volume in grape berries, has a pivotal role in the storage of organic acid and sugars as well as in the control of cytoplasmic pH [311]. Indeed, the SNP on chr18 significantly associated to berry juice pH falls within the V-type proton ATPase subunit a3 gene. The V-ATPase is one of the primary electrogenic pumps on tonoplast [312] and converts the chemical energy of ATP in an electrochemical proton gradient allowing the transport of many solutes against their electrochemical gradient by specific transport systems [313].

An overall measure of the solutes (largely sugars) in flesh berries is the Brix degree [283], which is usually used as an indicator of the proper berry maturity for quality wines [314]. A long

LD block of 150 kb around the genomic regions associated to Brix° trait was identified on chr14. Twelve genes are located within this region. The cytochrome p450 724b1 gene is implicated in the biosynthesis of brassinosteroid (BR) [315], plant hormones essential for normal plant development. A dramatic increase in endogenous BR levels was observed at the onset of berry ripening, as indicated by the simultaneous increase in berry weight and soluble solids (Brix°) [316]. A role of BR in the regulation of anthocyanin biosynthesis during ripening of grape berries has been invistigated recently [317], showing their effect mainly on downstream genes of anthocyanin biosynthesis. Indeed, anthocyanins accumulation in red grape varieties occurred since véraison. Accordingly another candidate gene for Brix° was the one encoding the methyltransferase PMT9, putatively involved in Arabidopsis anthocyanin biosynthesis [318]. Moreover, the genomic region of 150 kb associated with berries solutes content includes the α-l-fucosidase 2 gene, which is involved in the metabolism of the hemicellulosic polysaccharide xyloglucan (XyG), the dominant component of plant cell wall [319]. Indeed, the α-fucosidase is a glycosylhydrolase that acts on the XyGs once deposited on the cell wall, contributing to its reassembling during cell elongation and releasing fucose residuals in the cytosol.

Candidate genes discriminating cultivated and wild grapevines

Grape berries has K^+ as major cation, which is involved in several physiological processes, such as enzyme activation, cellular transport processes, anion neutralisation, and osmotic potential regulation [295]. A cross correlation between K^+ and 'species' traits was identified on chr4. The gene encoding a tpx2 (targeting protein for xklp2) family protein is located 9.7 kb apart from the SNP significantly associated with both K^+ and 'species' traits. TPX2 acts as a spindle assembly factor during mitosis as well as partecipates as a microtubule associated protein (MAP) in microtubule dynamics [320]. Therefore, the efficiency of spindle formation during cell proliferation as well as the microtubule metabolism during cell elongation may be addressed among the factors which influence berry size, one of the main domestication traits in grapevine [25]. The highest number of marker-trait associations was identified for the trait 'species', accounting for the level of genetic differentiation between cultivated and wild grapevines [155]. Notably, on chr15 we identified the nitrate transporter –like NRT1 gene, which showed significant p-values in both GLM-Q2 and GLM-Q3. The variation of NRT1.1B has been correlated with divergence in nitrate-use between the subspecies *Oryza sativa* L. *indica* and *japonica* [321]. NRT1.1B encodes a protein containing a peptide-transporter domain and is localized to the plasma membrane. The analysis of nucleotide diversity within this gene indicated that NRT1.1B underwent a positive selection during *indica* domestication process, leading to the higher nitrate-use efficiency of *indica* compared to *japonica* [321]. In agreement with the genome scan for signatures of selection reported in chapter 3, the GWAS test on 'species' trait led to identify genes involved in the response to environmental stresses. The salt overly sensitive 1 (SOS1) gene encodes a Na^+/H^+ antiporter, which is the downstream target of the Salt Overly Sensitive (SOS) signaling pathway, involved in controlling ion homeostasis during salt stress [322]. In particular, SOS1 acts by extruding the toxic excessive Na^+ from the cytosol [323]. In this sense, SOS1 is critically required for salt tolerance [322]. In addition, we identified the hypoxia up-regulated protein 1-like (HRE1) gene, whose product is an ERF transcription factor. HRE1 responds rapidly to oxygen deprivation by maintaining the expression of some anaerobic genes, such as the alcohol dehydrogenase (ADH) gene [324]. Hypoxia has been also associated with cell death in mesocarp

of winegrapes late in berry ripening due to high temperature and water stress [325]. Finally, the identification of the arginase gene, involved in the biosynthesis of polyamines [295], and the sugar transporter erd6-like 16-like gene, which encodes a monosaccharide transporter [326], highlighted how the *sativa* and *sylvestris* may present differences in the metabolisms of polyamins and sugars.

Conclusions

We scan the grapevine genome for significant allelic variation underlying domestication-related traits by applying GWAS approach. A considerable phenotypic variation was observed between and within the two *V. vinifera* subspecies, highlighting how our association panel will be useful in future GWA studies to further explore the consistent genetic variation still maintained within natural populations of grapevine. Several candidate genes were identified for most of the traits analyzed. In particular, our findings provided further evidence of how differences in the complicated interplay between transcription regulators, cell signalling factors and hormones, may be the basis of the phenotypic variation observed in berry and bunch weight between *sativa* and *sylvestris* individuals. Moreover, the significant allelic variation identified in candidate genes directly involved in the control of berry composition, notably of pH, malic acid concentration and titratable acidity, highlights multiple avenues for further works about the functional roles of the genes implicated, putative genetic pleiotropy between traits and GxE interactions. Finally, we presented a proof-of-concept of association mapping in grapevine, supporting its relevance as an efficient genetic tool to discover and reconstruct the genetic architecture of complex traits in a challenging genetic system.

Supplementary Data

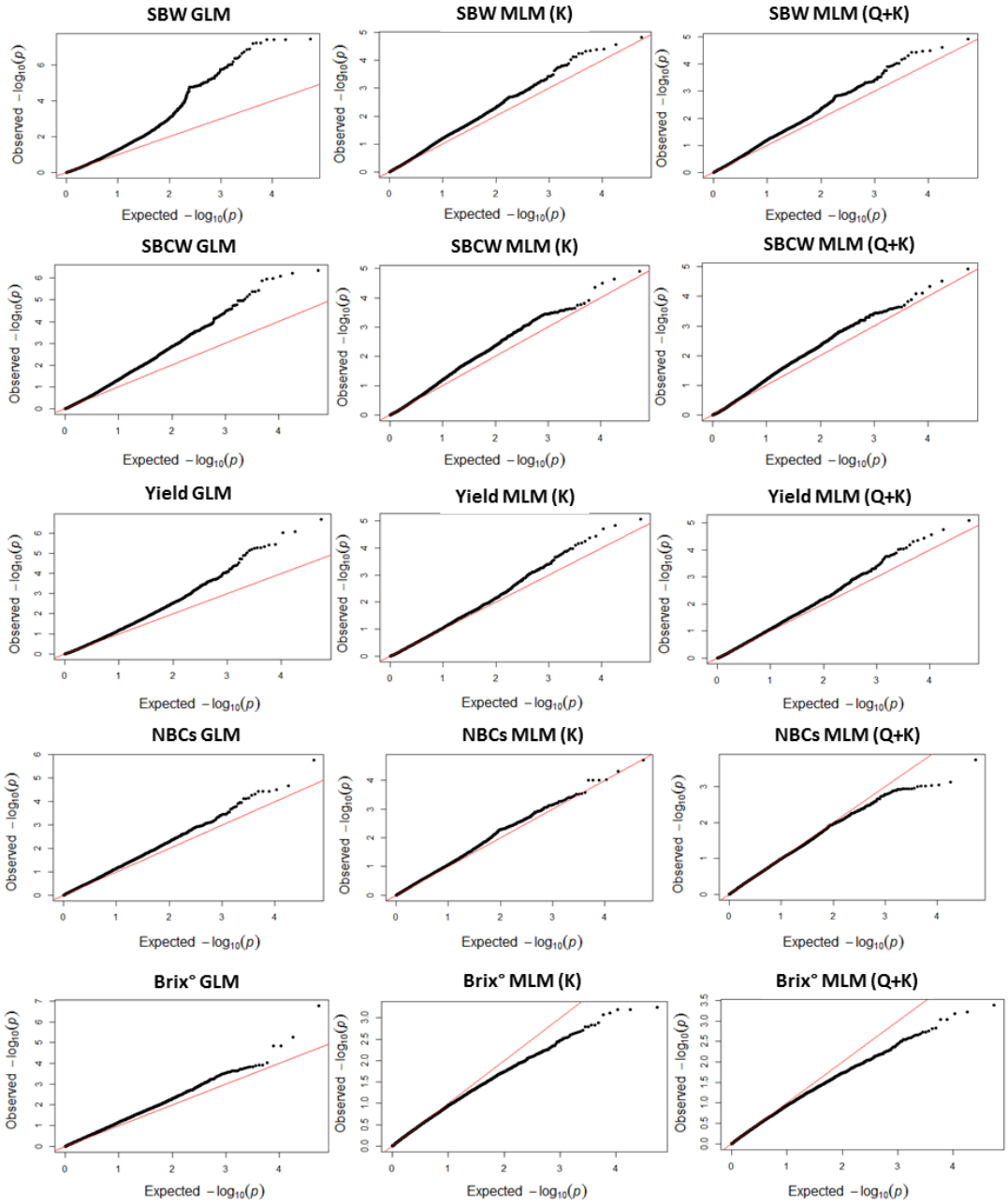

Figure S1: Q-Q plot of GLM, MLM (Q+K) and MLM (K) models used for GWAS test for SBW, SBCW, yield, NBCs, Brix° traits.

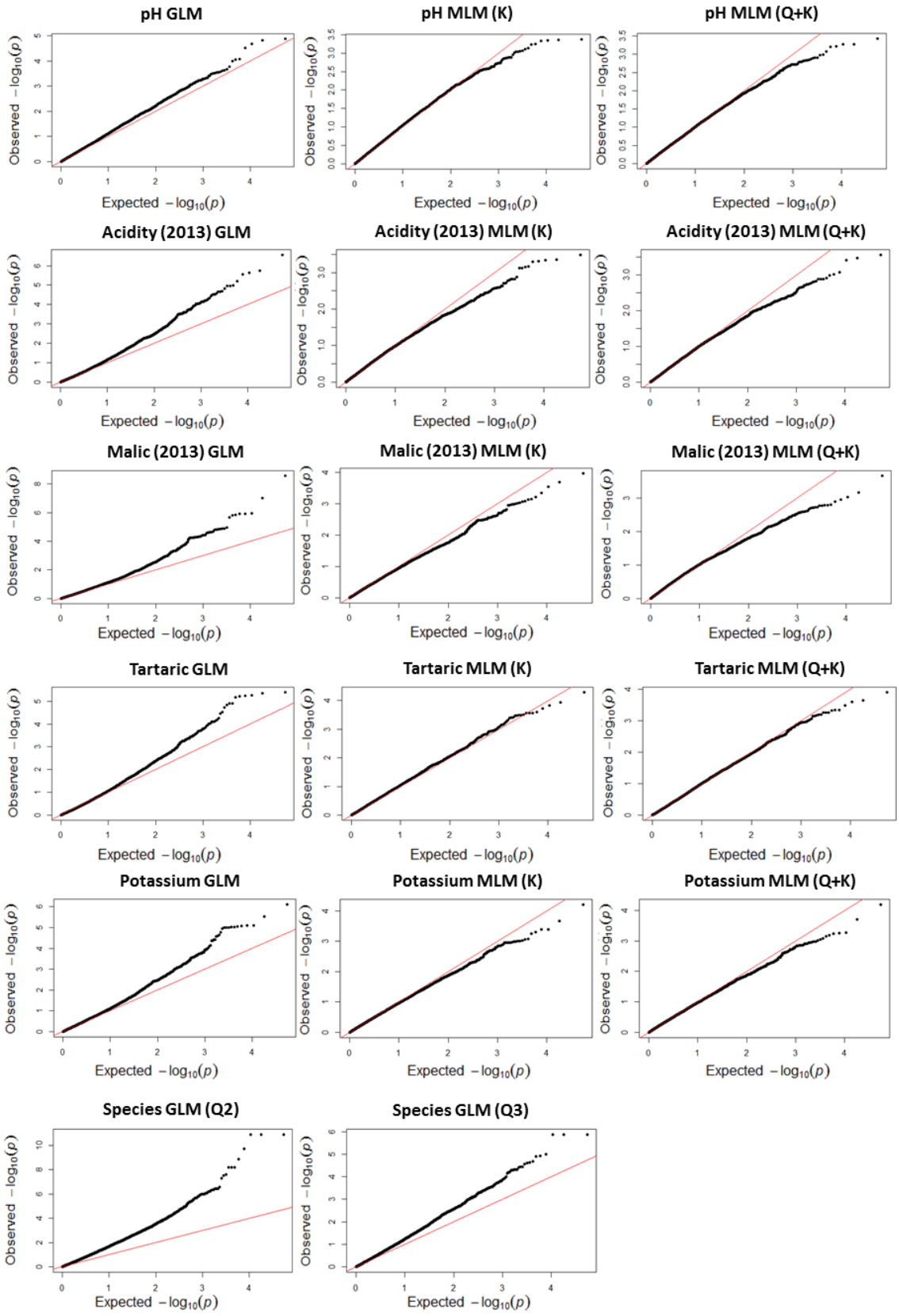

Figure S2: Q-Q plot of GLM, MLM (Q+K) and MLM (K) models used for GWAS test for pH, titratable acidity, malic and tartaric acid concentrations, K^+ content and 'species' traits.

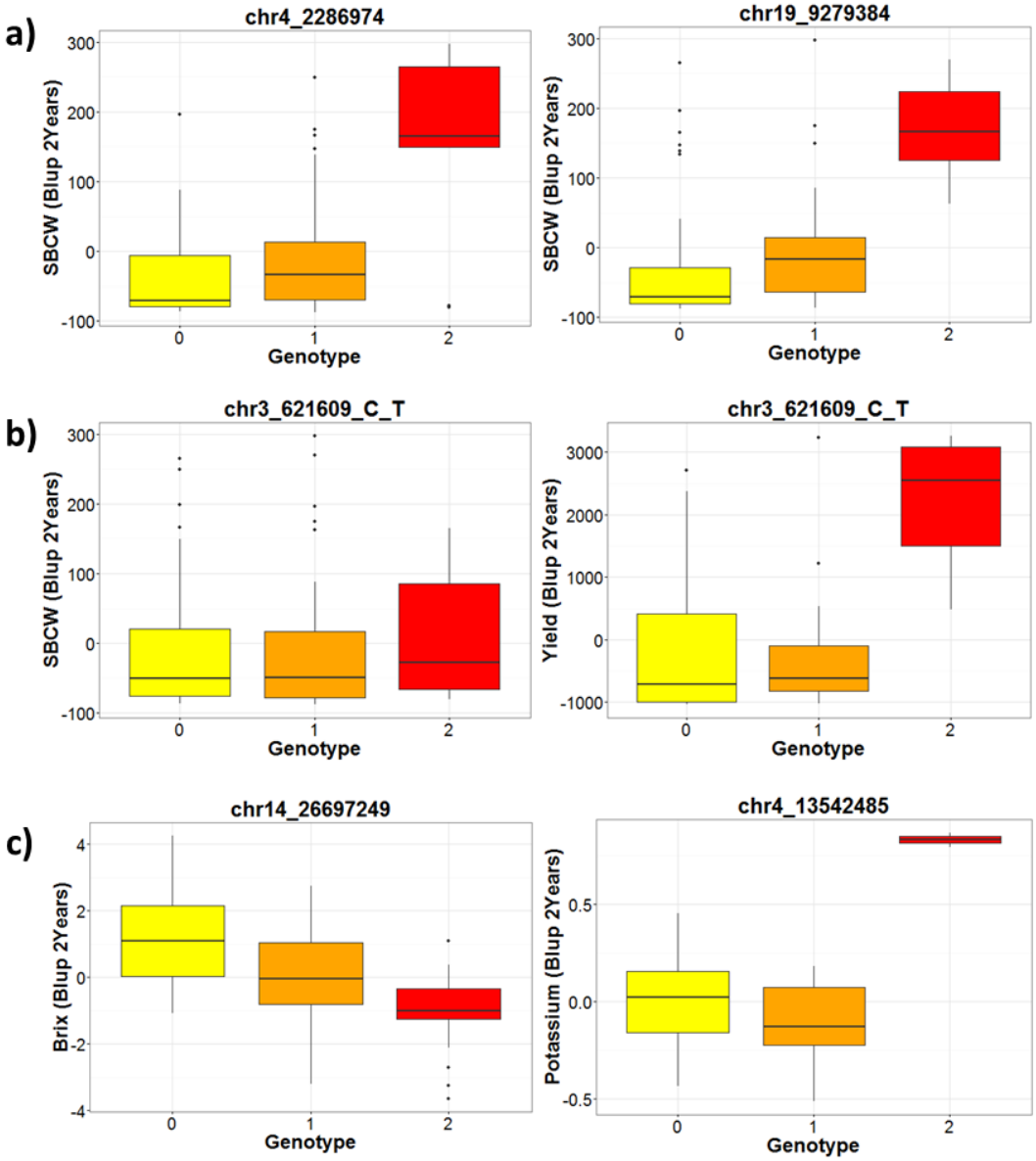

Figure S3: differences in SBCW **(a-b)**, yield (b), Brix° and K^+ **(c)** content between the three genotypes AA (0), AB (1) and BB (2) of the most associated SNPs.

