UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS AGRÓNOMOS

**Association study of phenology, yield and
quality related traits in table grapes using
SSR and SNP markers**

TESIS DOCTORAL

Belkacem Zarouri

Ingeniero agrónomo

2016

Association study of phenology, yield and quality related traits in table grapes using SSR and SNP markers

Autor

Belkacem Zarouri, Ingeniero Agrónomo

Directores

José Antonio Cabezas Martínez, Doctor en Ciencias Biológicas

María Teresa de Andrés Domínguez, Doctora en Ciencias Biológicas

CAMPUS
DE EXCELENCIA
INTERNACIONAL

UNIVERSIDAD
POLITÉCNICA DE MADRID
RECTORADO - EDIFICIOS B Y C
Paseo de Juan XXIII, 11
28040 Madrid

VICERRECTOR DE PLANIFICACION ACADEMICA Y
DOCTORADO

Madrid, 08 de Abril de 2016

Firmado digitalmente por: ENTIDAD UNIVERSIDAD
POLITECNICA DE MADRID - CIF Q2818015F - NOMBRE
ELORZA TENREIRO FCO JAVIER - NIF 32431055C
Motivo: CDUPM, 7 de abril de 2016
Localización: Madrid
Fecha y hora: 08.04.2016 20:37:27

Fdo. FRANCISCO JAVIER ELORZA TENREIRO

La Comisión de Doctorado de la UPM en su reunión del 7 de abril de 2016, ha aprobado la 'Solicitud de propuesta de tribunal' para la tesis titulada '*Association study of phenology, yield and quality related traits in table grapes using SSR and SNP markers*' de BELKACEM ZAROURI (09935777), alumno/a del Programa 'Doctorado en Biotecnología y Recursos Genéticos de Plantas y Microorganismos Asociados' de la 'ESCUELA TECNICA SUPERIOR DE INGENIEROS AGRONOMOS' (02D8).

A continuación se muestra el tribunal asignado para la defensa de tesis:

PRESIDENTE:

D/D^a JESUS M. ORTIZ MARCIDE
PROFESOR EMERITO. ETS INGENIERIA AGRONOMICA, ALIMENTARIA Y DE BIOSISTEMAS -
UNIVERSIDAD POLITECNICA DE MADRID

SECRETARIO:

D/D^a SANTIAGO MORENO VAZQUEZ
PROFESOR TITULAR DE UNIVERSIDAD. ETS INGENIERIA AGRONOMICA, AGROALIMENTARIA Y DE
BIOSISTEMAS - UNIVERSIDAD POLITECNICA DE MADRID

VOCALES:

D/D^a YOLANDA GOGORCENA AOIZ
INVESTIGADOR CIENTIFICO DEL CSIC. ESTACION EXPERIMENTAL DE AULA DEI - CONSEJO SUPERIOR
DE INVESTIGACIONES CIENTIFICAS

D/D^a FELIX CABELLO SAENZ DE SANTA MARIA
DIRECTOR DEPARTAMENTO DE INVESTIGACION AGROALIMENTARIA. INSTITUTO MADRILEÑO DE
INVESTIGACION Y DESARROLLO RURAL, AGRARIO Y ALIMENTARIO (IMIDRA) - COMUNIDAD DE MADRID

D/D^a NILO MEJIA
INVESTIGADOR. CENTRO REGIONAL LA PLATINA - INSTITUTO DE INVESTIGACIONES AGROPECUARIAS.
SANTIAGO DE CHILE

SUPLENTE:

D/D^a MARIA TERESA CERVERA GOY
CIENTIFICO TITULAR DE POIS. INIA-CIFOR (CENTRO DE INVESTIGACION FORESTAL) - INSTITUTO
NACIONAL DE INVESTIGACION Y TECNOLOGIA AGRARIA Y ALIMENTARIA (INIA)

D/D^a DIEGO LIJAVETZKY
CIENTIFICO INDEPENDIENTE Y JEFE DE GRUPO. CONSEJO NACIONAL DE INVESTIGACIONES

D. ANTONIO MOLINA FERNANDEZ Presidente de la Comisión Académica del programa de 'Doctorado en Biotecnología y Recursos Genéticos de Plantas y Microorganismos Asociados' (02D8). ESCUELA TECNICA SUPERIOR DE INGENIEROS AGRONOMOS. C/c Subdirección responsable de los Estudios de Doctorado ESCUELA TECNICA SUPERIOR DE INGENIEROS AGRONOMOS.

CAMPUS
DE EXCELENCIA
INTERNACIONAL

UNIVERSIDAD
POLITÉCNICA DE MADRID
RECTORADO · EDIFICIOS B Y C
Paseo de Juan XXIII, 11
28040 Madrid

CIENTIFICAS Y TECNOLOGICAS - UNIVERSIDAD NACIONAL DE CUYO - UNIVERSIDAD NACIONAL DE CUYO. ARGENTINA

Lo que le comunico para su conocimiento y efectos oportunos.

POLITÉCNICA

**AUTORIZACIÓN DEL DIRECTOR DE TESIS PARA SU ADMISIÓN A TRÁMITE DE
LECTURA
(ART. 5A DEL REGLAMENTO DE ELABORACIÓN Y EVALUACIÓN DE LA TESIS DOCTORAL)**

Título de la tesis doctoral:

Nombre del autor/a de la tesis: D. Belkacem Zarouri

Título de la tesis presentada: Association study of phenology, yield and quality related traits in table grapes using SSR and SNP markers

Director/es de la tesis:

Nombre: D. José Antonio Cabezas Martínez

Centro y Departamento al que pertenece: INIA-CIFOR. Departamento de Ecología y Genética Forestal, 28040, Madrid / INIA-UPM, Unidad mixta de Genómica y Ecofisiología Forestal, Madrid, España.

Nombre: D^a. María Teresa de Andrés Domínguez

Centro y Departamento al que pertenece: Laboratorio de Biología Molecular. Departamento de Investigación Agroalimentaria, Instituto Madrileño de Investigación y Desarrollo Rural, Agrario y Alimentario (IMIDRA), Finca "El Encín", Alcalá de Henares (Madrid), España.

Valoración de la tesis doctoral por parte del director/de los directores

Los estudios de asociación a nivel de todo el genoma (Genome-Wide Association Studies; GWAS) son una herramienta que en los últimos años ha demostrado su utilidad para la identificación de loci implicados en el control de la expresión de caracteres relevantes en especies cultivadas. En esta tesis doctoral nos planteamos comprobar la utilidad que este tipo de estudios puede tener para la identificación de regiones implicadas en la regulación de la expresión de caracteres de interés en la vid, y más concretamente en la uva de mesa. Durante su periodo de formación, el doctorando, Belkacem Zarouri, ha desarrollado con plena satisfacción las tareas de investigación inherentes a su capacitación dentro de las líneas de trabajo que le fueron encomendadas. Como resultado final, está esta tesis, llevada a cabo bajo nuestra dirección en el IMIDRA, que consideramos reúne todos los requisitos para ser defendida ante el Tribunal.

POLITÉCNICA

El director/Los directores de la tesis autorizan su presentación para:

- ☒ Su admisión a trámite de lectura.
- ☒ La solicitud de la mención "Doctor Europeo".