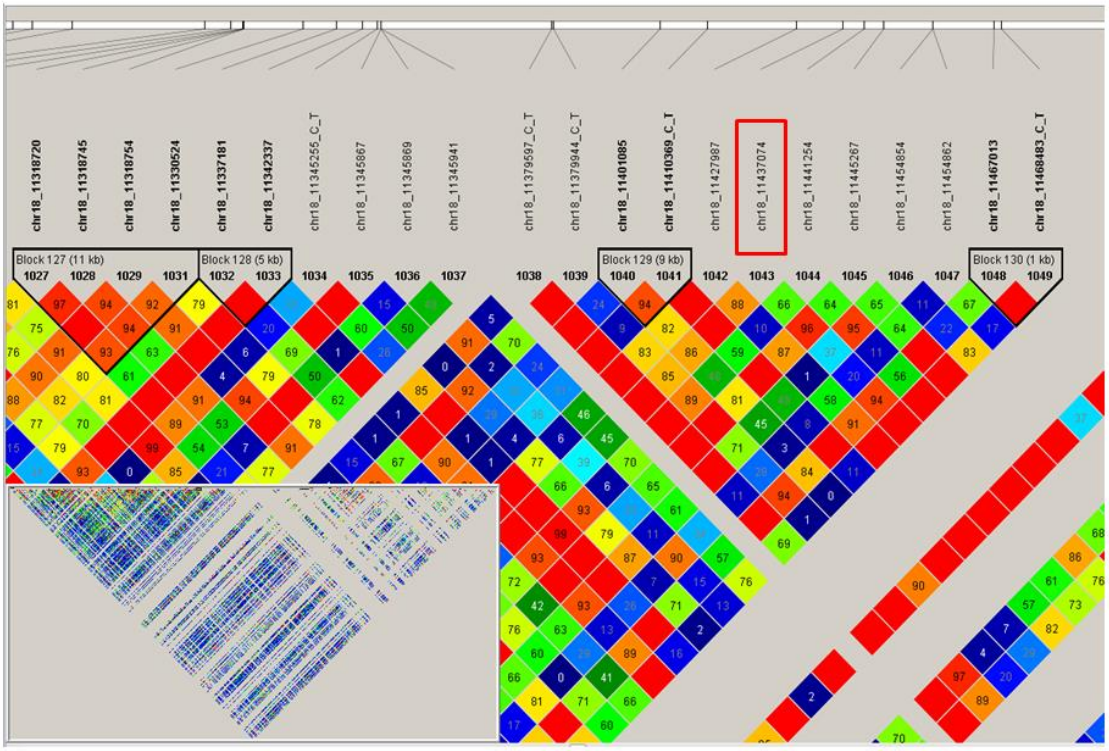

Figure S4: LD block on chr18 around the marker (in the red box) associated to pH (2012)

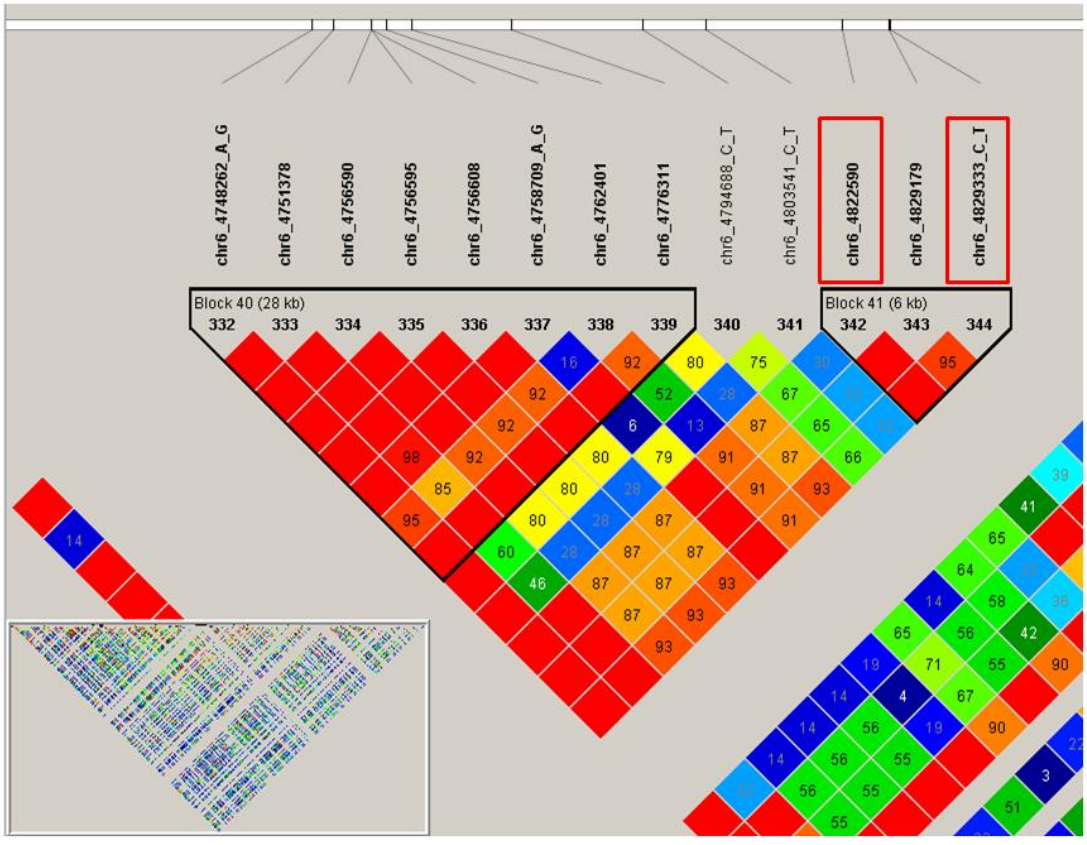

Figure S5: LD block on chr6 around the two markers (in the red box) associated to SBW.

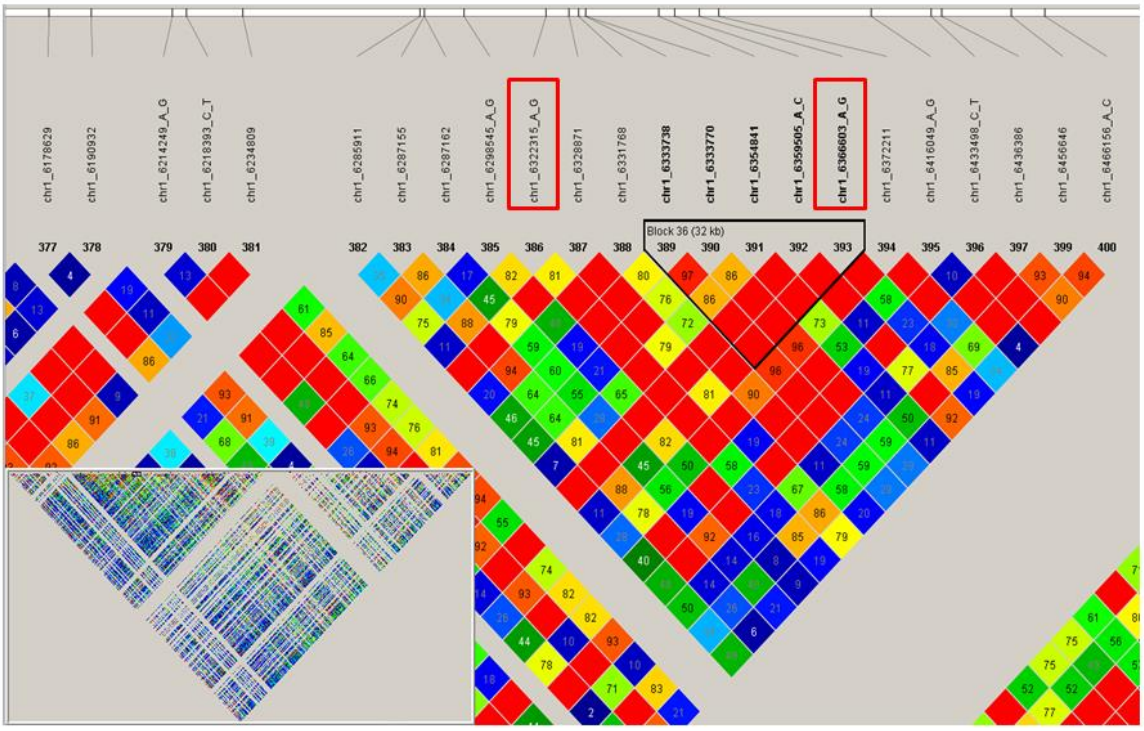

Figure S6: LD block on chr1 around the two SNPs (in the red box) associated to 'species' trait.

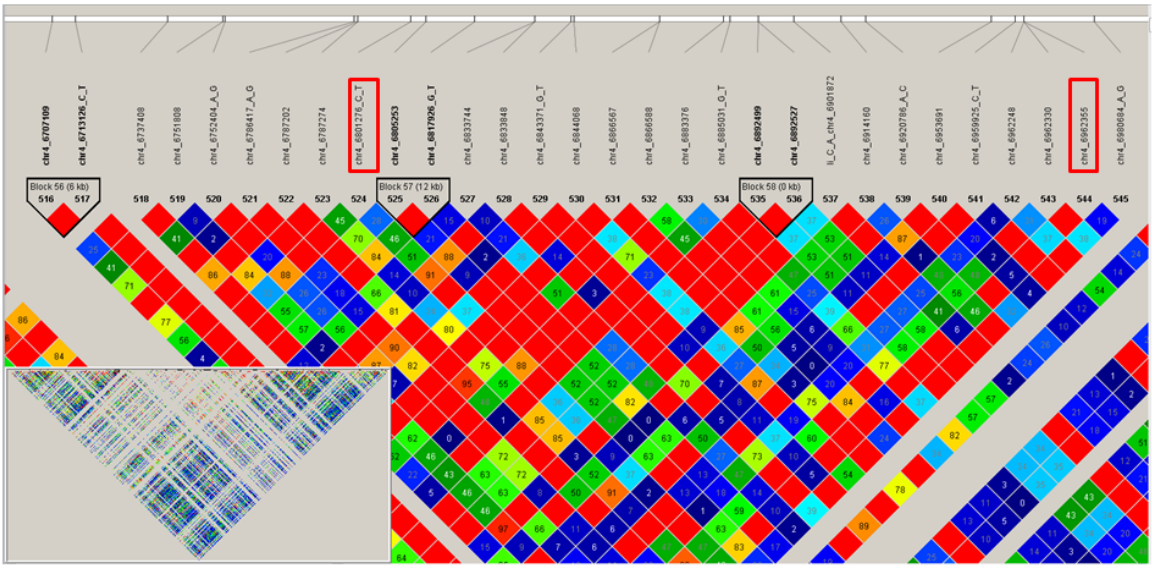

Figure S7: LD block on chr4 around the two SNPs (in the red box) associated to 'species' trait.

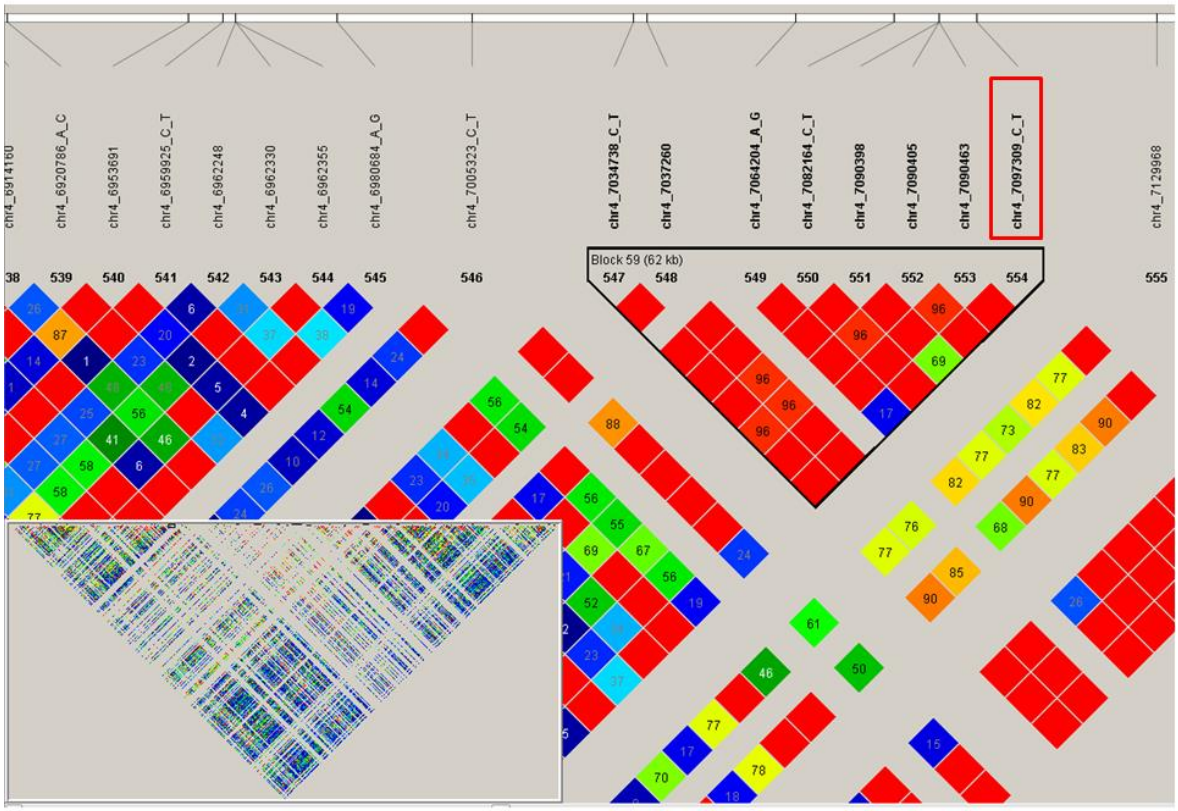

Figure S8: LD block on chr4 around the SNPs (in the red box) associated to 'species' trait.

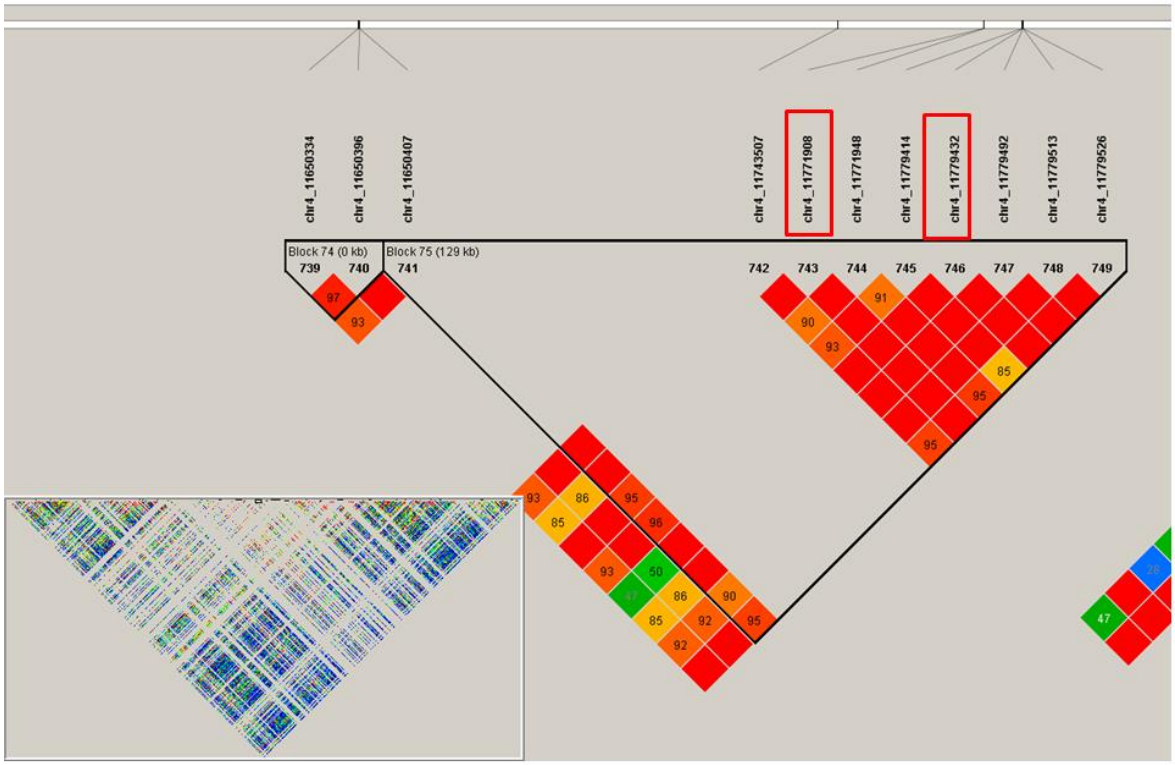

Figure S9: LD block on chr4 around the two SNPs (in the red box) associated to 'species' trait.

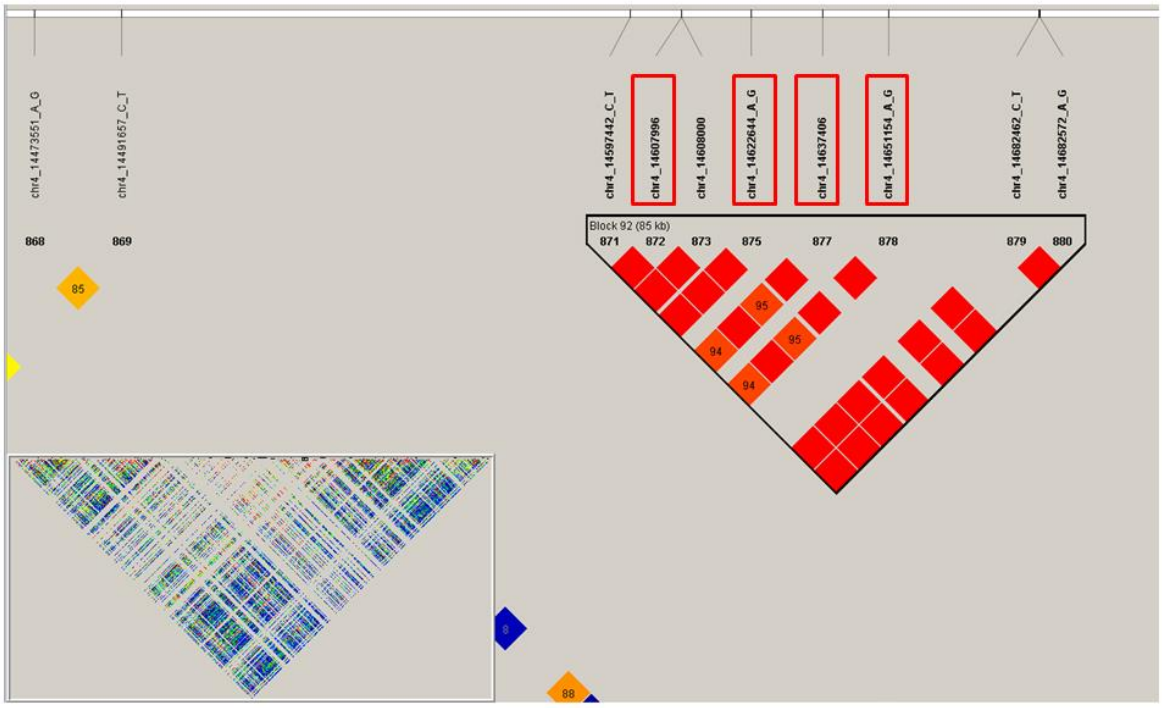

Figure S10: LD block on chr4 around the four SNPs (in the red box) associated to 'species' trait.