Madrid, a 28 de octubre de 2015

El director

Fdo: D. José Antonio Cabezas Martínez

La directora

Fdo: Dña. María Teresa de Andrés Domínguez

POLITÉCNICA

INFORME EXPERTO EXTERNO / EXTERNAL REVIEWER REPORT
(ART. 5B DEL REGLAMENTO DE DEFENSA Y EVALUACIÓN DE LA TESIS DOCTORAL)

Nombre del experto/Name of the reviewer

Patricio Hinrichsen

Categoría/Position

Director R&D, Researcher

Universidad a la que pertenece/Departamento, University /Department

Depto Biotecnología y Mejora Genética, Centro de Investigación La Platina, Instituto de Investigaciones Agropecuarias, INIA-Chile

Título de la tesis/Title of the Ph.D. Thesis

Association study of phenology, yield and quality related traits in table grapes using SSR and SNP markers

Nombre del doctorando que presenta la tesis/Name of the Ph.D. candidate

Belkacem Zarouri

Especificar los motivos que avalan la calidad de la tesis mencionada para su defensa:

Specify reasons endorsing the quality of the above-mentioned thesis for its defense:

Se trata de un trabajo que tanto en su enfoque experimental como en volumen de trabajo representan a mi entender un aporte significativo al conocimiento en genética aplicada en especies leñosas, con resultados publicados y por publicar en revistas de corriente principal.

Específicamente, se trata de uno de los primeros estudios genéticos de asociación basado en una colección nuclear de vid (colección per se de alto valor porque incluye genotipos representativos de los diferentes "sortotipos" de *Vitis vinifera* L.), que se basa en un enfoque de tipo GWAS, para lo cual ha sido necesario establecer los niveles genómicos de desequilibrio de ligamiento usando dos tipos de marcadores diferentes, SNPs y SSRs. Los resultados determinaron el número más conveniente de marcadores que sería deseable tener para cubrir el genoma y enfrentar adecuadamente un estudio de tipo GWAS en la especie. Sin embargo, aun cuando la cantidad de marcadores fue bastante inferior a lo deseable teóricamente, se pudo identificar las regiones y grupos de marcadores que se asociaron con varios caracteres de interés en esta especie.

POLITÉCNICA

¿Qué objetivos se han logrado con la tesis presentada?
What are the contributions of the Ph.D. thesis?

Existen muy pocos trabajos en los cuales se haya evaluado el fenómeno de desequilibrio de ligamiento (LD) a una escala genómica en una especie de la importancia de la vid. Con estos antecedentes, se puede enfrentar con mayor conocimiento de causa la aplicación de GWAS en esta especie, lo que aun está en sus comienzos. Para estos fines, se ha identificado una cantidad apreciable de SNPs basado en GBS (sobre 36.000), los que filtrados adecuadamente llevaron a disponer de un set de marcadores uniformemente distribuidos en el genoma de la vid.

Por otra parte, la definición de un panel multiplex de 45 PCRs para abarcar el análisis de cerca de 300 marcadores SSR que cubren de forma homogénea el genoma de la especie en estudio, es también un aporte valioso.

El uso de una colección de 274 genotipos de vid que representan la diversidad de la especie es una plataforma muy potente para estudios de asociación entre fenotipos de interés y sus determinantes genéticos.

El trabajo realizado ha demostrado que esta combinación de SNPs o SSRs lleva a resultados más o menos equivalentes en cuanto a su capacidad para definir índices de LD (aun cuando el número de marcadores SSR es más bien limitado y no permite definir los niveles de LD con la misma potencia que los SNPs basados en GBS).

En resumen, se pudo identificar asociaciones de sets discretos de SSRs con caracteres simples como color de baya y caracteres más complejos como sexo de las flores, con 33 SSRs distribuidos en 10 LGs significativamente asociados. Se evidenció además las fortalezas y puntos débiles de este tipo de estudio (GWAS) para futuros estudios que aborden el análisis de diversos caracteres.

Originalidad del trabajo:
Originality of the work:

Trabajo muy original, puesto que es primera vez que se usa un panel de genotipos de estas dimensiones para estudios de asociación basados en los principios de GWAS/LD, combinando sets significativos de marcadores SSR y SNPs basados en GBS, y todo esto aplicado a una especie de importancia agrícola y de tipo leñoso, como la vid. Además, están apuntando a caracteres de alta importancia para la industria de la uva de mesa, tales como tamaño de baya y firmeza de pulpa.

Metodología usada / hipótesis contrastadas:
Methodology used / hypotheses tested

Como ya se ha indicado más arriba, se ha combinado el análisis genético de una colección nuclear de 274 genotipos con técnicas de alto rendimiento, que incluyeron (i) la secuenciación masiva para la búsqueda de SNPs basado en GBS, y (ii) el genotipado semi-automatizado de ca. 300 SSRs mediante PCR multiplex, con análisis estadísticos de última generación que llevaron a identificar los niveles de LD por cromosoma, cubriendo todo el genoma. Se probaron y contrastaron varios modelos estadísticos para evaluar posibles asociaciones basadas en GWAS. A mi parecer, se trata de un trabajo muy innovativo que sentará las bases para la búsqueda de marcadores y genes que den cuenta de fenómenos biológicos diversos y caracteres de interés productivo.

Se ha demostrado que un enfoque de tipo GWAS es posible de ser aplicado para estudios genéticos de asociación usando una colección de un tamaño y diversidad apropiadas.

Observaciones:
Observations:

Trabajo de excelente nivel que a mi entender amerita el grado de Doctor.

Considera que la tesis anteriormente mencionada es apta para su defensa pública?
In consideration of all the above, is the Ph.D. Thesis ready for its defense?

☒ Si / Yes ☐ No

Firma y fecha Octubre 30, 2015.
Signature and date

POLITÉCNICA

INFORME EXPERTO EXTERNO / EXTERNAL REVIEWER REPORT
(ART. 5B DEL REGLAMENTO DE DEFENSA Y EVALUACIÓN DE LA TESIS DOCTORAL)

Nombre del experto/Name of the reviewer José M Jiménez-Gómez, PhD.
Categoría/Position Group Leader
Universidad a la que pertenece/Departamento, University /Department Max Planck Institute for Plant Breeding Research / Department of Plant Breeding and Genetics
Titulo de la tesis/Title of the Ph.D. Thesis Association study of phenology, yield and quality related traits in table grapes using SSR and SNP markers
Nombre del doctorando que presenta la tesis/Name of the Ph.D. candidate Belkacem Zarouri

Especificar los motivos que avalan la calidad de la tesis mencionada para su defensa:
Specify reasons endorsing the quality of the above-mentioned thesis for its defense:

This work represents an advance in its field by developing tools and methods that have the potential to advance genetic analysis of large germoplasm banks of grape and other crops. It uses the latest technology to characterize sequence variation, effectively genotypes large populations of grapes and performs genome wide association studies in a grape germoplasm collection. This work will be of interest to research groups and companies working with grapes and for a wide range of researchers working in plant genetics.

POLITÉCNICA

¿Qué objetivos se han logrado con la tesis presentada?
What are the contributions of the Ph.D. thesis?

This Thesis contributes a good number of novel methods to the community, ranging from marker development using next generation of sequencing datasets to multiplex PCR for efficient genotyping of large populations. In addition, it contributes important resources for researchers and companies interested in grapes, such as list of polymorphisms and markers for grape accessions hosted at a germplasm bank in Spain. In addition, it has demonstrated the usefulness of genome wide association analysis for grapevine genetic analysis, highlighting and discussing strengths and weaknesses that need to be taken in consideration in future GWAS studies. And, finally, the association analysis between the markers developed and a set of traits of agricultural interest, has allowed to identify a number of associations of potential interest for grape breeders.