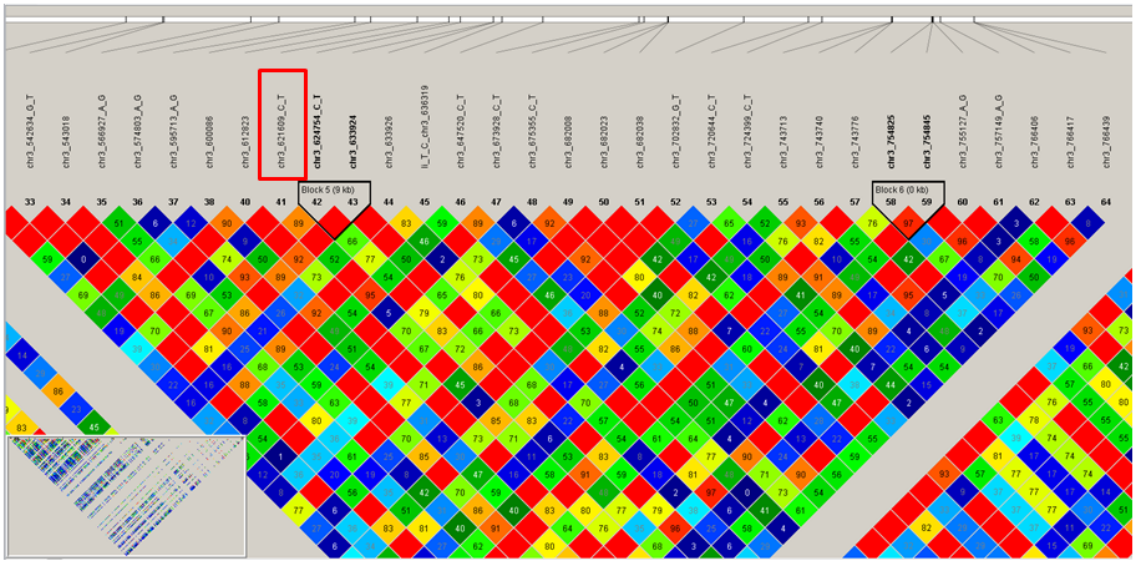

Figure S11: LD block on chr3 around the SNP (in the red box) associated to both SBCW and yield traits.

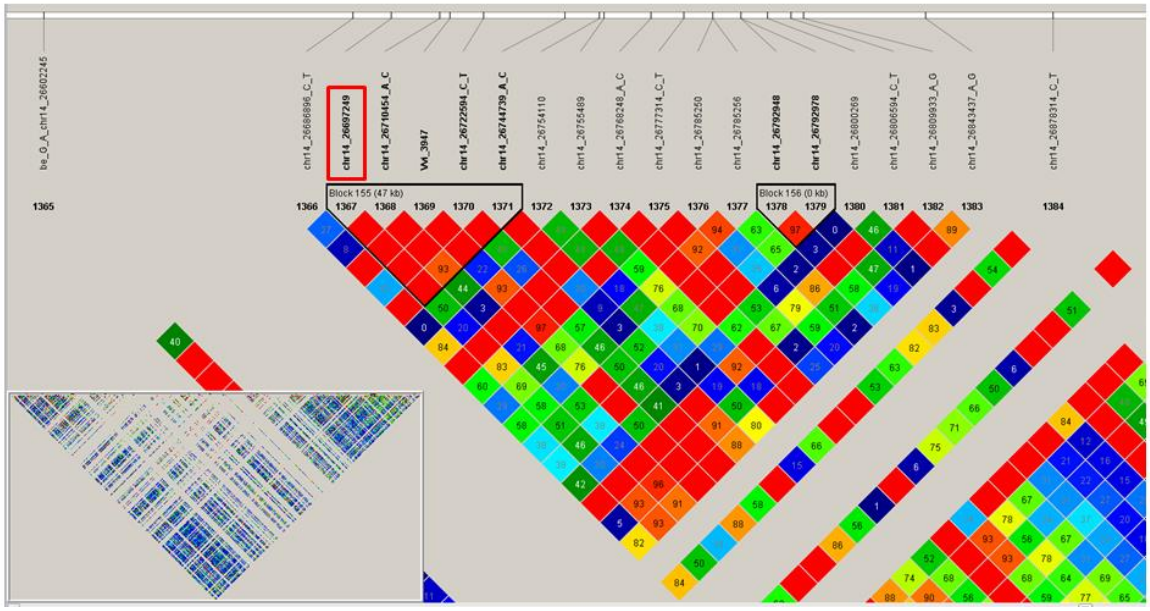

Figure S12: LD block on chr14 around the SNP (in the red box) associated to Brix°.

Conclusions

Crop plants used nowadays in modern civilization resulted from several thousand years of conscious as well as unintentional human selection, which transformed wild ancestors into high-yielding and useful domesticated descendants [142]. During this domestication process, crops underwent several phenotypic changes, commonly known as the ‘‘domestication syndrome’’ [144]. Characterizing the genetic architecture of domestication-related phenotypes gives a powerful lens for understanding the process of adaptation in nature, as Charles Darwin noted in the introduction to his famous book [327]:

“*At the commencement of my observations it seemed to me probable that a careful study of domesticated animals and of cultivated plants would offer the best chance of making out this obscure problem. Nor have I been disappointed; in this and in all other perplexing cases I have invariably found that our knowledge, imperfect though it be, of variation under domestication, afforded the best and safest clue*.”

Moreover, the identification of the genes underlying the phenotypic evolution associated with plant domestication is becoming of great economic importance, since it may facilitate trait manipulation through precise breeding strategies.

This thesis reports the characterization of the relationship between cultivated grapevine (*V.*subsp. *sativa*) and its supposed wild ancestor (*V.* subsp. *sylvestris*) at both genomic and phenotypic levels. The study has been organized in three main milestones, that is (i) the genotyping of a germplasm collection including wild and cultivated grapevines by using the latest Vitis20K SNP array and through the development a novel protocol of RAD-seq; (ii) the genome scan for signatures of selection with population genetic methods; (iii) the use of GWAS approach to identify the genetic bases of domestication-related traits in grapevine. The main conclusions drawn from these experiments are:

- both strategies of genotyping have presented some drawbacks. The array-based technology produced an excess of low frequency alleles, which may represent an underestimation of the real genetic diversity within the investigated population [161]. On the other hand, a high rate of missing data was observed in the SNP panel produced by RAD-seq. This result can be ascribed to the high level of heterozygosity of the grapevine genome [89], which is known to limit the performance of RRL technologies in discovering and genotyping genome-wide polymorphisms [192]. Nevertheless, we gained genetic profiles at 26K SNPs in almost one hundred grapevine individuals, half of which were *V. sylvestris*. This big amount of genetic information for such numerous individuals has not been obtained in grapevine so far, even though the full genome sequences of a few other individual grapevine cultivars have been published [95, 94].

- A significant variation in allele frequencies between wild and cultivated *V. vinifera* has been discovered at genomic regions including genes with roles in the adaptation to environmental stimuli. Indeed, the application of both population genetics and GWAS approaches led to the identification of genes encoding the ERF2, RAP2 and HRE1 transcription factors, chitinases,

- the CPN10 and the Na^+/H^+ antiporter of the SOS signaling pathway, which are involved in the response to salinity stess, high temperature, drought and pathogen attack.

- Most of the genomic regions identified as putative signatures of adaptation to domestication showed less genetic diversity in the wild compartment compared to grapevine cultivars. These findings raised some questions: is the genetic reduction in wild grapevine related to a higher stress tolerance? If yes, which physiological mechanisms are responsible for these abilities of adapting to environmental changes? If a balancing selection is acting, as suggested by the Tajima's D test, which is the evolutionary advantage of keeping both alleles at intermediate frequencies?

- Our findings on the genetic basis of domestication-related traits in grapevine support the prediction that changes in developmentally and morphologically complex traits, including single berry and single bunch weight in grapevine, occurred through selection on transcriptional regulators [i.e. MYB-H1-like and bZIP22 genes] as well as on proteins involved in hormone-dependent processes [i.e. NTF2 gene], and cell division [i.e. TPX2 gene] [328].

- The application of both top-down and bottom-up strategies to dissect the genomic basis of the phenotypic differentiation between wild and cultivated grapevine allowed to overcome some limitiations that each strategy presents individually. Indeed, when selection acts on standing genetic variation instead of a newly arisen mutation, undetectable "soft selective sweep" are generated by domestication, reducing the power of bottom-up approaches to detect signatures of selection [329]. In such cases, GWAS in populations of wild and cultivated plants is a suitable alternative to identify domestication genes. On the other hand, if the casual variants underlying domestication traits arise from de novo mutations as well as the trait is highly correlated with population structure, population genetic analysis is strongly recommended rather than association mapping.

Taken together, a step forwards to the acquisition of much more genetic information among thousands of grapevine individuals has been done in the present research. Moving from a single reference genome to multiple reference genomes is fundamental in grapevine in order to reconstruct its evolutionary history and for better interpreting the phenotypic variation observed nowadays in natural populations [330]. Our results point the attention towards wild grapevines as a model for understanding the mechanisms of adaptation to natural conditions. Future functional genomics studies accompanied by a broad phenotypic screening of stress tolerance in *V. sylvestris* are necessary to clarify how wild and cultivated grapevine react to environmental stimuli and stresses. In addition, the ongoing decline of wild grapevine populations encourages their preservation in germplasm collection, since they represent an opportunity for re-discovering resilience factors in view of a sustainable agriculture.

References

1. Adam-Blondon A-F, Martínez-Zapater JM, Kole C: *Genetics, Genomic and Breeding of Grapes*. 2012.

2. Iriti M, Rossoni M, Faoro F: **Melatonin content in grape: Myth or panacea?** *J Sci Food Agric* 2006, **86**:1432–1438.

3. OIV: **Statistical Report on World Viniviticulture**. 2015:1–15.

4. OIV: **OIV report on the world vitivinicultural situation**. 2015:3.

5. Nie Z-L, Sun H, Manchester SR, Meng Y, Luke Q, Wen J: **Evolution of the intercontinental disjunctions in six continents in the Ampelopsis clade of the grape family (Vitaceae)**. *BMC Evol Biol* 2012, **12**:17.

6. Kubitzki K: *The Families and Genera of Vascular Plants. Volume 53*; 2007.

7. Jansen RK, Kaittanis C, Saski C, Lee S-B, Tomkins J, Alverson AJ, Daniell H: **Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids.** *BMC Evol Biol* 2006, **6**:32.

8. Rossetto M, Jackes BR, Scott KD, Henry RJ: **Intergeneric relationships in the Australian Vitaceae: New evidence from cpDNA analysis**. *Genet Resour Crop Evol* 2001, **48**:307–314.

9. Ingrouille MJ, Chase MW, Fay MF, Bowman D, Van Der Bank M, Bruijn ADE: **Systematics of Vitaceae from the viewpoint of plastid rbcL DNA sequence data**. *Bot J Linn Soc* 2002, **138**:421–432.

10. Soejima A, Wen J: **Phylogenetic analysis of the grape family (Vitaceae) based on three chloroplast markers**. *Am J Bot* 2006, **93**:278–287.

11. Wen J, Nie Z-L, Soejima A, Meng Y: **Phylogeny of Vitaceae based on the nuclear GAI1 gene sequences**. *Can J Bot* 2007, **85**:731–745.

12. Rossetto M, Jackes BR, Scott KDS, Henry RJ, Michx A: **Is the Genus Cissus (Vitaceae) Monophyletic ? Evidence from Plastid and Nuclear Ribosomal DNA**. *Syst Bot* 2002, **27**:522–533.

13. Ren H, Lu LM, Soejima A, Luke Q, Zhang DX, Chen ZD, Wen J: **Phylogenetic analysis of the grape family (Vitaceae) based on the noncoding plastid trnC-petN, trnH-psbA, and trnL-F sequences**. *Taxon* 2011, **60**:629–637.

14. Wen J, Xiong Z, Nie ZL, Mao L, Zhu Y, Kan XZ, Ickert-Bond SM, Gerrath J, Zimmer EA, Fang XD: **Transcriptome Sequences Resolve Deep Relationships of the Grape Family**. *PLoS One* 2013, **8**.

15. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, et al.: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449**:463–7.

16. Zhang N, Wen J, Zimmer EA: **Congruent deep relationships in the grape family (vitaceae) based on sequences of chloroplast genomes and mitochondrial genes via genome skimming**. *PLoS One* 2015, **10**:1–12.

17. Wan Y, Schwaninger HR, Baldo AM, Labate JA, Zhong G-Y, Simon CJ: **A phylogenetic analysis of the grape genus (Vitis L.) reveals broad reticulation and concurrent diversification during neogene and quaternary climate change.** *BMC Evol Biol* 2013, **13**:141.

18. Reisch B, Pratt C: **Grapes**. In *Fruit breeding. 2nd edition*; 1996:297–369.

19. Barret H, Cramer S, Rhodes A: **A taximetric study of interspecific variation in Vitis.** *Vitis* 1969:177–187.

20. Tröndle D, Schröder S, Kassemeyer HH, Kiefer C, Koch MA, Nick P: **Molecular phylogeny of the genus Vitis (Vitaceae) based on plastid markers**. *Am J Bot* 2010, **97**:1168–1178.

21. Myles S, Chia JM, Hurwitz B, Simon C, Zhong GY, Buckler E, Ware D: **Rapid genomic characterization of the genus Vitis**. *PLoS One* 2010, **5(1): e821**.

22. Aradhya M, Wang Y, Walker MA, Prins BH, Koehmstedt AM, Velasco D, Gerrath JM, Dangl GS, Preece JE: **Genetic diversity, structure, and patterns of differentiation in the genus Vitis**. *Plant Syst Evol* 2013, **299**:317–330.

23. Bowen GJ, Beerling DJ, Koch PL, Zachos JC, Quattlebaum T: **A humid climate state during the Palaeocene / Eocene thermal maximum**. *Nature* 2004, **432**(November):495–499.

24. Hewitt G: **The genetic legacy of the Quaternary ice ages.** *Nature* 2000, **405**:907–913.

25. This P, Lacombe T, Thomas MR: **Historical origins and genetic diversity of wine grapes**. *Trends Genet* 2006, **22**:511–519.

26. Fleming SJ, Katz SH: *The Origins and Ancient History of Wine*. 1996.

27. Griffith MP: **Ancient Wine: The Search for the Origins of Viniculture**. *Econ Bot* 2004, **58**:488–488.

28. Arroyo-García R, Ruiz-García L, Bolling L, Ocete R, López MA, Arnold C, Ergul A, Söylemezoğlu G, Uzun HI, Cabello F, Ibáñez J, Aradhya MK, Atanassov A, Atanassov I, Balint S, Cenis JL, Costantini L, Gorislavets S, Grando MS, Klein BY, McGovern PE, Merdinoglu D, Pejic I, Pelsy F, Primikirios N, Risovannaya V, Roubelakis-Angelakis KA, Snoussi H, Sotiri P, Tamhankar S, et al.: **Multiple origins of cultivated grapevine (Vitis vinifera L. ssp. sativa) based on chloroplast DNA polymorphisms**. *Mol Ecol* 2006, **15**:3707–3714.

29. Olmo H: **Evolution of Crop Plants (ed. Simmonds NW).** 1976:294–298.

30. Mullins MG, Bouquet A, Williams LE: *Biology of the Grapevine*. 1992.

31. Vavilov NI: **Dikie rodichi plodovykh dereviev Aziatskoi chasti SSSR i Kavkaza i problema proiskhozhdenia plodovykh dereviev (Wild progenitors of the fruit trees of Turkestan and the Caucasus and the problem of the origin of fruit trees).** *Bull Appl Bot Genet Plant Breed* 1931, **6**:85–134.

32. Vouillamoz J, McGovern PE, Ergul A, Soylemezoglu G, Tevzadze G, Meredith CP, Grando MS: **Genetic characterization and relationships of traditional grape cultivars from Transcaucasia and Anatolia.** .

33. Grassi F, Labra M, Imazio S, Spada A, Sgorbati S, Scienza A, Sala F: **Evidence of a secondary grapevine domestication centre detected by SSR analysis**. *Theor Appl Genet* 2003, **107**:1315–1320.

34. Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D, Bustamante CD, Buckler ES: **Genetic structure and domestication history of the grape.** *Proc Natl Acad Sci U S A* 2011, **108**:3530–3535.

35. Drori E, Rahmi O, Henig Y, Lorenzi S, Brauner H, Marrano A, Amar Z, Netzer Y, Failla O, Grando MS: **Ampelographic and genetic characterization of the Israeli grapevine germplasm collection**. *Vitis* 2015, **54**(RN_32 ACCEPTED):107–110.

36. Drori E, Marrano A, Rahimia O, Heniga Y, Braunera H, Salmon-Divone M, Prazzoli, Maria Lucia Karpujf MV, Stanevskya M, Levia D, Feingerschf R, Failla O, Havivf I, Ehud W, Grando MS: **Genetic and phenotypic analysis of local populations points towards autonomous domestication of Israeli grapevine (Vitis vinifera)**. *Submitted* 2016.

37. Töpfer R, Sudharma KN, Kecke S, Marx G, Eibach R, Maghradze D, Maul E: **The Vitis International Varietey Catalogue (VIVC) – New design and more information**. In *XXXI Congreso Mundial de la Viña y el Vino*; 2008:9.

38. Emanuelli F, Battilana J, Costantini L, Le Cunff L, Boursiquot J-M, This P, Grando MS: **A candidate gene association study on muscat flavor in grapevine (Vitis vinifera L.).** *BMC Plant Biol* 2010, **10**:241.

39. Pelsy F, Merdinoglu D: **Complete sequence of Tvv1, a family of Ty1 copia-like retrotransposons of Vitis vinifera L., reconstituted by chromosome walking**. *Theor Appl Genet* 2002, **105**:614–621.

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

41. Fernandez L, Torregrosa L, Segura V, Bouquet A, Martinez-Zapater JM: **Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine**. *Plant J* 2010, **61**:545–557.

42. Doligez A, Bertrand Y, Farnos M, Grolier M, Romieu C, Esnault F, Dias S, Berger G, François P, Pons T, Ortigosa P, Roux C, Houel C, Laucou V, Bacilieri R, Péros J-P, This P: **New stable QTLs for berry weight do not colocalize with QTLs for seed traits in cultivated grapevine (Vitis vinifera L.).** *BMC Plant Biol* 2013, **13**:217.

43. Chen J, Wang N, Fang L-C, Liang Z-C, Li S-H, Wu B-H: **Construction of a high-density genetic map and QTLs mapping for sugars and acids in grape berries**. *BMC Plant Biol* 2015, **15**:1–14.

44. Costantini L, Malacarne G, Lorenzi S, Troggio M, Mattivi F, Moser C, Grando MS: **New candidate genes for the fine regulation of the colour of grapes**. *J Exp Bot* 2015, **66**:4427–4440.

45. Chitwood DH, Ranjan A, Martinez CC, Headland LR, Thiem T, Kumar R, Covington MF, Hatcher T, Naylor DT, Zimmerman S, Downs N, Raymundo N, Buckler ES, Maloof JN, Aradhya M, Prins B, Li L, Myles S, Sinha NR: **A Modern Ampelography: A Genetic Basis for Leaf Shape and Venation Patterning in Grape**. *Plant Physiol* 2014, **164**:259–272.

46. Grzeskowiak L, Costantini L, Lorenzi S, Grando MS: **Candidate loci for phenology and fruitfulness contributing to the phenotypic variability observed in grapevine**. *Theor Appl Genet* 2013, **126**:2763–2776.

47. Negrul A: **Origin and classification of cultivated grape**. In *The Ampelography of the USSR. vol 1*; 1946:159–216.

48. Dalmasso G, Tyndalo V: **Viticoltura e ampelografia dell'U.R.S.S.** *Accad Ital della Vite e del Vino* 1957, **Vol. IX**.

49. Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troggio M, Myles S, Martinez-Zapater JM, Zyprian E, Moreira FM, Grando MS: **Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape.** *BMC Plant Biol* 2013, **13**:39.

50. Bacilieri R, Lacombe T, Le Cunff L, Di Vecchi-Staraz M, Laucou V, Genna B, Péros J-P, This P, Boursiquot J-M: **Genetic structure in cultivated grapevines is linked to geography and human selection.** *BMC Plant Biol* 2013, **13**:25.

51. McClure KA, Sawler J, Gardner KM, Money D, Myles S: **Genomics: A potential panacea for the perennial problem**. *Am J Bot* 2014, **101**:1780–1790.

52. Marrano A, Grzeskowiak L, Sanz-Moreno P, Lorenzi S, Prazzoli ML, Arzumanov A, Amanova M, Failla O, Maghradze D, Grando MS: **Genetic diversity and relationships in the grapevine germplasm collection from Central Asia**. 2015, **54**:233–237.

53. Basheer-Salimia R, Lorenzi S, Batarseh F, Moreno-Sanz P, Emanuelli F, Grando MS: **Molecular identification and genetic relationships of Palestinian grapevine cultivars**. *Mol Biotechnol* 2014, **56**:546–556.

54. De Lorenzis G, Chipashvili R, Failla O, Maghradze D: **Study of genetic variability in Vitis vinifera L. germplasm by high-throughput Vitis18kSNP array: the case of Georgian genetic resources.** *BMC Plant Biol* 2015, **15**:154.

55. Santana JC, Heuertz M, Arranz C, Rubio JA, Martinez-Zapater JM, Hidalgo E: **Genetic structure, origins, and relationships of grapevine cultivars from the castilian plateau of spain**. *Am J Enol Vitic* 2010, **61**:214–224.

56. El Oualkadi A, Ater M, Laucou V, Boursiquot JM, Lacombe T, Peros JP, This P: **Study of genetic relationships between wild and domesticated grapevine in the north of Morocco**. *Int J Biodivers Conserv* 2011, **3**:512–526.

57. Laucou V, Lacombe T, Dechesne F, Siret R, Bruno JP, Dessup M, Dessup T, Ortigosa P, Parra P, Roux C, Santoni S, Varès D, Péros JP, Boursiquot JM, This P: **High throughput analysis of grape genetic diversity as a tool for germplasm collection management**. *Theor Appl Genet* 2011, **122**:1233–1245.

58. Picó FX, Méndez-Vigo B, Martínez-Zapater JM, Alonso-Blanco C: **Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the Iberian peninsula.** *Genetics* 2008, **180**:1009–1021.

59. Gao LZ, Zhang CH, Li DY, Pan DJ, Jia JZ, Dong YS: **Genetic diversity within Oryza rufipogon germplasms preserved in Chinese field gene banks of wild rice as revealed by microsatellite markers**. *Biodivers Conserv* 2006, **15**:4059–4077.

60. Vigouroux Y, Mitchell S, Matsuoka Y, Hamblin M, Kresovich S, Smith JSC, Jaqueth J, Smith OS, Doebley J: **An analysis of genetic diversity across the maize genome using microsatellites**. *Genetics* 2005, **169**:1617–1630.

61. Cabezas J a, Ibáñez J, Lijavetzky D, Vélez D, Bravo G, Rodríguez V, Carreño I, Jermakow AM, Carreño J, Ruiz-García L, Thomas MR, Martinez-Zapater JM: **A 48 SNP set for grapevine cultivar identification**. *BMC Plant Biol* 2011, **11**:153.

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

63. Ching A, Caldwell K., Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski A.: **SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines**. *BMC Genet* 2002, **3**:19.

64. Heywood V, Zohary D: **A catalogue of wild relatives of cultivated plants native to Europe**. *Flora Mediterr* 1991, **5**:375–415.

65. This P et al.: **Caractérisation de la diversité d'une population de vignes sauvages du Pic Saint-Loup (Hérault) et relations avec le compartiment cultive**. *Genet Selec Evol* 2001, **33**:289–304.

66. De Andrés MT, Benito A, Pérez-Rivera G, Ocete R, Lopez MA, Gaforio L, Munoz G, Cabello F, Martìnez Zapater JM, Arroyo-Garcìa R: **Genetic diversity of wild grapevine populations in Spain and their genetic relationships with cultivated grapevines**. *Mol Ecol* 2012, **21**:800–816.

67. Grando MS, De Micheli L, Biasetto L, Scienza A: **RAPD markers in wild and cultivated Vitis vinifera**. *Vitis* 1995, **34**:37–39.

68. Levadoux L: **Les populations sauvages et cultivées de Vitis vinifera L**. *Ann l'amélioration des plantes* 1956:59–118.

69. Arroyo-García R, Revilla E: **The Current Status of Wild Grapevine Populations (Vitis vinifera ssp sylvestris) in the Mediterranean Basin**. *Mediterr Genet Code - Grapevine Olive* 2013:51–72.

70. Di Vecchi-Staraz M, Laucou V, Bruno G, Lacombe T, Gerber S, Bourse T, Boselli M, This P: **Low level of pollen-mediated gene flow from cultivated to wild grapevine: Consequences for the evolution of the endangered subspecies Vitis vinifera L. subsp. silvestris**. *J Hered* 2009, **100**:66–75.

71. Ocete R, Cantos M, López G, Pérez A, A. T, Lara M, Failla O, Ferragut FJ, Liñán J: *Caracterización Y Conservación Del Recurso Fitogenético Vid Silvestre En Andalucía*. 2007.

72. Ergül A, Perez-Rivera G, Söylemezoğlu G, Kazan K, Arroyo-Garcia R: **Genetic diversity in Anatolian wild grapes (Vitis vinifera subsp. sylvestris) estimated by SSR markers**. *Plant Genet Resour* 2011, **9**:375–383.

73. Pipia I, Gamkrelidze M, Gogniashvli M, Tabidze V: **Genetic diversity of Georgian varieties of Vitis vinifera subsp. sylvestris**. *XXXIII Congr Mund la Viña y el Vino* 2010:5.

74. Revilla E, Carrasco D, Benito A, Arroyo-García R: **Anthocyanin composition of several wild grape accessions**. *Am J Enol Vitic* 2010, **61**:536–543.

75. Revilla E, Carrasco D, Carrasco V, Benito A, Arroyo-García R: **On the absence of acylated anthocyanins in some wild grapevine accessions**. *Vitis - J Grapevine Res* 2012, **51**:161–165.

76. Mattivi F, Guzzon R, Vrhovsek U, Stefanini M, Velasco R: **Metabolite profiling of grape: Flavonols and anthocyanins**. *J Agric Food Chem* 2006, **54**:7692–7702.

77. Bodor P, Ladányi M, Grzeskowiak L, Grando MS, Bisztray GD: **Ampelometric evaluation of wild grape (Vitis vinifera L. ssp. sylvestris (C.C. Gmel.) Hegi) accessions in the germplasm collection of FEM-IASMA, Italy**. 2015, **54**:213–215.

78. Guan X, Essakhi S, Laloue H, Nick P, Bertsch C, Chong J: **Mining new resources for grape resistance against Botryosphaeriaceae: a focus on Vitis vinifera subsp. sylvestris**. *Plant Pathol* 2015:n/a–n/a.

79. Riaz S, Boursiquot J-M, Dangl GS, Lacombe T, Laucou V, Tenscher AC, Walker MA: **Identification of mildew resistance in wild and cultivated Central Asian grape germplasm.** *BMC Plant Biol* 2013, **13**:149.

80. Coleman C, Copetti D, Cipriani G, Hoffmann S, Kozma P, Kovacs L, Morgante M, Testolin R, Di GG: **The powdery mildew resistance gene REN1 co-segregates with an NBS-LRR gene cluster in two Central Asian grapevines**. *BMC Genet* 2009, **10**(1471-2156 (Electronic)):89.

81. Riaz S, Tenscher AC, Ramming DW, Walker MA: **Using a limited mapping strategy to identify major QTLs for resistance to grapevine powdery mildew (Erysiphe necator) and their use in marker-assisted breeding**. *Theor Appl Genet* 2011, **122**:1059–1073.

82. Tisch C, Nick P, Kortekamp A: **Rescue to be rescued: European wild grape as genetic resources of resistance towards fungal diseases**. In *Proceedings of the 7th International Workshop on Gapevine Downy and Powdery Mildew*; 2014:61–62.

83. Duan D, Halter D, Baltenweck R, Tisch C, Tröster V, Kortekamp A, Hugueney P, Nick P: **Genetic diversity of stilbene metabolism in Vitis sylvestris**. *J Exp Bot* 2015, **66**:3243–3257.

84. Ocete R, Arroyo-Garcia R, Morales ML, Cantos M, Gallardo A, P??rez MA, G??mez I, L??pez MA: **Characterization of Vitis vinifera L. subspecies sylvestris (Gmelin) Hegi in the Ebro river Basin (Spain)**. *Vitis - J Grapevine Res* 2011, **50**:11–16.

85. Ellstrand NC, Heredia SM, Leak-Garcia JA, Heraty JM, Burger JC, Yao L, Nohzadeh-Malakshah S, Ridley CE: **Crops gone wild: Evolution of weeds and invasives from domesticated ancestors**. *Evol Appl* 2010, **3**:494–504.

86. Earl DA, vonHoldt BM: **STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method**. *Conserv Genet Resour* 2012, **4**:359–361.

87. Regner F, Stadlbauer A, Eisenheld C, Kaserer Herw: **Genetic relationships among Pinots and Related Cultivars.** 2000, **51**:7–14.

88. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome**. *Nature* 2005, **436**:793–800.

89. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Dematté L, Mraz A, et al.: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety**. *PLoS One* 2007, **2(12): e13**.

90. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A: **Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize.** *Nat Genet* 2005, **37**:997–1002.

91. Tuskan G, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Others: **The genome of black cottonwood, Populus trichocarpa (Torr. \& Gray)**. *Science (80-)*

2006, **313**:1596.

92. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unrevealing angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events**. *Nature* 2003, **422**(March):433–438.

93. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van De Peer Y: **Modeling gene and genome duplications in eukaryotes**. *Proc Natl Acad Sci U S A* 2005, **102**:5454–5459.

94. Da Silva C, Zamperin G, Ferrarini A, Minio A, Dal Molin A, Venturini L, Buson G, Tononi P, Avanzato C, Zago E, Boido E, Dellacassa E, Gaggero C, Pezzotti M, Carrau F, Delledonne M: **The High Polyphenol Content of Grapevine Cultivar Tannat Berries Is Conferred Primarily by Genes That Are Not Shared with the Reference Genome**. *Plant Cell* 2013, **25**:4777–4788.

95. Di Genova A, Almeida AM, Muñoz-Espinoza C, Vizoso P, Travisany D, Moraga C, Pinto M, Hinrichsen P, Orellana A, Maass A: **Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants.** *BMC Plant Biol* 2014, **14**:7.

96. Corso M, Vannozzi A, Maza E, Vitulo N, Meggio F, Pitacco A, Telatin A, D'Angelo M, Feltrin E, Negri AS, Prinsi B, Valle G, Ramina A, Bouzayen M, Bonghi C, Lucchin M: **Comprehensive transcript profiling of two grapevine rootstock genotypes contrasting in drought susceptibility links the phenylpropanoid pathway to enhanced tolerance**. *J Exp Bot* 2015, **66**:5739–5752.

97. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, et al.: **Real-time DNA sequencing from single polymerase molecules.** *Science* 2009, **323**:133–8.

98. Rhoads A, Au KF: **PacBio Sequencing and Its Applications**. *Genomics, Proteomics Bioinforma* 2015, **13**:278–289.

99. Cantu D: **Genomics of the grapevine and its associated microorganisms enables systems biology studies of grape-microbe interactions in the vineyard.** In *X International Symposiom on Grapevine Phisiology and Biotechnology*; 2016.

100. Vitulo N, Forcato C, Carpinelli EC, Telatin A, Campagna D, D'Angelo M, Zimbello R, Corso M, Vannozzi A, Bonghi C, Lucchin M, Valle G: **A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype.** *BMC Plant Biol* 2014, **14**:99.

101. Forcato C: **Gene prediction and functional annotation in the Vitis vinifera genome**. *PhD Thesis Univ Degli Stud Di Padova* 2010.

102. Petit RJ, Hampe A: **Some Evolutionary Consequences of Being a Tree**. *Annu Rev Ecol Evol Syst* 2006, **37**:187–214.

103. Martìnez-Zapater JM, Carmona MJ, Dìaz-Riquelme J, Fernàndez L, Lijavetzky D: **Grapevine genetics after the genome sequence: Challenges and limitations**. *Aust J Grape Wine Res* 2010, **16**:33–46.

104. Boss PK, Thomas MR: **Association of dwarfism and floral induction with a grape "green revolution" mutation.** *Nature* 2002, **416**:847–850.

105. Chaïb J, Torregrosa L, MacKenzie D, Corena P, Bouquet A, Thomas MR: **The grape microvine - A model system for rapid forward and reverse genetics of grapevines**. *Plant J* 2010, **62**:1083–1092.

106. Hamilton MB: *Population Genetics*. 2009.

107. Flint-Garcia S a, Thornsberry JM, Buckler ES: **Structure of linkage disequilibrium in plants.** *Annu Rev Plant Biol* 2003, **54**:357–374.

108. Lewontin RC: **The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.** *Genetics* 1964, **49**:49–67.

109. Hill WG, Robertson A: **Linkage disequilibrium in finite populations**. *Theor Appl Genet* 1968, **38**:226–

231.

110. Jorde LB: **Linkage Disequilibrium and the Search for Complex Disease Genes**. *Genome Res* 2000, **10**:1435–1444.

111. Slatkin M: **Linkage disequilibrium--understanding the evolutionary past and mapping the medical future.** *Nat Rev Genet* 2008, **9**:477–85.

112. Hamblin MT, Buckler ES, Jannink JL: **Population genetics of genomics-based crop improvement methods**. *Trends Genet* 2011, **27**:98–106.

113. Nicolas SD, Péros J-P, Lacombe T, Launay A, Le Paslier M-C, Bérard A, Mangin B, Valière S, Martins F, Le Cunff L, Laucou V, Bacilieri R, Dereeper A, Chatelet P, This P, Doligez A: **Genetic diversity, linkage disequilibrium and power of a large grapevine (Vitis vinifera L) diversity panel newly designed for association studies**. *BMC Plant Biol* 2016, **16**:74.

114. Vitti JJ, Grossman SR, Sabeti PC: **Detecting Natural Selection in Genomic Data**. *Annu Rev Genet* 2013, **47**:97–120.

115. Costantini L, Moreira FM, Zyprian E, Martínez-Zapater JM, Grando MS: **Molecular maps, QTL mapping and association mapping in grapevine. In: Molecular biology and biotechnology of the grapevine.** In *Molecular biology and biotechnology of the grapevine.*; 2009:535–564.

116. Duchêne E, Butterlin G, Claudel P, Dumas V, Jaegli N, Merdinoglu D: **A grapevine (Vitis vinifera L.) deoxy-d-xylulose synthase gene colocates with a major quantitative trait loci for terpenol content**. *Theor Appl Genet* 2009, **118**:541–552.

117. Pauquet J, Bouquet a., This P, Adam-Blondon a.-F: **Establishment of a local map of AFLP markers around the powdery mildew resistance gene Run 1 in grapevine and assessment of their usefulness for marker assisted selection**. *TAG Theor Appl Genet* 2001, **103**:1201–1210.

118. Merdinoglu D, Wiedemann-Merdinoglu S, Coste P, Dumas V, Haetty S, Butterlin G, Greif C, Adam-Blondon AF, Bouquet A, Pauquet J: **Genetic analysis of downy mildew resistance derived from Muscadinia rotundifolia**. In *Acta Horticulturae*. *Volume 603*; 2003:451–456.

119. Barba P, Cadle-Davidson L, Harriman J, Glaubitz JC, Brooks S, Hyma K, Reisch B: **Grapevine powdery mildew resistance and susceptibility loci identified on a high-resolution SNP map**. *Theor Appl Genet* 2014, **127**:73–84.

120. Dalbó M a, Ye GN, Weeden NF, Steinkellner H, Sefc KM, Reisch BI: **A gene controlling sex in grapevines placed on a molecular marker-based genetic map.** *Genome* 2000, **43**:333–340.

121. Fournier-Level a, Lacombe T, Le Cunff L, Boursiquot J-M, This P: **Evolution of the VvMybA gene family, the major determinant of berry colour in cultivated grapevine (Vitis vinifera L.).** *Heredity (Edinb)* 2010, **104**:351–362.