Originalidad del trabajo:

Originality of the work:

This work represent an original contribution to the field, since a good number of the methods and datasets described in this Thesis are novel and represent a valuable resource for other research groups and breeders working with grapes, and for plant biologists in general, because they have never been applied to this system and enable effective high-throughput genome-wide analyses of variation in grape.

Metodología usada / hipótesis contrastadas:

Methodology used / hypotheses tested

The main hypothesis tested focused on how the patterns of extend of the linkage disequilibrium on grapevine genome can be used to develop genome wide association studies in this species, and how efficient these studies can be to identify loci involved in the regulation of the expression of relevant traits. A number of methods that represent the forefront of genetics and genomics in other systems have been adapted to the particular necessities of this work. These are interesting protocols for the grape community and table grape breeding companies in particular, and to other researchers in general, since some of the strategies and conclusions can be used to planify similar studies in other species. Among these methods, the most valuable are the determination of sequence variation among grape cultivars by the use of next generation sequencing (genotyping by sequencing) and the development of a multiplexing method to genotype hundreds of individuals for hundreds of markers in a time and cost efficient manner. Other methods worth mentioning that are relevant to modern plant science is one of the first instances of genome-wide association analysis performed in grape

Observaciones:

Observations:

This thesis is a solid piece of work of interest for grapevine researchers and the breeding community, but thanks to its applied nature has potential to reach colleagues working in other systems. The introduction and discussion represent an updated revision of the current literature. The language used is correct.

Considera que la tesis anteriormente mencionada es apta para su defensa pública?

In consideration of all the above, is the Ph.D. Thesis ready for its defense?

X Si / Yes ☐ No

Firma y fecha Octubre 30, 2015.

Signature and date

José M Jiménez-Gómez, PhD.

Cologne, 30th of October 2015.

إلى واللى عزي زين
إلى عذو جتي

ACKNOWLEDGEMENTS

This research has been carried out at the Instituto Madrileño de Investigación y Desarrollo Rural, Agrario y Alimentario (IMIDRA, Alcalá de Henares, Madrid, España) and financed by the Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA, Madrid, España). Thanks to IMIDRA for providing all the resources I used to accomplish this work, and thanks to INIA for supporting me through the doctoral scholarship during four years.

I'd like to express my great knowledge to my supervisors, Dr. José Antonio Cabezas Martínez, and Dr. María Teresa de Andrés Domínguez, for giving me the opportunity to enter in a field of scientific research so precious to me, for guiding me along five large years, and for all what I learned from them.

Alba Vargas, Laura Gaforio and María Aller, you were greatly involved in all the research work performed in the framework of this thesis, from phenotypic assessments in the vineyard and in the lab to the hard work of SSR genotyping. Without your participation, this thesis would have never been accomplished. Find here the expression of my profound gratitude.

Gregorio Muñoz and Felix Cabello, you supervised and participated as well in the phenotypic evaluations of the mapping collection. Gregorio, I can't estimate how many kilometers we walked together through the vineyard evaluating the phenological traits. Felix and Gregorio, Thank you so much for all what you taught me about Ampelography, and especially, for providing a family-like work environment.

Kiki participated energetically in the phenotypic assessments, and so did the trainee students that were as many that I cannot remember their names. Thanks to all of you for the hard and nice moments we spent together.

I'd like to express my thanks to Lance Cadle-Davidson, Bruce Reisch and their research groups for making me so welcome in their lab during the research visit to Cornell University, for the bioinformatics training, for the invaluable discussions about GBS and GWAS and for the guided visits around the area (Niagara Falls especially). Lance, thank you for supervising and collaborating in the GBS work. Thanks also to the Grape Research and Coordination Network (GRCN) for their participation in the financial support of this visit.

Manuel Cantos received me in his lab at IRNAS (Sevilla) to perform a research work that has nothing to do with this thesis, Even so, I abused in spending a lot of time in writing the thesis. Thank you so much Manolo, I hope that this didn't compromise my real task in your lab. This research stay was supported financially by a research grant from the International Office of Vine and Wine (OIV). Thank you OIV.

Finally, and more importantly, I have to thank my beloved wife who suffered the real agony. Without forgetting my “Sumaya”, forgive me sweetie for neglecting you during all this time! To each person who participated directly or indirectly to the accomplishment of this work and here/his name is not mentioned here; Thank you.

ABSTRACT

The advent of cheaper high throughput genotyping technologies and the availability of large germplasm collections encouraged the extension of Genome-Wide Association Studies (GWAS) to crop plants. However, to date these strategies have not yet been tested in grapevine (*Vitis vinifera* L.). Taking advantage of the availability of a large grapevine germplasm collection maintained at the germplasm bank of El Encín (Alcalá de Henares, Madrid, Spain) and the relatively affordable genotyping tools, in this study we proposed to assess the feasibility of GWAS in grapevine, using both a high density Single Nucleotide Polymorphisms (SNPs) dataset and a sparse Simple Sequence Repeats (SSRs) coverage.

A diversity panel of 274 grapevine accessions representing most of the cultivated grapevine gene pool was genotyped using the Genotyping by Sequencing strategy (GBS). A total of 358,454 genome-wide distributed SNPs were identified. After filtering, a reliable dataset of 36,332 SNPs across 242 individuals, with a minimum minor allele frequency (MAF) of 0.01 and an average missingness of 12%, was obtained. On the other hand, we developed a panel of multiplex PCRs (Polymerase Chain Reaction) that allowed to genotype the mapping collection for 264 genome-wide distributed SSR markers using only 45 PCRs and sequencer runs. The genetic information generated with this panel (allele size ranges, diversity parameters, presence of null alleles and ease of scoring) will serve as the basis for the identification of the most suitable loci for future microsatellite-based genetic studies in grapevine.

GWAS relies on Linkage Disequilibrium (LD) to detect genotype-phenotype associations. LD across the grapevine genome was examined using both GBS-SNPs and SSR markers. SNP-based LD decayed to $r^2 = 0.2$ within an average distance of 1 kb, with considerable variation among chromosomes (from 0.03 to 6.79 kb). Therefore, millions of SNPs would be needed to perform efficient GWAS in grapevine. Although the set of SSR markers did not allow to obtain reliable estimates of LD decay throughout the genome because of insufficient marker density, the results suggest that SSR-based LD would not extend much more than SNP-based LD.

SNP-based GWAS across 21 phenotypic and agronomic traits permitted to identify genetic associations at the genome-wide Bonferroni corrected threshold for simply to moderately inherited traits such as berry color, flower sex, muscat flavor, seed content or berry size. Most of these associations overlapped with previously known quantitative trait loci (QTLs), validating hence the effectiveness of this approach, whereas others correspond to new loci that have not been identified before. Putative QTLs were also found for some traits such as ripening time,

cluster architecture or berry shape and firmness. However no associations were detected for others such as flowering time or cluster length and diameter.

The SSR-based association analysis allowed to detect consistent associations at the nominal significance level of 0.01 for more than one year that agree with SNP-based associations found in the present research and/or with QTLs reported in the literature, but also provided some evidence of new associations.

Overall, the present research demonstrates the potential of association studies in grapevine, highlighting and discussing strengths and weaknesses that would need to be taken into consideration in future GWAS projects.