122. Doligez A, Bouquet A, Danglot Y, Lahogue F, Riaz S, Meredith CP, Edwards KJ, This P: **Genetic mapping of grapevine (Vitis vinifera L.) applied to the detection of QTLs for seedlessness and berry weight**. *Theor Appl Genet* 2002, **105**:780–795.

123. Costantini L, Battilana J, Lamaj F, Fanizza G, Grando MS: **Berry and phenology-related traits in grapevine (Vitis vinifera L.): from quantitative trait loci to underlying genes.** *BMC Plant Biol* 2008, **8**:38.

124. Houel C, Chatbanyong R, Doligez A, Rienth M, Foria S, Luchaire N, Roux C, Adivèze A, Lopez G, Farnos M, Pellegrino A, This P, Romieu C, Torregrosa L: **Identification of stable QTLs for vegetative and reproductive traits in the microvine (Vitis vinifera L.) using the 18 K Infinium chip.** *BMC Plant Biol* 2015, **15**:205.

125. Xu S: **Theoretical Basis of the Beavis Effect**. *Genetics* 2003, **165**:2259–2268.

126. Ross-Ibarra J, Morrell PL, Gaut BS: **Plant domestication, a unique opportunity to identify the genetic basis of adaptation.** *Proc Natl Acad Sci U S A* 2007, **104 Suppl**:8641–8.

127. Abdurakhmonov IY, Abdukarimov A: **Application of association mapping to understanding the genetic diversity of plant germplasm resources**. *Int J Plant Genomics* 2008, **2008**.

128. Kraakman ATW, Niks RE, Van Den Berg PMMM, Stam P, Van Eeuwijk FA: **Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars**. *Genetics* 2004, **168**:435–446.

129. Ingvarsson PK, Street NR: **Association genetics of complex traits in plants**. *New Phytol* 2011, **189**:909–922.

130. Ehrenreich IM, Hanzawa Y, Chou L, Roe JL, Kover PX, Purugganan MD: **Candidate gene association mapping of Arabidopsis flowering time**. *Genetics* 2009, **183**:325–335.

131. Wen Z, Tan R, Yuan J, Bales C, Du W, Zhang S, Chilvers MI, Schmidt C, Song Q, Cregan PB, Wang D: **Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean.** *BMC Genomics* 2014, **15**:809.

132. Zhang Z, Ersoz E, Lai C, Todhunter RJ, Tiwari HK, Gore M a, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES: **Mixed linear model approach adapted for gneome-wide association studies**. *Nat Genet* 2010, **42**:355–360.

133. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nat Genet* 2006, **38**:203–8.

134. Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, Nordborg M: **Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes.** *PLoS Genet* 2005, **1**.

135. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S: **Population genomic and genome-wide association studies of agroclimatic traits in sorghum.** *Proc Natl Acad Sci U S A* 2013, **110**:453–8.

136. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li WW, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li WW, Lin Z, Buckler ES, Qian Q, Zhang Q-F, Li J, Han B: **Genome-wide association studies of 14 agronomic traits in rice landraces.** *Nat Genet* 2010, **42**:961–967.

137. Ma Y, Evans D, Logue S, Langridge P: **Mutations of barley B-amylase that improve substrate-binding affinity and thermostability**. *Mol Genet Genomics* 2001, **266**:345–352.

138. Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE: **Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize.** *Proc Natl Acad Sci U S A* 2012, **109**:8872–7.

139. Kumar S, Garrick DJ, Bink MC, Whitworth C, Chagné D, Volz RK: **Novel genomic approaches unravel genetic architecture of complex traits in apple.** *BMC Genomics* 2013, **14**:393.

140. Sardos J, Rouard M, Hueber Y, Cenci A, Hyma KE, van den Houwe I, Hribova E, Courtois B, Roux N: **A Genome-Wide Association Study on the Seedless Phenotype in Banana (Musa spp.) Reveals the Potential of a Selected Panel to Detect Candidate Genes in a Vegetatively Propagated Crop.** *PLoS One* 2016, **11**:e0154448.

141. Maynard Smith J, Haigh J: **The hitch-hiking effect of a favorable gene**. *Genet Res* 1974, **23**:23–35.

142. Olsen KM, Wendel JF: **A bountiful harvest: genomic insights into crop domestication phenotypes.** *Annu Rev Plant Biol* 2013, **64**:47–70.

143. Delph LF, Kelly JK: **On the importance of balancing selection in plants**. *New Phytol* 2014, **201**:45–56.

144. Gepts P, Papa R: **Evolution during Domestication**. *Encycl Life Sci* 2002:1–7.

145. Huang X, Han B: **Natural variations and genome-wide association studies in crop plants.** *Annu Rev Plant Biol* 2014, **65**(November 2013):531–51.

146. Cutter AD, Payseur BA: **Genomic signatures of selection at linked sites: unifying the disparity among species.** *Nat Rev Genet* 2013, **14**:262–74.

147. Wright S: **The Genetical Structure of Populations**. *Ann Eugen* 1949, **15**:323–354.

148. Nielsen R: **Molecular signatures of natural selection.** *Annu Rev Genet* 2005, **39**:197–218.

149. Nei M: **Analysis of gene diversity in subdivided populations**. *Proc Natl Acad Sci U S A* 1973, **70**:3321–3323.

150. Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Stadler T, Li J, Ye Z, et al.: **Genomic analyses provide insights into the history of tomato breeding**. *Nat Genet* 2014, **46**:1220–1226.

151. Tajima F: **Evolutionary relationship of DNA sequences in finite populations**. *Genetics* 1983, **105**:437–460.

152. Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, Ben C, Denny R, Sadowsky MJ, Ronfort J, Bataillon T, Young ND, Tiffin P: **Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula.** *Proc Natl Acad Sci U S A* 2011, **108**:E864–70.

153. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, et al.: **A map of rice genome variation reveals the origin of cultivated rice**. *Nature* 2012, **490**:497–501.

154. Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J: **The Genomic Signature of Crop-Wild Introgression in Maize**. *PLoS Genet* 2013, **9**.

155. Leforestier D, Ravon E, Muranty H, Cornille A, Lemaire C, Giraud T, Durel CE, Branca A: **Genomic basis of the differences between cider and dessert apple varieties**. *Evol Appl* 2015, **8**:650–661.

156. Deschamps S, Llaca V, May GD: **Genotyping-by-Sequencing in Plants**. *Biology (Basel)* 2012, **1**:460–483.

157. Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S: **SNP markers and their impact on plant breeding**. *Int J Plant Genomics* 2012, **2012**.

158. Gupta PK, Rustgi S, Mir RR: **Array-based high-throughput DNA markers for crop improvement.** *Heredity (Edinb)* 2008, **101**:5–18.

159. Kumar S, Banks TW, Cloutier S: **SNP discovery through next-generation sequencing and its applications**. *Int J Plant Genomics* 2012, **2012**.

160. Le Paslier MC, Choisne N, Bacilieri R, Bounon R, Boursiquot J, Bras M, Brunel D, Di Gaspero G, Hausmann L, Lacombe T, Laucou V, Launay A, Martinez-Zapater JM, Morgante M, Raj P., Ponnaiah M, Quesneville H, Scalabrin S, Torres-Perez R, Adam-Blondon AF: **The GrapeReSeq 18 k Vitis genotyping chip.** In *In 9th International symposium grapevine physiology and biotechnology: International Society for Horticultural Science*; 2013:123.

161. Wang Y, Nielsen R: **Estimating population divergence time and phylogeny from single-nucleotide polymorphisms data with outgroup ascertainment bias**. *Mol Ecol* 2012, **21**:974–986.

162. Miller M, Dunham J, Amores A, Cresko W, Johnson E: **Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers**. *Genome Res* 2007, **17**:240–248.

163. Davey JL, Blaxter MW: **RADseq: Next-generation population genetics**. *Brief Funct Genomics* 2010,

9:416–423.

164. Nelson JC, Wang S, Wu Y, Li X, Antony G, White FF, Yu J: **Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum.** *BMC Genomics* 2011, **12**:352.

165. Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, Toppino L, Rotino G: **Identification of SNP and SSR markers in eggplant using RAD tag sequencing**. *BMC Genomics* 2011, **12**:304.

166. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species**. *PLoS One* 2011, **6**:1–10.

167. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE: **Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species**. *PLoS One* 2012, **7(5): e371**.

168. Wang S, Meyer E, McKay JK, Matz M V: **2b-RAD: a simple and flexible method for genome-wide genotyping**. *Nat Methods* 2012, **9**:808–810.

169. Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S: **LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms.** *G3 (Bethesda)* 2015, **5**:2383–90.

170. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet* 2007, **81**:1084–97.

171. Stölting KN, Nipper R, Lindtke D, Caseys C, Waeber S, Castiglione S, Lexer C: **Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species**. *Mol Ecol* 2013, **22**:842–855.

172. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: **Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags**. *PLoS Genet* 2010, **6**.

173. Thermo Fisher Scientific Inc.: **Fragment Library Preparation 5500**. 2015:1–16.

174. **Vitis vinifera mitochondrion, complete genome** [http://www.ncbi.nlm.nih.gov/nuccore/224365609/]

175. **Vitis vinifera chloroplast, complete genome.** [http://www.ncbi.nlm.nih.gov/nuccore/91983971/]

176. Homer N, Merriman B, Nelson SF: **BFAST: An alignment tool for large scale genome resequencing**. *PLoS One* 2009, **4(11): e77**.

177. Development Core Team R: *R: A Language and Environment for Statistical Computing*. *Volume 0*; 2011.

178. Wickham H: *ggplot2: Elegant Graphics for Data Analysis. Volume 3*. Springer-Verlag New York; 2009.

179. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**:1297–1303.

180. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**:2156–2158.

181. Sanger F, Nicklen S, Coulson a R: **DNA sequencing with chain-terminating inhibitors.** *Proc Natl Acad Sci U S A* 1977, **74**:5463–7.

182. **Primer-BLAST.** [http://www.ncbi.nlm.nih.gov/tools/primer-blast/]

183. Staden R: **The Staden sequence analysis package.** *Mol Biotechnol* 1996, **5**:233–241.

184. **Plant and Fungi Data Integration. GrapeReSeq_Illumina_20K** [Http://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K]

185. Illumina Inc.: **Infinium ® Genotyping Data Analysis**. 2014.

186. Prazzoli ML, Marrano A, Lorenzi S, Failla O, Grando MS: **Genetic investigation of Caucasian grapevine germplasm with low susceptibility to downy mildew**. In *X International Symposiom on Grapevine Phisiology and Biotechnology*; 2016.

187. Fodor A, Segura V, Denis M, Neuenschwander S, Fournier-Level A, Chatelet P, Homa FAA, Lacombe T, This P, Cunff L Le: **Genome-wide prediction methods in highly diverse and heterozygous species: Proof-of-concept through simulation in grapevine**. *PLoS One* 2014, **9(11): e11**.

188. Martinez-Zapater JM, Carmona MJ, Diaz-Riquelme J, Fernandez L, Lijavetzky D: **Grapevine genetics after the genome sequence: Challenges and limitations**. *Aust J Grape Wine Res* 2010, **16**:33–46.

189. Morgante M, De Paoli E, Radovic S: **Transposable elements and the plant pan-genomes**. *Curr Opin Plant Biol* 2007, **10**:149–155.

190. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Dematt?? L, Mraz A, et al.: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety**. *PLoS One* 2007, **2**.

191. Wang N, Fang L, Xin H, Wang L, Li S: **Construction of a high-density genetic map for grape using next generation restriction-site associated DNA sequencing**. *BMC Plant Biol* 2012, **12**:148.

192. Myles S: **Improving fruit and wine: What does genomics have to offer?** *Trends Genet* 2013, **29**:190–196.

193. Ilut DC, Nydam ML, Hare MP: **Defining loci in restriction-based reduced representation genomic data from nonmodel species: Sources of bias and diagnostics for optimal clustering**. *Biomed Res Int* 2014, **2014**.

194. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR: **Low-coverage sequencing: implications for design of complex trait association studies**. *Genome …* 2011, **21**:940–951.

195. Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML: **Special features of RAD Sequencing data: Implications for genotyping**. *Mol Ecol* 2013, **22**:3151–3164.

196. Grimplet J, Van Hemert J, Carbonell-Bejerano P, Díaz-Riquelme J, Dickerson J, Fennell A, Pezzotti M, Martínez-Zapater JM: **Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences**. *BMC Res Notes* 2012, **5**:213.

197. Mattia F, Imazio S, Grassi F, Doulati Baneh H, Scienza A, Labra M: **Study of genetic relationships between wild and domesticated grapevine distributed from Middle East Regions to European countries**. *Rend Lincei* 2008, **19**:223–240.

198. Castañeda-Álvarez NP, Khoury CK, Achicanoy HA, Bernau V, Dempewolf H, Eastwood RJ, Guarino L, Harker RH, Jarvis A, Maxted N, Müller J V., Ramirez-Villegas J, Sosa CC, Struik PC, Vincent H, Toll J: **Global conservation priorities for crop wild relatives**. *Nat Plants* 2016:1–6.

199. McGovern PE, Glusker DL, Exner LJ, Voigt MM: **Neolithic resinated wine**. *Nature* 1996, **381**:480–481.

200. Fechter I, Hausmann L, Daum M, Rosleff Sörensen T, Viehöver P, Weisshaar B, Töpfer R: **Candidate genes within a 143 kb region of the flower sex locus in Vitis**. *Mol Genet Genomics* 2012, **287**:247–259.

201. Askri H, Daldoul S, Ammar A Ben, Rejeb S, Jardak R, Rejeb MN, Mliki A, Ghorbel A: **Short-term response of wild grapevines (Vitis vinifera L. ssp. sylvestris) to NaCl salinity exposure: Changes of some physiological and molecular characteristics**. *Acta Physiol Plant* 2012, **34**:957–968.

202. Doulati Baneh H, Hassani A, Abdollahi R: **Growth and physiological responses of some wild grapevine (Vitis vinifera L . SSP . sylvestris) genotypes to salinity**. 2015, **21**:530–535.

203. Cambrollé J, Garcìa JL, Figueroa ME, Cantos M: **Physiological responses to soil lime in wild grapevine (Vitis vinifera ssp. sylvestris)**. *Environ Exp Bot* 2014, **105**:25–31.

204. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: **Genomic scans for selective sweeps using SNP data**. *Genome Res* 2005, **15**:1566–1575.

205. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner RA, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J: **Comparative population genomics of maize domestication and improvement**. *Nat Genet* 2012, **44**:808–811.

206. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, et al.: **A map of rice genome variation reveals the origin of cultivated rice**. *Nature* 2012, **490**:497–501.

207. Bonhomme M, Boitard S, San Clemente H, Dumas B, Young N, Jacquet C: **Genomic Signature of Selective Sweeps Illuminates Adaptation of Medicago truncatula to Root-Associated Microorganisms.** *Mol Biol Evol* 2015, **32**:2097–2110.

208. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: A tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**:559–575.

209. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets**. *Gigascience* 2015, **4**:7.

210. Raj A, Stephens M, Pritchard JK: **FastSTRUCTURE: Variational inference of population structure in large SNP data sets**. *Genetics* 2014, **197**:573–589.

211. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data**. *Genetics* 2000, **155**:945–959.

212. Jakobsson M, Rosenberg NA: **CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure**. *Bioinformatics* 2007, **23**:1801–1806.

213. Rosenberg NA: **DISTRUCT: A program for the graphical display of population structure**. *Mol Ecol Notes* 2004, **4**:137–138.

214. Jombart T: **Adegenet: A R package for the multivariate analysis of genetic markers**. *Bioinformatics* 2008, **24**:1403–1405.

215. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: Analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**:263–265.

216. Weir B, Cockerham CC: **Estimating F-Statistics for the Analysis of Population Structure Author (s): B . S . Weir and C . Clark Cockerham**. *Evolution (N Y)* 1984, **38**:1358–1370.

217. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism**. *Genetics* 1989, **123**:585–595.

218. Alexa A, Rahnenfuhrer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure**. *Bioinformatics* 2006, **22**:1600–1607.

219. Kole C: **Genetics, Genomics, and Breeding of Grapes**. 2011(Ld):390.

220. Barnaud a, Laucou V, This P, Lacombe T, Doligez a: **Linkage disequilibrium in wild French grapevine, Vitis vinifera L. subsp. silvestris.** *Heredity (Edinb)* 2010, **104**:431–437.

221. Koch E, Ristroph M, Kirkpatrick M: **Long range linkage disequilibrium across the human genome**. *PLoS One* 2013, **8**.

222. Battilana J, Lorenzi S, Moreira FM, Moreno-Sanz P, Failla O, Emanuelli F, Grando MS: **Linkage mapping and molecular diversity at the flower sex locus in wild and cultivated grapevine reveal a prominent ssr haplotype in hermaphrodite plants**. *Mol Biotechnol* 2013, **54**:1031–1037.

223. Riaz S, Krivanek AF, Xu K, Walker MA: **Refined mapping of the Pierce's disease resistance locus, PdR1, and Sex on an extended genetic map of Vitis rupestris x V. arizonica**. *Theor Appl Genet* 2006, **113**:1317–1329.

224. Lowe KM, Walker MA: **Genetic linkage map of the interspecific grape rootstock cross Ramsey (Vitis champinii) Riparia Gloire (Vitis riparia)**. *Theor Appl Genet* 2006, **112**:1582–1592.

225. Zinelabidine LH, Haddioui A, Bravo G, Arroyo-García R, Martínez Zapater JM: **Genetic origins of cultivated and wild grapevines from morocco**. *Am J Enol Vitic* 2010, **61**:83–90.

226. Aradhya MK, Dangl GS, Prins BH, Boursiquot J-M, Walker MA, Meredith CP, Simon CJ: **Genetic structure and differentiation in cultivated grape, Vitis vinifera L.** *Genet Res (Camb)* 2003, **81**:179–192.

227. Riahi L, Zoghlami N, Dereeper A, Laucou V, Mliki A, This P: **Single nucleotide polymorphism and haplotype diversity of the gene NAC4 in grapevine**. *Ind Crops Prod* 2013, **43**:718–724.

228. Houel C, Bounon R, Chaïb J, Guichard C, Péros J-P, Bacilieri R, Dereeper A, Canaguier A, Lacombe T, N'Diaye A, Le Paslier M-C, Vernerey M-S, Coriton O, Brunel D, This P, Torregrosa L, Adam-Blondon A-F: **Patterns of sequence polymorphism in the fleshless berry locus in cultivated and wild Vitis vinifera accessions.** *BMC Plant Biol* 2010, **10**:284.