RESUMEN

Las últimas innovaciones en tecnologías de genotipado de alto rendimiento y la disponibilidad de grandes colecciones de germoplasma han impulsado la realización de estudios de asociación conocidos como “Genome-Wide Association Studies” (GWAS) en las especies cultivadas. Sin embargo, hasta la fecha no se han llevado a cabo trabajos que evalúen la utilidad de este tipo de estrategias en la vid (*Vitis vinífera* L.). En este estudio nos propusimos comprobar la utilidad del GWAS para análisis genético en la vid, empleando para ello la colección del banco de germoplasma de vid de El Encín (Alcalá de Henares, Madrid, España) y dos aproximaciones: utilizando por un lado marcadores SNPs (Single Nucleotide Polymorphism) en alta densidad; y por otro una baja cobertura de marcadores microsatélites (SSR; Simple Sequence Repeat).

Una colección de 274 accesiones de vid, que representa gran parte de la diversidad genética existente en la vid cultivada fue genotipada empleando la estrategia de Genotyping by Sequencing (GBS), lo que permitió identificar un total de 358.454 SNPs distribuidos a lo largo de todo el genoma. Tras un proceso de filtrado de datos se obtuvo un set de 36.332 SNPs de alta calidad para 242 individuos, con una frecuencia mínima del alelo minoritario (MAF) de 0,01 y una proporción media de datos perdidos del 12%.

Por otro lado, se desarrolló un panel de PCRs (Polymerase Chain Reaction) multiplex que permitió analizar 264 marcadores SSRs distribuidos a lo largo del genoma de la vid empleando sólo 45 reacciones de PCR. La información genética y alélica generada con este panel (rangos de amplificación, parámetros de diversidad, presencia de alelos nulos, facilidad de genotipado) serán de gran utilidad para seleccionar los marcadores SSRs más adecuados para futuros estudios genéticos en la vid.

Los estudios de tipo GWAS se fundamentan en los patrones de distribución del desequilibrio de ligamiento (LD) a lo largo del genoma. En este trabajo hemos estudiado los patrones de distribución del LD a lo largo del genoma de la vid empleando tanto marcadores de tipo SNPs como microsatélites. Los resultados demostraron que el LD basado en SNPs decae para un r^2 de 0.2, dentro de una distancia media de 1 kb, aunque se observó una variación considerable entre cromosomas (de 0,03 a 6,79 kb). Por lo tanto, para desarrollar GWAS eficientemente en la vid sería necesario utilizar millones de SNPs. Aunque el análisis con SSRs no permitió obtener estimas fiables del decaimiento del LD a lo largo de todo el genoma debido a la insuficiente densidad de marcadores, los resultados indican que el LD basado en SSRs no se extendería mucho más que el LD basado en SNPs.

El GWAS para 21 caracteres fenotípicos e agronómicos basado en SNPs permitió detectar asociaciones con una significación muy alta (corrección de Bonferroni) para caracteres de herencia sencilla o moderadamente compleja como el color de la baya, el sexo de la flor, el sabor

amoscatelado, contenido en semillas o el tamaño de bayas. Muchas de estas asociaciones se localizan en regiones en las que estudios anteriores habían ya identificado QTLs (Quantitative Trait Loci), confirmando así la validez de los resultados obtenidos, mientras que otras corresponden a nuevos QTLs no identificados hasta la fecha. Se detectaron también posibles asociaciones para algunos caracteres, como la fecha de maduración, arquitectura del racimo o la forma y firmeza de la baya. Sin embargo, no se detectaron asociaciones para otros caracteres, como fecha de floración o longitud y diámetro del racimo.

El análisis de asociación basado en SSRs permitió detectar asociaciones consistentes al nivel de significación del 0.01 para más de un año, coincidiéndose con asociaciones encontradas en el análisis con SNPs y/o con QTLs identificados en estudios anteriores. El análisis de asociación basado en SSRs también proporcionó algunas pruebas de nuevos QTLs.

Como conclusión, el trabajo desarrollado demuestra el potencial de los estudios de asociación en la vid, poniendo de manifiesto sus fortalezas, pero también puntos débiles, que deberían ser objeto de especial consideración en futuros proyectos de GWAS.

LIST OF FIGURES

Figure 1. GBS protocol for library preparation and sequencing	11
Figure 2. GBS pipeline work flow.....	12
Figure 3. Schematic representation of the filtering process used to filter the GBS-SNP data.	32
Figure 4. Monthly Cumulative rainfall (a) and average temperatures (b) recorded at the El Encín weather station during the four years of study.....	43
Figure 5. Phenotypic traits distribution within the mapping collection across three growing seasons.....	49
Figure 6. Pairwise Spearman correlations between the studied traits.	52
Figure 7. Distribution of GBS read coverage, genotype quality scores and SNP call rates within the mapping collection	54
Figure 8. Comparison of depth of coverage, proportion of missing values and minor allele frequency between the raw (DS0, left) and the filtered (DS5, right) GBS datasets.	56
Figure 9. Distribution of GBS-SNPs across the 19 grapevine chromosomes	57
Figure 10. Relationship between number of GBS-SNPs per chromosome and chromosome size.	57
Figure 11. Distribution of observed (H_{obs}) and expected (H_{exp}) heterozygocities in the SNP dataset DS5.....	58
Figure 12. Examples of electrophoretic profiles showing the possible amplification of multiple loci in a single genotype by the SSR markers VMC3G7 and UDV-020.	59
Figure 13. Distribution of the 264 genotyped SSR loci along the grapevine genome.	63
Figure 14. Example of PCR profile obtained with the multiplex PCR Mx01.....	64
Figure 15. Distribution of multiplexing level and genetic diversity parameters of the 264 genotyped SSR loci.	65
Figure 16. Population structure of the mapping collection used in this study.....	67
Figure 17. Phenotypic traits variation among the three subpopulations determined by STRUCTURE.....	69
Figure 18. Distribution of kinship coefficients among the 274 individuals of the mapping collection.	70

Figure 19. Distribution of background LD as estimated using a sample of 994,181 unlinked SNP pairs over 242 genotypes.....	71
Figure 20. Distribution of LD within chromosomes as estimated using a set of 25,726 SNPs over 242 genotypes.....	72
Figure 21. Effect of MAF (a), sample size (b) and population structure (c) on SNP-based LD estimates.	73
Figure 22. Distribution of background and linked LD estimated using a set of 250 SSRs over 207 genotypes.	74
Figure 23. Genome-wide association profiles (Manhattan plots) for five traits obtained using four analysis models.	81
Figure 24. Chromosome-wide association patterns showing the colocalization of association peaks with major loci for berry color, flower sex, muscat flavor and seed content	82
Figure 25. Quantile-Quantil (QQ) plots for four analysis models across four traits.....	83
Figure 26. Genome-wide association profiles obtained using the Q+K model for nine grape traits across three growing seasons.....	91
Figure 27. Localization of SSR gaps regarding the distribution of gene density along the grapevine chromosomes.	109
Figure 28. Distances of LD decay below $r^2 = 0.1$ as estimated using SNP vs. SSR markers	118
Figure 29. Association profiles of SNPs (a, c) and SSRs (b, d) around the candidate genes responsible for berry color and muscat flavor	134

LIST OF FIGURES IN THE ANNEXES

Figure A.1. Relationship between SNP and gene density distributions in the grapevine genome.....	181
Figure A.2. LD decay plots across the grapevine genome as estimated using a set of 25,726 SNPs over 242 genotypes.....	200
Figure A.3. LD decay plots across the grapevine genome as estimated using a set of 250 SSRs over 207 genotypes..	202
Figure A.4. Genome-wide association profiles obtained using four analysis models for seed content traits in 2010 and 2012	206
Figure A.5. Genome-wide association profiles obtained using the Q+K model for seven grape traits across three growing seasons.....	209

Figure A.6. GLM model-based association profiles for 16 grape traits across three growing seasons.....	215
---	-----

LIST OF TABLES

Table 1. List of the evaluated traits with their abbreviated names, OIV codes and summarized descriptions	29
Table 2. Pairwise Spearman correlations between the three years of study for each trait.	50
Table 3. Pattern of decline of the SNP dataset dimensions after filtering by combination of minimum call rate per individuals and by sites.....	54
Table 4. Microsatellite markers included in each multiplex PCR.	60
Table 5. Extent of SNP-based LD across the grapevine genome and estimated number of SNPs for whole-genome covering.....	71
Table 6. Estimated distances of SSR-based LD decay across the grapevine genome and estimated number of SSRs for whole-genome covering.	75
Table 7. Variation of the number of significantly associated SNPs according to the analysis model and to their position regarding the major loci governing the trait.	84
Table 8. Main features of genetic associations identified for 13 grape traits across three growing seasons using a mapping collection of 242 accessions and a set of 36,332 SNPs.	92
Table 9. Genome-wide associated SSRs to 18 grape traits for three growing seasons in comparison with SNP association results and previous QTL studies.	96
Table 10. Descriptive statistics of quantitative traits as evaluated in our mapping collection vs. in a collection of 322 table grape accessions (GLOBAL322) by Vargas (2009).	104
Table 11. Comparison between SNP and SSR-based LD at different categorical physical distances.....	118

LIST OF TABLES IN THE ANNEXES

Table A.1. List of the 274 grapevine accessions included in the mapping collection and their main features	167
Table A.2. List of trio pedigrees included in the mapping collection	176
Table A.3. Mendelian inheritance checking of SNP alleles in trio pedigrees	177
Table A.4. Chromosome-wise summary statistics of the filtered SNP dataset DS5	178
Table A.5. Main features of the developed SSR multiplex PCRs	182
Table A.6. Thermocycling programs optimized to run the designed multiplex PCRs.....	188

Table A.7. Quality of amplification profiles, data checking and diversity parameters of the analyzed SSR markers and their utilization in the study	189
Table A.8. Positions and <i>P</i> values of all the significant trait-SNP associations identified on chr1 using Q+K model	216
Table A.9. Positions and <i>P</i> values of all the significant trait-SNP associations identified on chr2 using Q+K model	217
Table A.10. Positions and <i>P</i> values of all the significant trait-SNP associations identified on chr5 using Q+K model	219
Table A.11. Positions and <i>P</i> values of all the significant trait-SNP associations identified on chr8 using Q+K model	219
Table A.12. Positions and <i>P</i> values of all the significant trait-SNP associations identified on chr17 using Q+K model.	220
Table A.13. Positions and <i>P</i> values of all the significant trait-SNP associations identified on chr18 using Q+K model	220
Table A.14. List of new unstable genetic associations detected by SSR markers in comparison with SNP association results.	221

INDEX

ACKNOWLEDGEMENTS.....	xvii
ABSTRACT	xxi
RESUMEN	xxiii
LIST OF FIGURES	xxvii
LIST OF FIGURES IN THE ANNEXES	xxviii
LIST OF TABLES.....	xxx
LIST OF TABLES IN THE ANNEXES	xxxi
1. INTRODUCTION	3
1.1. GRAPEVINE: TAXONOMY, ORIGIN AND DOMESTICATION	3
1.2. THE GRAPEVINE GENOME	4
1.3. GENETIC DIVERSITY OF GRAPEVINE: CULTIVARS, CLONES AND GERMPLASM CONSERVATION.....	5
1.4. GRAPE PRODUCTION: CLASIFICATION AND ECONOMIC IMPORTANCE	6
1.5. GRAPE BREEDING	7
1.6. MOLECULAR MARKERS FOR GENETIC ANALYSIS IN GRAPEVINE.....	8
1.6.1. Microsatellite markers	9
1.6.2. Single nucleotide polymorphism markers	10
1.7. STUDY OF THE GENETIC DETERMINISM OF PHENOTYPIC TRAITS	13
1.7.1. Linkage mapping.....	13
1.7.2. Association mapping	14
1.7.2.1. Linkage Disequilibrium.....	15
1.7.2.2. Methodology of Association Mapping	16
1.7.2.3. Current status of association mapping in plants	19
2. OBJECTIVES.....	23
3. MATERIAL AND METHODS.....	27
3.1. PLANT MATERIAL	27
3.2. PHENOTYPIC CHARACTERIZATION	27
3.2.1. Experimental conditions.....	27
3.2.2. Traits evaluation	28
3.2.3. Statistical analysis of phenotypic data	28
3.3. SNP GENOTYPING USING GENOTYPING BY SEQUENCING	30
3.3.1. DNA extraction	30
3.3.2. Library preparation and sequencing.....	30

3.3.3. SNP calling.....	30
3.3.4. Data filtering and error checking.....	31
3.4. SSR GENOTYPING.....	32
3.4.1. DNA extraction	32
3.4.2. Markers selection	33
3.4.3. Multiplex PCRs design and genotyping.....	33
3.4.3.1. Amplification tests.....	33
3.4.3.2. Multiplex PCRs optimization	34
3.4.4. Data checking.....	35
3.4.5. SSR data analysis	35
3.5. POPULATION STRUCTURE AND RELATEDNESS ANALYSIS.....	36
3.6. LINKAGE DISEQUILIBRIUM ANALYSIS	36
3.6.1. SNP-based LD.....	36
3.6.2. SSR-based LD	37
3.7. GENOME-WIDE ASSOCIATION ANALYSIS	38
3.7.1. SNP-based association analysis and model testing	38
3.7.2. SSR-based association analyses	39
4. RESULTS.....	43
4.1. PHENOTYPIC DATA.....	43
4.1.1. Traits distribution	44
4.1.2. Phenotypic correlations between years and between traits	50
4.2. GENOTYPIC DATA.....	53
4.2.1. SNP genotyping.....	53
4.2.1.1. Library sequencing, SNP calling and filtering	53
4.2.1.2. SNP density and genetic diversity	55
4.2.2. SSR genotyping.....	58
4.2.2.1. Amplification tests.....	58
4.2.2.2. Multiplex PCRs design, optimization and validation	59
4.2.2.3. Genetic diversity and quality of the studied SSR markers	64
4.3. POPULATION STRUCTURE AND RELATEDNESS.....	66
4.4. LINKAGE DISEQUILIBRIUM ANALYSIS	70
4.4.1. LD analysis using SNP markers.....	70
4.4.1.1. Background and linked LD distribution	70
4.4.1.2. Effect of MAF, sample size and population structure on the extent of LD.....	72