229. Charlesworth D: **Balancing selection and its effects on sequences in nearby genome regions**. *PLoS Genet* 2006, **2**:379–384.

230. Van Minnebruggen A, Neyt P, Groeve S De, Coussens G, Ponce MR, Micol JL, Van Lijsebettens M: **The ang3 mutation identified the ribosomal protein gene RPL5B with a role in cell expansion during organ growth**. *Physiol Plant* 2010, **138**:91–101.

231. Devis D, Firth SM, Liang Z, Byrne ME: **Dosage Sensitivity of RPL9 and Concerted Evolution of Ribosomal Protein Genes in Plants**. *Front Plant Sci* 2015, **6**(December):1–12.

232. Bonan GB: **Ecological Climatology: Concepts and Applications, 2nd Edition**. *Geogr Res* 2008, **48**:221–222.

233. Agudelo-Romero P, Erban A, Sousa L, Pais MS, Kopka J, Fortes AM: **Search for Transcriptional and Metabolic Markers of Grape Pre-Ripening and Ripening and Insights into Specific Aroma Development in Three Portuguese Cultivars**. *PLoS One* 2013, **8**.

234. Deluc LG, Grimplet J, Wheatley MD, Tillett RL, Quilici DR, Osborne C, Schooley D a, Schlauch K a, Cushman JC, Cramer GR: **Transcriptomic and metabolite analyses of Cabernet Sauvignon grape berry development.** *BMC Genomics* 2007, **8**:429.

235. Leng P, Yuan B, Guo Y, Chen P: **The role of abscisic acid in fruit ripening and responses to abiotic stress**. *J Exp Bot* 2014, **65**:4577–4588.

236. Fortes A, Teixeira R, Agudelo-Romero P: **Complex Interplay of Hormonal Signals during Grape Berry Ripening**. *Molecules* 2015, **20**:9326–9343.

237. Skopelitis DS, Paranychianakis N V, Paschalidis KA, Pliakonis ED, Delis ID, Yakoumakis DI, Kouvarakis A, Papadakis AK, Stephanou EG, Roubelakis-Angelakis KA: **Abiotic stress generates ROS that signal expression of anionic glutamate dehydrogenases to form glutamate for proline synthesis in tobacco and grapevine**. *Plant Cell* 2006, **18**:2767–2781.

238. Yang T, Poovaiah BW: **Arabidopsis chloroplast chaperonin 10 is a calmodulin-binding protein.** *Biochem Biophys Res Commun* 2000, **275**:601–7.

239. Hemmingsen SM, Woolford C, van der Vies SM, Tilly K, Dennis DT, Georgopoulos CP, Hendrix RW, Ellis RJ: **Homologous plant and bacterial proteins chaperone oligomeric protein assembly**. *Nature* 1988, **333**:330–334.

240. Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette M-L, Mireau H, Peeters N, Renou J-P, Szurek B,

Taconnat L, Small I: **Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis.** *Plant Cell* 2004, **16**:2089–103.

241. Grennan AK: **To Thy Proteins Be True: RNA Editing in Plants**. *Plant Physiol* 2011, **156**:453–454.

242. Azevedo C, Santos-Rosa MJ, Shirasu K: **The U-box protein family in plants**. *Trends Plant Sci* 2001, **6**:354–358.

243. Callis J: **The ubiquitination machinery of the ubiquitin system.** *Arabidopsis Book* 2014, **12**:e0174.

244. Bonaventure G, Salas JJ, Pollard MR, Ohlrogge JB: **Disruption of the FATB gene in Arabidopsis demonstrates an essential role of saturated fatty acids in plant growth.** *Plant Cell* 2003, **15**:1020–33.

245. Raffaele S, Vailleau F, Leger A, Joubes J, Miersch O, Huard C, Blee E, Mongrand S, Domergue F, Roby D: **A MYB transcription factor regulates very-long-chain fatty acid biosynthesis for activation of the hypersensitive cell death response in Arabidopsis**. *Plant Cell* 2008, **20**:752–767.

246. Li Q, Zhang N, Zhang L, Ma H: **Differential evolution of members of the rhomboid gene family with conservative and divergent patterns**. *New Phytol* 2015, **206**:368–380.

247. De Carolis E, De Luca V: **Purification, Characterization, and Analysis of a 2-Oxoglutarate-dependent Dioxygenase Involved in Vindoline Biosynthesis from Catharanthus roseus**. *J Biol Chem* 1993, **268**:5504–5511.

248. Cordell GA: **Fifty years of alkaloid biosynthesis in Phytochemistry**. *Phytochemistry* 2013:29–51.

249. Zhu W, Yang B, Komatsu S, Lu X, Li X, Tian J: **Binary stress induces an increase in indole alkaloid biosynthesis in Catharanthus roseus**. *Front Plant Sci* 2015, **6**(July):1–12.

250. Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K: **AP2/ERF family transcription factors in plant abiotic stress responses**. *Biochim Biophys Acta - Gene Regul Mech* 2012, **1819**:86–96.

251. Nakano T, Suzuki K, Fujimura T, Shinshi H: **Genome-Wide Analysis of the ERF Gene Family**. *Plant Physiol* 2006, **140**(February):411–432.

252. Nakano T, Nishiuchi T, Suzuki K, Fujimura T, Shinshi H: **Studies on transcriptional regulation of endogenous genes by ERF2 transcription factor in tobacco cells**. *Plant Cell Physiol* 2006, **47**:554–558.

253. Sharma M, Pandey GK: **Expansion and Function of Repeat Domain Proteins During Stress and Development in Plants**. 2016, **6**(January):1–15.

254. Chervin C, El-Kereamy A, Roustan JP, Latché A, Lamon J, Bouzayen M: **Ethylene seems required for the berry development and ripening in grape, a non-climacteric fruit**. *Plant Sci* 2004, **167**:1301–1305.

255. Kaps ML, Cahoon G a: **Growth and fruiting of container-grown Seyval blanc grapevines modified by changes in crop level, leaf number and position, and light exposure**. *Am J Enol Vitic* 1992, **43**:191–199.

256. Chakrabarti M, Zhang N, Sauvage C, Muños S, Blanca J, Cañizares J, Diez MJ, Schneider R, Mazourek M, McClead J, Causse M, van der Knaap E, Knaap E: **A cytochrome P450 regulates a domestication trait in cultivated tomato.** *Proc Natl Acad Sci U S A* 2013, **110**:17125–17130.

257. He F, Mu L, Yan GL, Liang NN, Pan QH, Wang J, Reeves MJ, Duan CQ: **Biosynthesis of anthocyanins and their regulation in colored grapes**. *Molecules* 2010:9057–9091.

258. Hichri I, Barrieu F, Bogs J, Kappel C, Delrot S, Lauvergeat V: **Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway**. *Journal of Experimental Botany* 2011:2465–2483.

259. Morrell PL, Buckler ES, Ross-Ibarra J: **Crop genomics: advances and applications**. *Nat Rev Genet* 2011, **13**:85–96.

260. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps**. *Genetics* 2001, **157**:1819–1829.

261. Korte A, Farlow A: **The advantages and limitations of trait analysis with GWAS: a review.** *Plant Methods* 2013, **9**:29.

262. Brachi B, Morris GP, Borevitz JO: **Genome-wide association studies in plants: the missing heritability is in the field**. *Genome Biol* 2011, **12**:232.

263. Pearson TA, Manolio TA: **How to interpret a genome-wide association study**. *Jama* 2008, **299**:1335–1344.

264. Asimit J, Zeggini E: **Rare variant association analysis methods for complex traits.** *Annu Rev Genet* 2010, **44**:293–308.

265. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: **Rare Variants Create Synthetic Genome-Wide Associations**. *PLoS Biol* 2010, **8**.

266. Zhang X, Cal AJ, Borevitz JO: **Genetic architecture of regulatory variation in Arabidopsis thaliana**. *Genome Res* 2011, **21**:725–733.

267. Platt A, Vilhjálmsson BJ, Nordborg M: **Conditions under which genome-wide association studies will be positively misleading**. *Genetics* 2010, **186**:1045–1052.

268. Storey JD, Akey JM, Kruglyak L: **Multiple locus linkage analysis of genomewide expression in yeast**. *PLoS Biol* 2005, **3**.

269. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, Bouchez D, Dillmann C, Guerche P, Hospital F, Colot V: **Assessing the impact of transgenerational epigenetic variation on complex traits**. *PLoS Genet* 2009, **5**.

270. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**:747–53.

271. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES: **Association mapping: critical considerations shift from genotyping to experimental design.** *Plant Cell* 2009, **21**:2194–2202.

272. Vilhjálmsson BJ, Nordborg M: **The nature of confounding in genome-wide association studies.** *Nat Rev Genet* 2013, **14**:1–2.

273. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: **Efficient control of population structure in model organism association mapping.** *Genetics* 2008, **178**:1709–23.

274. Altshuler D, Daly MJ, Lander ES: **Genetic mapping in human disease.** *Science* 2008, **322**:881–8.

275. Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas M a, Zwonitzer JC, Kresovich S, McMullen MD, Ware D, Balint-Kurti PJ, Holland JB: **Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population**. *Nat Genet* 2011, **43**:163–168.

276. Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo Z, Kearsey MJ, Werner P, Harrap D, Tapsell C, Liu H, Hedley PE, Stein N, Schulte D, Steuernagel B, Marshall DF, Thomas WTB, Ramsay L, Mackay I, Balding DJ, Consortium A, Waugh R, O'Sullivan DM: **Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome.** *Proc Natl Acad Sci U S A* 2010, **107**:21611–21616.

277. Harrell Frank E Jr, Dupont C: **Hmisc: Harrell Miscellaneous. R package version 3.17-3. http://CRAN.R-project.org/package=Hmisc**. 2016.

278. Bates D, Maechler M, Bolker B, Walker S: **lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, http://CRAN.R-project.org/package=lme4.** *R Packag version* 2014.

279. Hadfield JD, Wilson AJ, Garant D, Sheldon BC, Kruuk LEB: **The Misuse of BLUP in Ecology and**

Evolution. *Am Nat* 2010, **175**:116–125.

280. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES: **TASSEL: Software for association mapping of complex traits in diverse samples**. *Bioinformatics* 2007, **23**:2633–2635.

281. Endelman JB, Jannink J-L: **Shrinkage Estimation of the Realized Relationship Matrix**. *Genes|Genomes|Genetics* 2013, **2**:1405–1413.

282. Turner SD: *Qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots*. 2014.

283. Coombe BG, McCarthy MG: **Dynamics of grape berry growth and physiology of ripening**. *Aust J Grape Wine Res* 2000, **6**:131–135.

284. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Oropeza Rosas M, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, et al.: **Genetic properties of the maize nested association mapping population.** *Science* 2009, **325**:737–40.

285. Dell'Acqua M, Gatti DM, Pea G, Cattonaro F, Coppens F, Magris G, Hlaing AL, Aung HH, Nelissen H, Baute J, Frascaroli E, Churchill GA, Inzé D, Morgante M, Pè ME: **Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays**. *Genome Biol* 2015, **16**:167.

286. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R: **A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana**. *PLoS Genet* 2009, **5**.

287. Sannemann W, Huang BE, Mathew B, L??on J: **Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept**. *Mol Breed* 2015, **35**.

288. Thépot S, Restoux G, Goldringer I, Hospital F, Gouache D, Mackay I, Enjalbert J: **Efficiently tracking selection in a multiparental population: The case of earliness in wheat**. *Genetics* 2015, **199**:609–623.

289. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Structure of linkage disequilibrium and phenotypic associations in the maize genome.** *Proc Natl Acad Sci U S A* 2001, **98**:11479–11484.

290. McCouch S: **Diversifying selection in plant breeding**. *PLoS Biol* 2004, **2**.

291. Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO: **Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana.** *Proc Natl Acad Sci USA* 2010, **107**:21199–204.

292. Busam G, Kassemeyer HH, Matern U: **Differential expression of chitinases in Vitis vinifera L. responding to systemic acquired resistance activators or fungal challenge**. *Plant Physiol* 1997, **115**:1029–1038.

293. Costanzo E, Trehin C, Vandenbussche M: **The role of WOX genes in flower development**. *Ann Bot* 2014, **114**:1545–1553.

294. Romera-Branchat M, Ripoll JJ, Yanofsky MF, Pelaz S: **The WOX13 homeobox gene promotes replum formation in the Arabidopsis thaliana fruit**. *Plant J* 2013, **73**:37–49.

295. Gerós H, Chaves MM, Delrot S: *The Biochemistry of the Grape Berry*. 2012.

296. Ferguson IB: **Calcium in plant senescence and fruit ripening**. *Plant, Cell Environ* 1984, **7**:477–489.

297. Tak H, Mhatre M: **Cloning and molecular characterization of a putative bZIP transcription factor VvbZIP23 from Vitis vinifera**. *Protoplasma* 2013, **250**:333–345.

298. Nicolas P, Lecourieux D, Kappel C, Cluzet S, Cramer G, Delrot S, Lecourieux F: **The basic leucine zipper transcription factor ABSCISIC ACID RESPONSE ELEMENT-BINDING FACTOR2 is an important transcriptional**

regulator of abscisic acid-dependent grape berry ripening processes. *Plant Physiol* 2014, **164**:365–83.

299. Hichri I, Heppel SC, Pillet J, Léon C, Czemmel S, Delrot S, Lauvergeat V, Bogs J: **The basic helix-loop-helix transcription factor MYC1 is involved in the regulation of the flavonoid biosynthesis pathway in grapevine**. *Mol Plant* 2010, **3**:509–523.

300. Agudelo-Romero P, Bortolloti C, Pais MS, Tiburcio AF, Fortes AM: **Study of polyamines during grape ripening indicate an important role of polyamine catabolism**. *Plant Physiol Biochem* 2013, **67**:105–119.

301. Shiozaki S, Ogata T, Horiuchi S: **Endogenous polyamines in the pericarp and seed of the grape berry during development and ripening**. *Sci Hortic (Amsterdam)* 2000, **83**:33–41.

302. Agudelo-Romero P, Ali K, Choi YH, Sousa L, Verpoorte R, Tiburcio AF, Fortes AM: **Perturbation of polyamine catabolism affects grape ripening of Vitis vinifera cv. Trincadeira**. *Plant Physiol Biochem* 2014, **74**:141–155.

303. Fortes AM, Agudelo-Romero P, Silva MS, Ali K, Sousa L, Maltese F, Choi YH, Grimplet J, Martinez-Zapater JM, Verpoorte R, Pais MS: **Transcript and metabolite analysis in Trincadeira cultivar reveals novel information regarding the dynamics of grape ripening.** *BMC Plant Biol* 2011, **11**:149.

304. Roubelakis-Angelakis KA: *Molecular Biology & Biotechnology of the Grapevine*. *Volume 162*; 2002.

305. Ruffner HP: **Metabolism of Tartaric and Malic Acids in Vitis: A Review-Part B**. *Vitis* 1982, **21**:346–358.

306. Kliewer WM: **Changes in the Concentration of Malates, Tartrates, and total free Acids in Flowers and Berries of Vitis Vinifera**. *Am J Enol Vitic* 1965, **16**:92–100.

307. Yoshida N, Yanai Y, Chen L, Kato Y, Hiratsuka J, Miwa T, Sung ZR, Takahashi S: **EMBRYONIC FLOWER2, a novel polycomb group protein homolog, mediates shoot development and flowering in Arabidopsis.** *Plant Cell* 2001, **13**:2471–2481.

308. Cerdán PD, Chory J: **Regulation of flowering time by light quality.** *Nature* 2003, **423**:881–885.

309. Etienne A, Génard M, Lobit P, Mbeguié-A-Mbéguié D, Bugaud C: **What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells**. *J Exp Bot* 2013, **64**:1451–1469.

310. Terrier N, Glissant D, Grimplet J, Barrieu F, Abbal P, Couture C, Ageorges A, Atanassova R, Léon C, Renaudin JP, Dédaldéchamp F, Romieu C, Delrot S, Hamdi S: **Isogene specific oligo arrays reveal multifaceted changes in gene expression during grape berry (Vitis vinifera L.) development**. *Planta* 2005, **222**:832–847.

311. Storey R: **Potassium Localization in the Grape Berry Pericarp by Energy-Dispersive X-Ray Microanalysis**. *Am J Enol Vitic* 1987, **38**:301–309.

312. Barkla BJ, Pantoja O: **Physiology of Ion Transport Across the Tonoplast of Higher Plants**. *Annu Rev Plant Physiol Plant Mol Biol* 1996, **47**:159–184.

313. Poole RJ: **Energy Coupling for Membrane Transport**. *Annu Rev Plant Physiol* 1978, **29**:437–460.

314. Jackson DI, Lombard PB, Kabinett LQ: **Environmental and Management Practices Affecting Grape Composition and Wine Quality - A Review**. *Am J Enol Vitic* 1993, **44**:409–430.

315. Tanabe S, Ashikari M, Fujioka S, Takatsuto S, Yoshida S, Yano M: **A Novel Cytochrome P450 Is Implicated in Brassinosteroid Biosynthesis via the Characterization of a Rice Dwarf Mutant , dwarf11 , with Reduced Seed Length**. 2005, **17**(March):776–790.

316. Symons GM, Davies C, Shavrukov Y, Dry IB, Reid JB, Thomas MR: **Grapes on steroids. Brassinosteroids are involved in grape berry ripening.** *Plant Physiol* 2006, **140**:150–158.

317. Luan LY, Zhang ZW, Xi ZM, Huo SS, Ma LN: **Brassinosteroids regulate anthocyanin biosynthesis in the ripening of grape berries**. *South African J Enol Vitic* 2013, **34**:196–203.

318. Shin DH, Cho M, Choi MG, Das PK, Lee SK, Choi SB, Park Y Il: **Identification of genes that may regulate the expression of the transcription factor production of anthocyanin pigment 1 (PAP1)/MYB75 involved in Arabidopsis anthocyanin biosynthesis**. *Plant Cell Rep* 2015, **34**:805–815.

319. Gunl M, Neumetzler L, Kraemer F, de Souza a., Schultink a., Pena M, York WS, Pauly M: **AXY8 Encodes an -Fucosidase, Underscoring the Importance of Apoplastic Metabolism on the Fine Structure of Arabidopsis Cell Wall Polysaccharides**. *Plant Cell* 2011, **23**:4025–4040.