4.4.2. LD analysis using SSR markers	74
4.5. GENOME-WIDE ASSOCIATION ANALYSIS	76
4.5.1. GWAS using SNP markers	76
4.5.1.1. Comparative assessment of association analysis methods	76
4.5.1.2. GWAS based on the mixed model Q+K.	84
4.5.2. Association analysis using SSR markers	93
5. DISCUSSION.....	101
5.1. PHENOTYPIC DATA.....	101
5.2. GENOTYPING BY SEQUENCING: ADVANTAGES AND LIMITATIONS.....	104
5.3. MICROSATELLITE MULTIPLEX PCRS: A TOOL FOR WHOLE-GENOME GENOTYPING OF GRAPEVINE	107
5.3.1. Critical points in the process of multiplex PCRs development.....	107
5.3.2. Potential applications of the developed multiplex PCRs	111
5.3.3. The problem of Null alleles.....	112
5.4. POPULATION STRUCTURE	113
5.5. PATTERNS OF LINKAGE DISEQUILIBRIUM IN GRAPEVINE.....	114
5.5.1. LD assessed using SNP markers	114
5.5.1.1. Effect of MAF, sample size and population structure on LD estimates.....	115
5.5.2. LD assessed using SSR markers	116
5.5.3. Comparison of SNP-based vs. SSR-based LD.....	117
5.5.4. Implications for GWAS	119
5.6. GENOME-WIDE ASSOCIATION ANALYSIS	121
5.6.1. GWAS using SNP markers	121
5.6.1.1. Association analysis models: efficiency in controlling for false positives.....	121
5.6.1.2. Detection power and resolution of GWAS.....	122
5.6.1.3. New genetic associations.....	126
5.6.2. SSR-based association analysis.....	128
5.6.3. Comparison of high density SNPs vs. low density SSRs for GWAS	130
6. CONCLUSIONS	137
7. REFERENCES	141
ANNEXES.....	167

INTRODUCTION

1. INTRODUCTION

1.1. GRAPEVINE: TAXONOMY, ORIGIN AND DOMESTICATION

Grapevine (*Vitis vinifera* L.) is a dicotyledonous species that belongs to the family *Vitaceae*, whose members are collectively known as grapevines (Keller, 2015). Grapevines are shrubs or woody lianas that climb by means of their leaf-opposed tendrils. Almost all the cultivated grapes in the world belong to the genus *Vitis* ($2n = 38$), and the rest belong to the genus *Muscadinia* ($2n = 40$) famous in Southern United States (Winkler *et al.*, 1974). There are about 70 known *Vitis* species, most of them originating from the northern hemisphere, and particularly abundant in Asia and North America (Alleweldt and Possingham, 1988). While *Vitis vinifera* L., is the principal species of the genus *Vitis* that gave rise to most of the cultivated grapes grown today, the American species of *Vitis* have contributed to viticulture mainly as rootstocks, but also as progenitors of the “hybrid direct producers”.

Vitis vinifera L. is subdivided into two subspecies: the cultivated form, *V. vinifera* ssp. *sativa* (or *vinifera*) and the wild form, *V. vinifera* ssp. *sylvestris* (or *silvestris*). The former is thought to be domesticated from the latter in the Near East about 7 to 8 thousand of years ago (Levadoux, 1956; Hardie, 2000; McGovern, 2003; Grassi *et al.*, 2006; This *et al.*, 2006; Myles *et al.*, 2011). Although other molecular analysis suggested the Western Mediterranean as an independent center of domestication (Grassi *et al.*, 2003; Arroyo-García *et al.*, 2006). The domestication process was marked by a considerable increase in cluster and berry size and the change of the floral structure from unisexual flowers in the wild ancestor to perfect flowers in cultivated grapes. Grapevine domestication was also accompanied by the adoption of vegetative propagation (by cuttings) instead of sexual reproduction (by seeds). Following domestication, thousands of cultivars derived from spontaneous or controlled crosses, but also from somatic variation (Torregrosa *et al.*, 2011) have been selected and spread by vegetative propagation throughout the world, from temperate to tropical climates (Bouquet, 2011). Despite the large period of time separating the currently cultivated grapes from their ancestor, and the natural disasters caused by the Phylloxera and mildews at the end of the 19th century, the genetic diversity of both compartments seem to be quite similar, probably as a consequence of a weak domestication bottleneck and the vegetative propagation that helped to maintain the initial diversity (Myles *et al.*, 2011).

According to the eco-geographic based classification proposed by Negrul (1946), cultivated grapes can be subdivided into three ecotypes (or *proles*): the *occidentalis* of Central and Western Europe, the *orientalis* extending from Central Asia to Azerbaijan and Armenia and the *pontica* of

Eastern Europe, southern Russia, Turkey, and Georgia. Modern grapevine clustering based on molecular data and statistical modeling agree substantially with this classification (Aradhya *et al.*, 2003; Bacilieri *et al.*, 2013; Emanuelli *et al.*, 2013). While *occidentalis* and *pontica* groups are composed mostly of wine grapes, *orientalis* grapes are generally used for fresh fruit consumption (Bacilieri *et al.*, 2013).

1.2. THE GRAPEVINE GENOME

The highly heterozygous diploid genome of grapevine is organized in 19 chromosome pairs. Its size, initially estimated at 475 Mb using flow cytometry (Lodhi and Reisch, 1995), is comparable to that of poplar (485 Mb; Tuskan *et al.*, 2006) and rice (389 Mb; International Rice Genome Sequencing Project, 2005), and four times larger than that of *Arabidopsis* (125 Mb; The Arabidopsis Genome Initiative, 2000). This has encouraged the grapevine research community to undertake several genome sequencing initiatives. The first one was launched by the French–Italian Public Consortium (Jaillon *et al.*, 2007) that selected the nearly homozygous inbred line PN40024 to produce what is now considered as the reference sequence of the grapevine genome. The selection of a nearly homozygote line was decisive in obtaining a high quality draft sequence (8x) that was published in 2007. Two years later, a higher coverage assembly (12x) was released (<http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/>); and more recently, further improvements regarding sequence ordering and orientation have been added (<https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences>; Canaguier *et al.*, 2014). In parallel, another union of Italian institutions sequenced the genome of the highly heterozygote elite cultivar, Pinot Noir (Velasco *et al.*, 2007). The genome sizes reported upon sequencing were concordant between the two projects (487 and 504 respectively) and agreed with the initial flow cytometry-based estimation. Moreover, with the objective of deciphering the grapevine genome polymorphism, several *Vitis* genotypes have been partially re-sequenced (Myles *et al.*, 2010; Le Paslier *et al.*, 2013), and other specific cultivars such Tannat (Da Silva *et al.*, 2013) or Sultanina (Di Genova *et al.*, 2014) have been completely re-sequenced.

Genome annotation revealed that the 8x sequence contains 30,434 protein-coding genes and that 41.4% of the grapevine genome is composed of repetitive/transposable elements unevenly distributed along the chromosomes, defining hence gene-poor and gene-rich regions (Jaillon *et al.*, 2007). As for genome assembly, genome prediction has been updated several times (Grimplet *et al.*, 2012), being the V2 at CRIBI (<http://genomes.cribi.unipd.it/grape/>) the latest version available so far, containing 31,922 coding genes (Vitulo *et al.*, 2014).

Paralogous regions distribution over the grapevine genome suggested that the present-day grapevine haploid genome had been originated from the contribution of three ancestral genomes.

Moreover, comparative genomics analysis showed that the hypothetical hexaploid ancestor of grapevine is shared with *Arabidopsis* and poplar (dicotyledonous), but not with rice (monocotyledonous) (Jaillon *et al.*, 2007).