320. Evrard J-L, Pieuchot L, Vos JW, Vernos I, Schmit A-C: **Plant TPX2 and related proteins.** *Plant Signal Behav* 2009, **4**:69–72.

321. Hu B, Wang W, Ou S, Tang J, Li H, Che R, Zhang Z, Chai X, Wang H, Wang Y, Liang C, Liu L, Piao Z, Deng Q, Deng K, Xu C, Liang Y, Zhang L, Li L, Chu C: **Variation in NRT1.1B contributes to nitrate-use divergence between rice subspecies**. *Nat Genet* 2015, **47**:834–838.

322. Ji H, Pardo JM, Batelli G, Van Oosten MJ, Bressan RA, Li X: **The salt overly sensitive (SOS) pathway: Established and emerging roles**. *Mol Plant* 2013, **6**:275–286.

323. Quintero FJ, Martinez-Atienza J, Villalta I, Jiang X, Kim W-Y, Ali Z, Fujii H, Mendoza I, Yun D-J, Zhu J-K, Pardo JM: **Activation of the plasma membrane Na/H antiporter Salt-Overly-Sensitive 1 (SOS1) by phosphorylation of an auto-inhibitory C-terminal domain.** *Proc Natl Acad Sci U S A* 2011, **108**:2611–6.

324. Licausi F, Van Dongen JT, Giuntoli B, Novi G, Santaniello A, Geigenberger P, Perata P: **HRE1 and HRE2, two hypoxia-inducible ethylene response factors, affect anaerobic responses in Arabidopsis thaliana**. *Plant J* 2010, **62**:302–315.

325. Tyerman S, Xiao Z, Liao S, Scharwies J, Caravia L: **Cell death in the grape berry**. In *X International Symposiom on Grapevine Phisiology and Biotechnology*; 2016.

326. Afoufa-Bastien D, Medici A, Jeauffre J, Coutos-Thévenot P, Lemoine R, Atanassova R, Laloi M: **The Vitis vinifera sugar transporter gene family: phylogenetic overview and macroarray expression profiling.** *BMC Plant Biol* 2010, **10**:245.

327. Darwin C: *On the Origin of the Species*. *Volume 5*; 1859.

328. Doebley JF, Gaut BS, Smith BD: **The Molecular Genetics of Crop Domestication**. *Cell* 2006, **127**:1309–1321.

329. Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD: **Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice**. *Mol Biol Evol* 2012, **29**:675–687.

330. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, Clark RM, Rätsch G, Mott R: **Multiple reference genomes and transcriptomes for Arabidopsis thaliana.** *Nature* 2011, **477**:419–423.

Appendix A

Vitis **54** (Special Issue), 233–237 (2015)

Genetic diversity and relationships in the grapevine germplasm collection from Central Asia

A. Marrano[1)], L. Grzeskowiak[1)], P. Moreno Sanz[1)], S. Lorenzi[1)], M. L. Prazzoli[1)], A. Arzumanov[2)], M. Amanova[2)], O. Failla[3)], D. Maghradze[4)] and M. S. Grando[1)]

[1)] Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige (Trento) Italy
[2)] Uzbek Research Institute of Plant Industry, Tashkent, Uzbekistan
[3)] Dipartimento di Scienze Agrarie e Ambientali, Università degli Studi di Milano, Italy
[4)] Institute of Horticulture, Viticulture and Oenology, Agricultural University of Georgia, Georgia

Summary

The mountainous region between the Caucasus and China is considered the center of diversity for many temperate fruit crops. Also the transitional types of grapes, including wild forms of the subsp. *Vitis sylvestris*, cultivated landraces and ancient local varieties, were once common in this region. Despite Central Asia is considered a focal region of the world regarding grapevine development, limited information about the extent and distribution of grapevine genetic variation is available.

Here we report the first assessment of genetic diversity, relationships and structure of 80 grapevine cultivars and 21 *V. sylvestris* accessions originated from the regions of Uzbekistan, Tajikistan and Kyrgyzstan. We expanded the coverage of this survey to include a set of 53 traditional Georgian varieties and homologous SSR genotypes of 107 cultivars representing four *V. vinifera* ancestral subpopulations. This allowed us to evaluate the contribution of the Central Asian grapevine germplasm to diversification of the cultivated grapevine gene pool.

K e y w o r d s : SSR marker profiles; Georgia; Uzbekistan; Tajikistan; Kyrgyzstan.

Introduction

Based on archaeological evidence dating from 8000 BC and the large genetic diversity, South Caucasus and Anatolia have long been regarded as homelands of viticulture (Vavilov 1931, Vouillamoz *et al.* 2006). Historical records suggest that cultivation of *V. vinifera* was spread to North Africa by the end of the fifth millennium BC, and it was established in Europe during the first millennium BC. Grape culture is supposed to have reached Afghanistan and the oases of Central Asia by the fourth century BC, and China in the second century BC (Lutz 1922, Vavilov 1931, Negrul 1946, Levadoux 1956, Mc Govern 2003).

According to Negrul (1946) who traveled widely throughout Europe and Central Asia, the grapevines found in the wide area extending from eastern Georgia, Armenia, Azerbaijan to the former Soviet republics in Central Asia and the region of the Near East have clear distinguishing features and were placed in the Proles *orientalis*. Negrul recognised two sub-proles within this main group: *caspica*, composed of ancient vines used for vinification before the advent of Islam (from AD 500-1100), and the *antasiatica* including cultivars for table grape and raisins of more recent origin. Varietal ecotypes found from Georgia to the Balkans were instead designated *P. pontica* sub-proles *georgica* and sub-proles *balkanica*, respectively.

Further extensive field investigations into natural populations of *V. vinifera* led Negrul to conclude that cultivars from the region of the Caspian Sea (sub-proles *caspica*) were so different from the Proles *pontica* that they must have arisen from a different wild form. He called it *V. sylvestris* var. *aberrans* the vine form with hairless leaf surface as opposed to the most widespread *V. sylvestris* var. *typical* having hairy leaves.

Molecular analysis has provided, for almost two decades, new insights on genetic diversity of *V. vinifera* in relation to wild relatives, origin of cultivars and specific alleles linked to selected traits (Arroyo-Garcia *et al.* 2006, This *et al.* 2006, Emanuelli *et al.* 2010). However, despite Central Asia is considered a focal region of grapevine development, information about the amount and distribution of grapevine genetic variation have only recently started to emerge and it is based on accessions from Central Asian countries maintained in European and USA germplasm repositories. These materials were included in genetic studies aimed at interpreting the population structure of cultivated varieties as well as to further investigate the intriguing resistance to *Erysiphe necator* found in some eastern *V. vinifera* forms (Hoffmann *et al.* 2008, Bacilieri *et al.* 2013, Riaz *et al.* 2013).

Here we report the first assessment of genetic diversity, relationships and structure of grapevine cultivars conserved in the local collection of the Uzbek Research Institute of Plant Industry ('UzRIPI'; Tashkent region, Republic of Uzbekistan) including several *V. sylvestris* accessions. We expanded the coverage of this preliminary survey to include a set of traditional Georgian varieties and homologous SSR genotypes of cultivars representing four *V. vinifera* ancestral subpopulations (Emanuelli *et al.* 2013). This allowed us to evaluate the contribution of Central Asian grapevine germplasm to diversification of the cultivated gene pool.

Correspondence to: Dr. M. S. Grando, Fondazione Edmund Mach, Research and Innovation Centre, Via E. Mach 1, 38010 San Michele all'Adige (Trento) Italy. E-mail: stella.grando@fmach.it

Material and Methods

Plant material and SSR analysis: A grapevine (*V. vinifera* L.) collection of 80 cultivated and 21 supposed wild accessions from the region of Central Asia (Uzbekistan, Tajikistan and Kyrgyzstan) and 53 cultivars from Georgia was analyzed (Tab. 1). Leaf samples were placed in 96-well microtube plates and freeze-dried. DNA extraction was performed using DNAeasy 96 plan mini kit (Qiagen, Germany).

Table 1

Origin of grapevine accessions analysed in the present study.

Area of origin	Number of accessions	Institutional source
Tajikistan	52	UzRIPi
Uzbekistan	23	UzRIPi
Kyrgyzstan	4	UzRIPi
Central Asia*	22	UzRIPi
Georgia	53	GEO015

*the exact geographic location is unknown

Twenty two SSR markers, including at least one locus per chromosome, were chosen to profile the whole collection of 154 accessions. This set includes the nine SSR markers proposed by the European Project GrapeGen06 for the characterization of regional cultivars (Maul *et al.* 2012) and the loci VVIQ52, VVIN16, VVIV37, VVIH54, VVIN73, VVIP31, VVIB01, VVIV67 (Merdinoglu *et al.* 2005), VVMD21, VVMD24 (Bowers and Meredith 1999), VMC4F3.1 (BV722689), VMC4F8 (BV102437) and VMC1B11 (BV681754).

Nine multiplex panels of fluorescently labeled markers were used as reported in Emanuelli *et al.* (2013). The PCR products were denatured and size fractionated using capillary electrophoresis on an ABI 3130 Genetic Analyzer (Applied Biosystems). GeneMapper v3.5 (Applied Biosystems) was used for the alleles size estimation.

Genetic diversity assessment: The final dataset of non-redundant genotypes was used to estimate the main diversity statistics, such as total number of different alleles per locus (N_A), number of effective alleles (N_E, the number of equally frequent alleles required to give the observed level of heterozygosity), observed (H_O) and expected (H_E) heterozygosity and fixation index (F, inbreeding coefficient) through GenAlex v6.501 (Peakall and Smouse 2006, 2012).

Analysis of population structure: The genetic structure was first assessed by principal coordinate analysis (PCoA), implemented in GenAlex v6.501. Genotypic data were then subjected to the Bayesian clustering analysis, implemented in STRUCTURE 3.2 (Pritchard *et al.* 2000) using the admixed and correlated allele frequency models. Ten independent runs for K values ranging from 1 to 10 were performed with a burn-in length of 10,000 followed by 100,000 iterations. The most likely subdivision (K) was established by plotting the log probability L(K) and ΔK of the data over ten runs, as implemented in Structure Harvester v0.6.94 (Evanno *et al.* 2005, Earl and von Holdt 2012).

The unique genetic profiles at 22 SSR loci were further subjected to cluster analysis using the Darwin software package v6.0 (Perrier and Jacquemoud-Collet 2006). A weighted neighbour-joining tree was constructed based on the simple matching dissimilarity matrix with 100 bootstrap replicates. Further cluster analysis was performed including the genetic profiles of 107 cultivars which belong to the FEM germplasm collection (ITA362) and represent four subpopulations of *V. vinifera*, in accordance with the eco-geographic origin of the cultivars (Emanuelli *et al.* 2013). In addition, the SSR profile of 11 grape rootstock (*Vitis* spp.) varieties were used for outgroup comparisons.

Results and Discussion

Pairwise comparisons based on SSR profiles at 9 loci led to the identification of 11 and 10 synonymous groups in the Central Asia and Georgian subsets respectively, comprising 35 accessions overall. The final dataset of distinct SSR profiles was composed of 13 wild and 66 cultivated genotypes from Central Asia and 40 Georgian cultivars.

A comparison of the SSR genotypes with those reported in the European *Vitis* Database (www.eu-vitis.de) revealed that the three different Georgian accessions 'Saperavi Budeshuriseburi' (a 'Saperavi' mutant), 'Kisi' and 'Ikaltos Tsiteli' matched at all the nine tested loci with the following entries: 'Saperavi' (DEU098-1993-253, ITA035-118), 'Goruli mtsvane' (ITA035-69) and 'Rkatsiteli' (DEU098-1980-083, AUT024-319), respectively. It is worth noting that the cultivated Uzbek varieties, 'Bishti' and 'Ruzbari', had identical SSR profiles as 'Lambrusque Carranques' 3, a *V. sylvestris* accession (FRA139-8500Mtp164), and 'Rund Weiss' (FRA139-0Mtp1002) a cultivar thought to have originated in Azerbaijan, respectively. These findings deserve additional investigation on the accessions' morphological descriptors which were not integrated in the dataset.

The panel of 119 unique genotypes was characterized at 13 more SSR loci in order to estimate the main indexes of genetic diversity separately in the three subsets: cultivated and wild accessions from Central Asia and cultivars from Georgia, and to assess the relationship among the accessions. The average number of different alleles per locus in the whole sample was 11.2 and 64 % of alleles were shared among these three groups. Genetic diversity parameters, summarized in Tab. 2, revealed higher levels of expected and observed heterozygosity in the cultivated compartment, compared to the small group of wild individuals. The amount of variation was similar to that reported for larger samples of *V. vinifera* germplasm (Bacilieri *et al*. 2013; Emanuelli *et al*. 2013).

Genetic relationships were investigated using the principle coordinate analysis (PCoA) and Structure approaches. The PCoA, based on a genetic distance matrix with data standardization, explored how the Georgian group may be differentiated from the Central Asia populations. Plotting of the first two principle coordinates showed a clear separation between cultivated accessions from

Table 2

Summary statistics for 119 genotypes from three populations assessed using 22 SSR markers (N = sample size; N_a = N° of different alleles per locus; N_e = N° of effective alleles; N_p = N° of private alleles; Ho = observed heterozygosity; He = expected heterozygosity; F = Fixation Index)

Population	N	N_a	N_e	Np	H_o	H_e	F
Central Asia	66	9.55	4.86	54	0.76	0.77	0.01
Wild Central Asia	13	4.91	3.39	0	0.64	0.66	0.02
Georgia	40	8.05	4.24	34	0.75	0.73	-0.03

Central Asia and Georgian varieties along the first axis, whereas the wild and cultivated genotypes of Central Asia were distinguished, though to a lesser degree, along the second coordinate (Fig. 1). The subdivision of populations originating from Central Asia, with respect to those from the Caucasus region, was not very evident in the previous structure analysis of the large grape collection of Vassal (INRA, France) performed by BACILIERI *et al*. (2013). However, similarly to our findings genotypes from the eastern regions subdivided into two sub-groups according to the main local use of grapevines: wine, for the Caucasian cultivars and table, for the Central Asian cultivars.

The unique profiles at 22 SSR loci were used for Bayesian clustering analysis implemented in STRUCTURE. The most likely number of clusters (K), obtained using the ΔK method proposed by Evanno et al. (2005), was equal to K = 2. Using a threshold of cluster membership coefficient equal to 0.80, 37 out of 40 Georgian genotypes were assigned to the cluster K_1 and all 79 individuals from Central Asia were included into the group K_2. Thus, a very low level of admixture characterizes the Caucasian sample analyzed in this study. This is in agreement with the findings of IMAZIO *et al*. (2013) on a different portion of the Georgian germplasm. The absence of admixture observed in the Central Asian populations is intriguing, and it may raise questions regarding the spatial and temporal patterns of grapevine domestication.

The topology of the weighted neighbor joining dendrogram including the genotypes from Georgia and Central Asia reflected as well two major groups, in accordance with the geographic origin of the samples (data not shown). Moreover, within the Central Asian grapevines, most of the supposed wild accessions clustered together, and no evidence of genetic differentiation was observed among the subsets from the UzRIPI germplasm collection considered to have originated from the regions of Uzbekistan, Tajikistan and Kyrgyzstan.

To gain a broader understanding of the genetic relationship of these Georgian and Central Asian grapevines we performed a further cluster analysis including the homologous genetic profiles of 107 additional cultivars at the same set of 22 SSR loci. These accessions belong to four ancestral subpopulations of *V. vinifera* ssp. *sativa* (VV) which were detected within a large sample of grapevine accessions following a hierarchical clustering approach (EMANUELLI *et al.* 2013). In particular, we included the cluster of Italian and Greek wine grapes (VV1), representing the proles *pontica*, the French and German wine varieties (VV4), representing the proles *occidentalis*, and the Muscat table and wine grapes (VV3) reflecting the proles *orientalis* subpr. *caspica*. The cluster VV2 was more heterogeneous, and it was composed of both table grape varieties related to 'Sultanina' (proles *orientalis* subpr. *antasiatica*) and Spanish wine grapes.

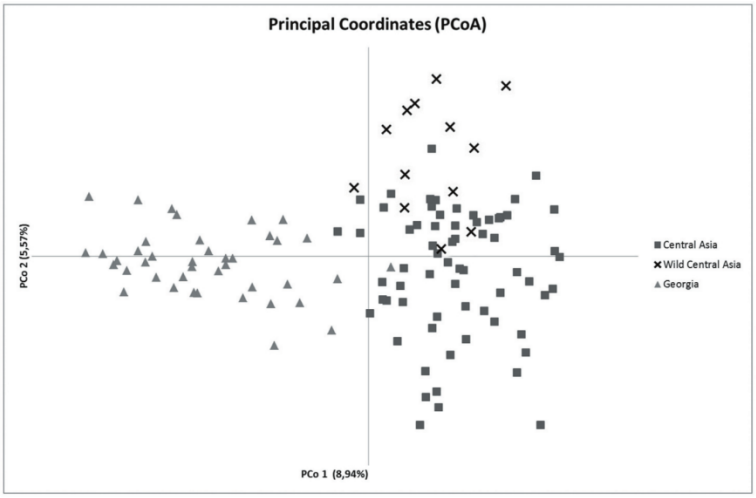

Fig. 1: Scatter plot of the first two principal coordinate analysis axes for the SSR data of 119 genotypes.