1.3. GENETIC DIVERSITY OF GRAPEVINE: CULTIVARS, CLONES AND GERMPLASM CONSERVATION

The number of grape cultivars existing today is estimated to be around 5,000 (This *et al.*, 2006), being many of them related to each other, because most of them are the result of continuous selection and vegetative propagation over many centuries of grapevines derived from spontaneous crosses and/or somatic mutations (Keller, 2015). Modern grapevine cultivars derived from planned breeding activities did not appear until the end of the 19 century; when phylloxera and fungal diseases began to decimate the vineyards of Europe (Alleweldt and Possingham, 1988).

Unlike annual crops, in perennial fruit crops such as grapevine, somatic mutations constitute an important source of genetic variation, because they have more chance to happen and to accumulate along the large lifetime of the plant (Torregrosa *et al.*, 2011). Additionally, when a somatic variant (or sport) appear in a given plant, it could be stabilized and propagated via vegetative multiplication. For instance, somatic variation has been almost the unique source of genetic variation exploited by breeders for clonal selection within classical wine cultivars (This *et al.*, 2006; Ibáñez *et al.*, 2015). In viticulture, the word “clone” refers to the vegetative descendant of a vine selected for its indisputable identity, its phenotypic characteristics and health status (Ibáñez *et al.*, 2015). Paradoxically, if those phenotypic characteristics are sufficiently distinct, the clone may come to be called cultivar (Keller, 2015); the case, for example, of the berry skin color mutants of Pinot: Pinot noir, Pinot gris, and Pinot blanc (Regner *et al.*, 2000). Therefore, there is no clear-cut distinction between clones and cultivars. Moreover, the large extension of cultivars growing in space and time generated a large confusion in their denomination. For instance, many cultivars are known by several synonyms (different names for the same cultivar) or homonyms (identical name for different cultivars). As an example, Afus Ali, Dattier de Beyrouth, Regina, Roseti, Rozaki, jointly with other 215 names correspond to the same cultivar as reported in the *Vitis* International Variety Catalogue (<http://www.vivc.de/>).

Several millennia of cultivation and repeated selection of spontaneous mutants and natural as well as controlled crosses in different regions of the world resulted in a vast range of cultivated grapevines. However, at the end of the nineteenth century, new pathogens such as downy and powdery mildews and phylloxera spread from North America to Europe, causing the destruction of many European (and surrounding regions) vineyards and radically changing the diversity of

both cultivated and wild grapes. The currently existing diversity is most likely a pale reflection of what existed before the arrival of phylloxera. European viticulture escaped extinction thanks to the introduction of pathogen resistant species of the genus *Vitis* originating in Central America that were used as rootstocks. Moreover, for decades, the globalization of the wine market and, to a lesser extent, table grape market has favored the expansion of certain elite varieties in the detriment of ancient local ones, leading to another reduction in grapevine diversity. This continuous erosion of genetic diversity raised the concern of its conservation. Hence, several grape germplasm collections have been established in different grape-growing countries (<http://www.vivc.de/>), being the French collection at Vassal (INRA, Montpellier) with more than 7,000 accessions the largest one world-wide (<http://www1.montpellier.inra.fr/vassal/>). The Spanish collection of El Encín (<http://www.madrid.org/coleccionvidencin/> Alcalá de Henares, Madrid, Spain), comes at the second place with 3.696 accessions corresponding to: 1,757 wine grapes, 489 table grapes, 848 rootstocks, 423 wild grapes, 107 hybrids direct producers and 75 from other species of the genus *Vitis* (data accessed in August 2015).

1.4. GRAPE PRODUCTION: CLASIFICATION AND ECONOMIC IMPORTANCE

The usual classification of grape cultivars is based on the destination of their final products. In this sense, grapevine cultivars are generally subdivided into three classes, although some cultivars are used for various purposes: wine grapes, table grapes and raisin grapes. The area dedicated to vineyards is increasing by about 2% per year. According to the OIV statistics (OIV, 2011), the total area under vineyards in the world rounded 7.52 million ha producing more than 69 million tons of grapes, from which 62.30% processed into wine (including juice and must, 3.60% approx.), 30.56% consumed fresh and 1.79% dried for raisins.

- **Wine grapes:** Characterized by small to medium sized and compact clusters, with small and seeded berries with thick skin and high juice content (vs. pulp). On average, wine grapes have more acidity and more sugar content than table grapes. Europe with 176.40 million hectoliters (Mhl) is the leading continent of wine production, followed by North and South America (51 Mhl), Asia (15 Mhl), Australia and New Zealand (13. 35 Mhl) and Africa (11 Mhl) (OIV, 2011). Italy, Spain and France were the principal wine grape producers in the EU. In 2011, these countries produced alone 126.93 Mhl of wine.
- **Table grapes:** Characterized by large and loose clusters with large, fleshy and crunchy berries, sometimes of particular flavors. They represent in average 25 to 30% of grape production in the world, being this proportion highly variable between countries. For example, while table grape production did not exceed 5% (0.28 million tons) in Spain, it represented 33.5% (0.85 million tons) in Chile and 60.7% (2.58 million tons) in Turkey

(OIV, 2011). The world production of table grapes shows a remarkable upward trend, increasing by 11% between 2007 and 2011. This increment is mainly accounted for by the Asian continent whose production in 2011 represented 59.2% of world production (half of which in China alone), ahead of Europe whose share was 17.3%. North and South America and Africa represented respectively 12.8% and 10% of global production (OIV, 2011). Despite the perishability of the fruit and the high cost of transportation, table grapes support a significant international trade of nearly 3 million tons with imports mainly into the USA, Germany and UK and exports mainly from Chile and Italy (Bouquet, 2011).

- **Raisin grapes:** Also called dried grapes, are predominantly seedless, either parthenocarpic such as Black Corinth or stenospermocarpic such as Sultanina. Nevertheless, some producers produce raisins using seeded grapes that are initially intended to be consumed fresh. In 2011, the global production of dried grapes reached 1.24 million tons, about 1.79% of the total grape production (OIV, 2011). As for table grapes, Asia is the leading continent of raisins production with the 52.9% of world production, followed by North and South America, which account for 35.9%, and Europe that represents 6.5% of the total. The top producers are Turkey (0.41 million tons), USA (0.34 million tons) and Iran (0.15 million tons).

1.5. GRAPE BREEDING

Although, from an economic point of view, table grape market is less important than wine market, it is much more dynamic regarding the cultivated cultivars. The range of wine grape cultivars has remained quite rigid for decades, mainly because wine producers in a particular locality have to use determined cultivars to achieve a certain label of quality. Consequently, breeding efforts for wine grapes have generally focused on clonal selection within the old cultivars (Ibáñez *et al.*, 2015). By contrast, such restrictions do not apply for table grape market, and most cultivars being grown today are different from the cultivated ones 30 years ago. This is why breeding programs are much more frequent and active in the table grape sector, and new cultivars are being continuously released. Traits of interest currently targeted for table grape improvement include seedlessness, muscat flavor, large and firm berries, loose clusters, early and late ripening and long shelf life. Moreover, there is an increased interest in developing cultivars capable to face the future challenges related to global warming and its effect on host-pathogen interactions.

On the other hand, many new cultivars are protected by Intellectual Property Rights, and therefore, growers have to pay royalties for their use. This scenario incited several countries

(through public institutions) and private companies to undertake their own breeding programs. In Spain for example, table grape breeding programs were initiated by the public institutions, IFAPA (Jerez) in the 1970s, and IMIDA (Murcia) in the 1990s, and recently the private company ITUM has joined the campaign as well (Ibáñez *et al.*, 2015).