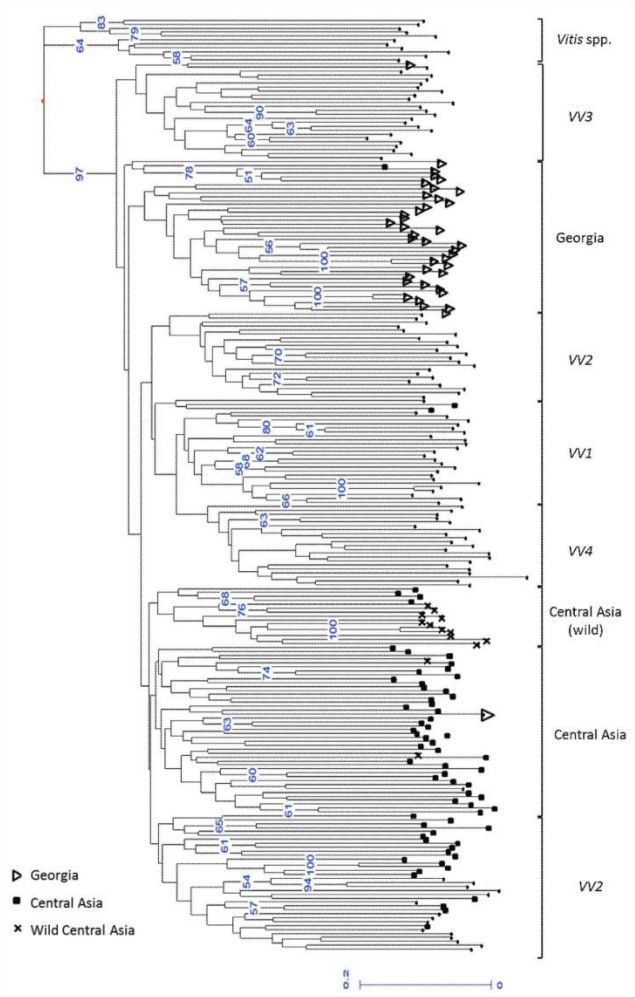

Fig. 2: Neighbor joining weighted tree based on a dissimilarities matrix calculated from SSR alleles at 22 loci for 226 *V. vinifera* genotypes and 11 rootstocks (*Vitis* spp.) as an outgroup. Only bootstraps superior to 50 are presented.

In this context, almost all the Georgian genotypes formed an additional well distinct clade, which likely corresponds to the proles *pontica* subpr. *georgica* Negr (Fig. 2). Most of Central Asian genotypes also grouped together in a large cluster composed of three subclusters, including the table grapes portion of the VV2 population. In fact, since 20 Central Asian genotypes fell within the VV2 population, the previous subgroup of the Spanish wine cultivars was now separated. Most of the cultivated and *V. sylvestris* accessions from the UzRIPI collection composed two distinct subgroups, although it is worth noting that some wild accessions were included within the subgroup of cultivars and *vice versa*.

In conclusion, the grapevine gene pool of Georgia and Central Asia surveyed in this study has a significant amount of genetic variation and exhibits high levels of population differentiation which may reflect a limited historical gene flow between the two regions. In addition to the first molecular description of the genetic diversity of the Central Asian grape germplasm collection currently maintained in local repository, this study contributed to the integration of genotype information on these extremely valuable grapevine genetic resources into The European *Vitis* Database.

References

ARROYO-GARCÍA, R.; RUIZ-GARCÍA, L.; BOLLING, L.; OCETE, R.; LÓPEZ, M. A.; ARNOLD, C.; ERGUL, A.; SÖYLEMEZOĞLU, G.; UZUN, H. I.; CABELLO, F.; IBÁÑEZ, J.; ARADHYA, M. K.; ATANASSOV, A.; ATANASSOV, I.; BALINT, S.; CENIS, J. L.; COSTANTINI, L.; GORIS-LAVETS, S.; GRANDO, M.

S.; Klein, B. Y.; McGovern, P. E.; Merdinoglu, D.; Pejic, I.; Pelsy, F.; Primikirios, N.; Risovannaya, V.; Roubelakis-Angelakis, K. A.; Snoussi, H.; Sotiri, P.; Tamhankar, S.; This, P.; Troshin, L.; Malpica, J. M.; Lefort, F.; Martinez-Zapater, J. M.; 2006: Multiple origins of cultivated grapevine (*Vitis vinifera* L. ssp. *sativa*) based on chloroplast DNA polymorphisms. Mol. Ecol. **15,** 3707-3714.

Bacilieri, R.; Lacombe, T.; Le Cunff, L.; Di Vecchi-Staraz, M.; Laucou, V.; Genna, B.; Péros, J. P.; This, P.; Boursiquot, J. M.; 2013: Genetic structure in cultivated grapevines is linked to geography and human selection. BMC Plant Biol. **13**, 25.

Belkhir, K.; Borsa, P.; Chikhi, L.; Raufaste, N.; Bonhomme, F.; Genetix 4.05. CNRS UMR 5171. Montpellier: Université de Montpellier II; http://www.genetix.univmontp2.fr/genetix/intro.htm.

Bowers, J. E.; Dangl, G. S.; Meredith, C. P.; 1999: Development and characterization of additional microsatellite DNA markers for grape. Am. J. Enol. Vitic. **50**, 243-246.

Earl, D. A.; von Holdt, B. M.; 2012: Structure Harvester: a website and program for visualizing Structure output and implementing the Evanno method. Conserv. Genet. Resour. **4**, 359-361.

Emanuelli, F.; Battilana, J.; Costantini, L.; Le Cunff, L.; Boursiquot, J. M.; This, P.; Grando, M. S.; 2010: A candidate gene association study on muscat flavor in grapevine (*Vitis vinifera* L.). BMC Plant Biol. **10**, 241.

Emanuelli, F.; Lorenzi, S.; Grzeskowiak, L.; Catalano, V.; Stefanini, M.; Troggio, M.; Myles, S.; Martinez-Zapater, J. M.; Zyprian, E.; Moreira, F. M.; Grando, M. S.; 2013: Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biol. **13**, 39.

Evanno, G.; Regnaut, S.; Goudet, J.; 2005: Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. **14**, 2611-2620.

Hoffmann, S.; Di Gaspero, G.; Kovács, L.; Howard, S.; Kiss, E.; Galbács, Z.; Testolin, R.; Kozma, P.; 2008: Resistance to *Erysiphe necator* in the grapevine 'Kishmish vatkana' is controlled by a single locus through restriction of hyphal growth. Theor. Appl. Genet. **116**, 427-438

Imazio, S.; Maghradze, D.; De Lorenzis, G.; Bacilieri, R.; Laucou, V.; This, P.; Scienza, A.; Failla, O.; 2013: From the cradle of grapevine domestication: molecular overview and description of Georgian grapevine (*Vitis vinifera* L.) germplasm. Tree Genet. Genomes **9**, 641-658.

Kimura, M.; Crow, J. F.; 1964: The number of alleles that can be maintained in a finite population. Genetics **49**, 725-738.

Laucou, V.; Lacombe, T.; Dechesne, F.; Siret, R.; Bruno, J. P.; Dessup, M.; Dessup, T.; Ortigosa, P.; Parra, P.; Roux, C.; Santoni, S.; Varès, D.; Péros, J. P.; Boursiquot, J. M.; This, P.; 2011: High throughput analysis of grape genetic diversity as a tool for germplasm collection management. Theor. Appl. Genet. **122**, 1233-1245.

Levadoux, L. D.; 1956: Wild and cultivated populations of *Vitis vinifera* L. Ann. Amelior. Plant **6**, 59-118.

Lutz, H. F.; 1922: Viticulture and Brewing in the Ancient Orient. J. C. Hinrichs'sche Buchhandlung, Leipzig.

Maul, E.; Töpfer, R.; Eibach, R.; 2012: Vitis International Variety Catalogue. Julius Kühn Institut. www.vivc.de.

Merdinoglu, D.; Butterlin, G.; Bevilacqua, L.; Chiquet, V.; Adam-Blondon, A. F.; Decroocq, S.; 2005: Development and characterization of a large set of microsatellite markers in grapevine (*Vitis vinifera* L.) suitable for multiplex PCR. Mol. Breed. **15,** 349-366.

Mcgovern, P. E.; 2003: Ancient Wine: The Search for the Origins of Viniculture. Princeton: Princeton University Press;.

Negrul, A. M.; 1946: Origin and classification of cultivated grape. In: A. K. Y. Baranov, M. A. Lazarevski, T. V. Palibin, N. N. Prosmoserdov (Eds): The Ampelography of the USSR, vol 1, 159-216. Pischepromizdat, Moscow.

Nei, M.; 1973: Analysis of gene diversity in subdivided populations. PNAS **70**, 3321–3323.

Peakall, R.; Smouse, P. E.; 2006: Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecol. Notes **6**, 288-295.

Peakall, R.; Smouse, P. E.; 2012: GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research - an update. Bioinformatics **19**, 2537-2539.

Perrier, X.; Jacquemoud-Collet, J. P.; 2006: DARwin software. http://darwin.cirad.fr

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

Riaz, S.; Boursiquot, J. M.; Dangl, G. S.; Lacombe, T.; Laucou, V.; Tenscher, A. C.; Walker, M. A.; 2013: Identification of mildew resistance in wild and cultivated Central Asian grape germplasm. BMC Plant Biol. **13**,149.

Sefc, K. M.; Regner, F.; Glössl, J.; Steinkellner, H.; 1999: Identification of microsatellite sequences in *Vitis riparia* and their applicability for genotyping of different *Vitis* species. Genome **42**, 367-373.

This, P.; Lacombe, T.; Thomas, M. R.; 2006: Historical origins and genetic diversity of wine grapes. Trends Genet **22,** 511-519.

Vavilov, N. I.; 1931: Dikie rodichi plodovykh dereviev Aziatskoi chasti SSSR i Kavkaza i problema proiskhozhdenia plodovykh dereviev (Wild progenitors of the fruit trees of Turkestan and the Caucasus and the problem of the origin of fruit trees). Bull. Appl. Bot. Genet. Plant Breed. **6**, 85-134 (in Russian).

Vouillamoz, J.; Mcgovern, P. E.; Ergul A Söylemezoğlu, G.; Tevzadze, G.; Meredith, C. P.; Grando M. S.; 2006: Genetic characterization and relationships of traditional grape cultivars from Transcaucasia and Anatolia. Plant Genet. Resour. **4**, 144-158.

Appendix B

Table A1. List of the grapevine accessions included in the research panels of chapters 2, 3 and 4 of the present thesis. 'True-to-type' varieties are marked in bold. Samples removed for the high missing rate at SNPs loci in chapters 3 and 4 are checked in red.

Sample ID	Specie	Accession name	Used in Chapter 2	Used in Chapter 3	Used in Chapter 4
GRAPE_01	sativa	Alba aganin isyoum	√	√	√
GRAPE_02	sativa	**Alarjie**	√	√	√
GRAPE_03	sativa	**Arnsburger**	√	√	√
GRAPE_04	sativa	**Brustiano**	√	√	√
GRAPE_05	sativa	**Forsellina**	√	√	√
GRAPE_06	sativa	**Gewuerztraminer**	√	√	√
GRAPE_07	sativa	**Leon Millot**	√	√	√
GRAPE_08	sativa	Beli Medenac	√	√	√
GRAPE_09	sativa	**Macabeu**	√	√	√
GRAPE_10	sativa	**Mornen noir**	√	√	√
GRAPE_11	sativa	**Lambrusco casetta**	√	√	√
GRAPE_12	sativa	**Corbera**	√	√	√
GRAPE_13	sativa	**Reze**	√	√	√
GRAPE_14	sativa	**Roussanne**	√	√	√
GRAPE_15	sativa	**Csaba gyongye**	√	√	√
GRAPE_16	sativa	**Pinot Grigio**	√		√
GRAPE_17	sativa	**Saperavi**	√	√	√
GRAPE_18	sativa	**Malvasia Istriana**	√	√	√
GRAPE_19	sativa	**Jacquere**	√	√	√
GRAPE_20	sativa	**Zilavka**	√	√	√
GRAPE_21	sativa	**Vernaccia di S.Gimignano**	√	√	√
GRAPE_22	sativa	**Shiraz**	√	√	√
GRAPE_23	sativa	**Claverie coulard**	√	√	√
GRAPE_24	sativa	**Ak chekerek**	√	√	√
GRAPE_25	sativa	**Ortrugo**	√	√	√
GRAPE_26	sativa	**Limnio**	√	√	√
GRAPE_27	sativa	**Canorroio**	√	√	√
GRAPE_28	sativa	**Pinot Meunier**	√		√
GRAPE_29	sativa	**Pinot Noir**	√	√	√
GRAPE_30	sativa	**Verdelet**	√	√	√
GRAPE_31	sativa	**Pignoletto**	√	√	√
GRAPE_32	sativa	Aris	√	√	√
GRAPE_33	sativa	Nevado	√	√	√
GRAPE_34	sativa	Moscato	√	√	√
GRAPE_35	sativa	Piè di Palombo	√	√	√
GRAPE_36	sativa	Rossola	√	√	√
GRAPE_37	sativa	Castor	√	√	√
GRAPE_38	sativa	Armenia chi 10	√	√	√
GRAPE_39	sativa	Trollinger Rot	√	√	√
GRAPE_40	sativa	Espadeiro blanco	√	√	√
GRAPE_41	sativa	Muscat Bleu	√	√	√
GRAPE_42	sativa	Bracciola nera	√	√	√
GRAPE_43	sativa	Semidano	√	√	√

GRAPE_44	sativa	Soleil Blanc	√	√	√
GRAPE_45	sativa	Buffalo	√	√	√
GRAPE_46	sativa	Ak ouzioum tagapskii	√	√	√
GRAPE_47	sativa	Ahmed	√	√	√
GRAPE_48	sativa	V.berlandieri Colombard	√	√	√
GRAPE_49	sativa	V,silvestris Lauri 2	√	√	√
GRAPE_50	sativa	V,silvestris cl, Guemuld 103-64	√	√	√
GRAPE_51	sativa	Pinot Noir line 40024	√		
GRAPE_52	sylvestris		√	√	√
GRAPE_53	sylvestris		√	√	√
GRAPE_54	sylvestris		√	√	√
GRAPE_55	sylvestris		√	√	√
GRAPE_56	sylvestris		√	√	√
GRAPE_57	sylvestris		√	√	√
GRAPE_58	sylvestris		√	√	√
GRAPE_59	sylvestris		√	√	√
GRAPE_60	sylvestris		√	√	√
GRAPE_61	sylvestris		√	√	√
GRAPE_62	sylvestris		√	√	√
GRAPE_63	sylvestris		√	√	√
GRAPE_64	sylvestris		√	√	√
GRAPE_65	sylvestris		√	√	√
GRAPE_66	sylvestris		√	√	√
GRAPE_67	sylvestris		√	√	√
GRAPE_68	sylvestris		√	√	√
GRAPE_69	sylvestris		√	√	√
GRAPE_70	sylvestris		√	√	√
GRAPE_71	sylvestris		√	√	√
GRAPE_72	sylvestris		√	√	√
GRAPE_73	sylvestris		√	√	√
GRAPE_74	sylvestris		√	√	√
GRAPE_75	sylvestris		√	√	√
GRAPE_76	sylvestris		√	√	√
GRAPE_77	sylvestris		√	√	√
GRAPE_78	sylvestris		√	√	√
GRAPE_79	sylvestris		√	√	√
GRAPE_80	sylvestris		√	√	√
GRAPE_81	sylvestris		√	√	√
GRAPE_82	sylvestris		√	√	√
GRAPE_83	sylvestris		√	√	√
GRAPE_84	sylvestris		√	√	√
GRAPE_85	sylvestris		√	√	√
GRAPE_86	sylvestris		√	√	√
GRAPE_87	sylvestris		√	√	√
GRAPE_88	sylvestris		√	√	√
GRAPE_89	sylvestris		√	√	√
GRAPE_90	sylvestris		√	√	√
GRAPE_91	sylvestris		√	√	√
GRAPE_92	sylvestris		√	√	√
GRAPE_93	sylvestris		√	√	√
GRAPE_94	sylvestris		√	√	√
GRAPE_95	sylvestris		√	√	√

Acknowledgements

If I look back to my first day of PhD, I see a young girl, hesitant, naive but greatly enthusiastic to follow one of her major passion: plant genetics. I was not expecting so many emotions: highs and lows in a constant cycle, which deeply changes me in a woman. The faces of all people I have met during the last four years follow one another in my mind: everyone has left and taken a piece of me, contributing to my huge metamorphosis in a PhD. I would like to thank all these faces: every single smile, tear, madness, fear, anger, disillusion and hope you gave to me in these years, have made me the person I am today. Thanks!

I wish to express my gratitude to my supervisors Dr Maria Stella Grando and Prof. Giorgio Valle for giving me the opportunity to conduct this PhD study and to introduce me in the research world. I would like also to thank Prof David Neale to have hosted me in his lab at UC Davis and to have showed me another side of research. Thanks to all of you for the time you have dedicated to me by listening, teaching and discussing together.

I would like to thank the member of my research group "Grapevine Applied Genomics", especially Silvia, Paula and Francesco for being much more than co-workers. Working in a friendly environment where there is not insane competition but just the necessity to improve our knowledge, grow up together and support each other, was for me essential during these four years. Thank to make me feeling at home and to be always on my side with your precious advices and help.

A special thanks goes to Lucia and Daniele, that have accompanied me in this curvy trip. Thanks for bearing and cheering me up during all my bad moments (that have been too many! I know...). Some personal relationships evolve from a simple harmony to a strong friendship to a solid bond, which will be there nevertheless the distances and the natural evolution of life. Our friendship is among the solid bonds of my life, with our jokes, fights and reciprocal respect. I would like also to thank the 'FEM PhD community', especially Mario, Veronica and Stefano. We have proven how ambition and collaboration are not opposites, but are the trick to grow up and reach every goal. Thanks for shearing all the moments of happiness, frustation, madness and comparison. I will always remember our coffee room, where every shared coffee was a request for help, talking, laughing or just for escaping from our PhD mess.

I would like to thank my friends in Davis, which was for me a second home. The months spent in CA represent the best experience in my life. I have learnt how there are not culture barriers but just the necessity of curiosity and tolerance. Cultural differences are a great resource that make this World so exciting to live and explore. A special thanks goes to Angela and Valentina, the other two founders of the Italian Trinity. Shearing this experience with you was a privilege: it would never be so intense and extraordinary without you.

I wish to thank all my family, especially my parents, to have left me free of choosing and building my own life up. I am sure you will be always behind me, looking to my steps, supporting and motivating me. I owed you my passionate nature, obstinacy and humility. Sometimes it's hard being away from home, but you are able to fill in this lack with our usual discussions, misunderstandings and laughs.

I would like to thank my best friend Maria Teresa (Anti) for your constant partecipation in my life. We are not anymore the dreaming girls, sat on the stairs outside the school, talking about our love troubles and the school strike of the day after. Some of our expectations went wrong, but other new dreams have cheered our life up. Let's try to enjoy the unpredictability of life and to take the much as we can from every experience our future will offer us!

Last but not the least, I wish to express my immense gratitude to Juan Manuel. You are the most precious gift this PhD experience gave me. Thanks for listening, understanding and supporting me everytime I need. You are my landmark, my best friend, my model in research and life. You are my alter ego that complete and enrich me. If today I am here, it's because of your advices and extreme patience. I will never be afraid of every choice, change and adventure, if you will be on my side.