The breeding of grapevines however present considerable difficulties, related mainly to: limitations to achieve large breeding populations due mostly to the complicated emasculation-pollination process; long biological cycle (the production phase is not attained until a minimum of three years); high cost and intensive labor required for growing, maintaining and assessing large breeding populations; and lack of knowledge about the genetics and the inheritance of the traits of interest (Eibach and Töpfer, 2015).

Nowadays, we are in a position to overcome at least part of these limitations. In fact, the adoption of molecular techniques since the early 1990s by grape geneticists has contributed largely to improve our understanding of the genetic determinism of some traits of interest (Welter *et al.*, 2011). Additionally, through marker assisted selection (MAS), molecular markers linked to major genes underlying these traits can be used since the first year of planting to discard less interesting seedlings. For instance, this has become a reality not only for disease resistance (Eibach *et al.*, 2007; Riaz *et al.*, 2008), but also for quality related traits such as seedlessness (Karaagac *et al.*, 2012; Bergamini *et al.*, 2013) and muscat flavor (Emanuelli *et al.*, 2014). On the other hand, whole-genome distributed markers can accelerate the elimination of the linkage drag and recover the recurrent parent genetic background when wild species are involved in the cross (Herzog *et al.*, 2013). Moreover, grapevine accessions conserved in germplasm collections are being fully characterized using molecular markers. This can help in the selection of adequate parental genotypes to be used for a given breeding objective (Töpfer *et al.*, 2011).

More recently, the availability of the reference genome sequence and high throughput genotyping technologies has facilitated the access to the whole-genome polymorphism. This offers the possibility to push further both the understanding of traits genetics and the efficiency of conventional breeding programs through the use of the strategies known as “genome-wide association mapping” and “genomic selection” (Fodor *et al.*, 2014).

1.6. MOLECULAR MARKERS FOR GENETIC ANALYSIS IN GRAPEVINE

Traditionally, the study of the botanical description, identification, and classification of grapevines have been based on a set of morphological characteristics in what is known as ampelography (Galet, 1998). In the last decades, these studies have been complemented by using different types of molecular markers from the initial isozymes, through RFLPs (Restriction

Fragment Length Polymorphism), RAPDs (Random Amplified Polymorphic DNA), AFLPs (Amplified Fragment Length Polymorphism), SSRs (Simple Sequence Repeat), and lastly to SNPs (Single Nucleotide Polymorphism) (Cipriani *et al.*, 2011). However, the great revolution in grape molecular genetics did not take place until the advent of SSR markers. Here we will only focus on SSRs and SNPs, the two types of markers used in this study.

1.6.1. Microsatellite markers

Microsatellite markers, also known as Simple Sequence Repeats (SSRs), are short tandemly repeated sequences, with repeat unit sizes ranging from one to six bases (Goldstein and Schlötterer, 1999). During the last decades, SSRs have been widely used for different genetic studies in plants (Kalia *et al.*, 2010). Their popularity is related to their monolocus and multiallelic features, codominant inheritance and high reproducibility (Rafalski and Tingey, 1993; Powell *et al.*, 1996; Schlötterer, 2004). In addition, they are hypervariable, easily transferable, and require basic equipment usually accessible for small to medium-sized laboratories. In spite of these advantages, microsatellite markers present some limitations; mainly the high cost and time-consume per individual SSR assay (Guichoux *et al.*, 2011). In order to increase cost and labor efficiency of SSR analysis, multiplex PCR approaches combined with fluorescence-based analysis and semi-automated allele calling have been developed (Butler, 2005b; Missiaggia and Grattapaglia, 2006; Blacket *et al.*, 2012). These strategies allow amplifying simultaneously more than one locus in a single reaction using multiple primer pairs. Multiplex PCR protocols have been developed and successfully applied for different uses in several organisms (Bryda and Riley, 2008; Jewell *et al.*, 2010; Sayama *et al.*, 2011; Liu and Wu, 2012; Patil *et al.*, 2012; Postolache *et al.*, 2013; Drašnarová *et al.*, 2014). Multiplex PCR design requires a priori knowledge of the range of allelic sizes for each marker in order to avoid amplicons overlapping, and the optimization of PCR conditions to obtain a balanced amplification of all the targeted fragments (Butler, 2005a). Another limitation for the use of microsatellites is the eventual presence of null alleles that challenges particularly population and linkage disequilibrium-based studies (Callen *et al.*, 1993). A microsatellite null allele is any allele that cannot be amplified via PCR, usually due to mutations in the primer binding sites (Dakin and Avise, 2004), and may lead to misinterpreting the results by: (1) considering a PCR failure when the null allele is present in homozygous state; (2) classifying a heterozygote individual as homozygote when the null allele is in heterozygous state (Wagner *et al.*, 2006).

Microsatellite repeats are abundant and diverse in the grapevine genome (Thomas *et al.*, 1993). This has allowed the identification of hundreds of them and the design of primers for their

analysis throughout the last two decades (Thomas and Scott, 1993; Bowers *et al.*, 1996; Bowers *et al.*, 1999; Sefc *et al.*, 1999; Scott *et al.*, 2000; Lefort *et al.*, 2002; Decroocq *et al.*, 2003; Arroyo-García and Martínez-Zapater, 2004; Di Gaspero *et al.*, 2005; Merdinoglu *et al.*, 2005; Cipriani *et al.*, 2008; Huang *et al.*, 2011). Currently, up to 1,079 *V. vinifera* SSR probes can be found at the NCBI database (<http://www.ncbi.nlm.nih.gov/probe>). Microsatellite markers have been used in grape for many purposes, including: germplasm characterization and pedigree reconstruction (Sefc *et al.*, 2009); construction of linkage maps (Cipriani *et al.*, 2011); identification and mapping of QTLs (Welter *et al.*, 2011) and marker assisted selection (Töpfer *et al.*, 2011).

1.6.2. Single nucleotide polymorphism markers

Single Nucleotide Polymorphisms (SNPs) are the most common type of genetic variation in any living organism. Each SNP represents a single genomic position (one nucleotide) at which two or more alternative bases occur (Wang *et al.*, 1998). Due to their abundance and amenability to automation, SNPs are considered as the markers of choice for genetic mapping and breeding (Rafalski, 2002b). The first generation of large scale SNP genotyping methods were based on two separate steps: (1) SNP discovery: polymorphic sites are identified by re-sequencing a set of genotypes or by screening pre-existing sequence libraries (Rafalski, 2002a); (2) SNP genotyping: a subset of the identified SNPs is selected (based on their position and other criteria such as allele frequency) to develop a genotyping platform (microarray) capable of assaying all the SNPs in parallel (Syvänen, 2001). In grapevine, three such platforms have been developed: several small SNPlex chips of 48 SNPs (Lijavetzky *et al.*, 2007), a Vitis9KSNP array (Myles *et al.*, 2010) and the latest one GRAPERESQ 18K chip that allows to genotype 18,071 SNPs, of which 14,817 are polymorphic in *V. vinifera* (Le Paslier *et al.*, 2013). However, these platforms suffer from two major limitations: firstly, the relatively small number of polymorphic sites that can be genotyped; and secondly, it is challenging to obtain reliable genotype calls in grapevine using array-based technologies due to the high diversity displayed by this species, which complicates the finding of invariant primer-binding regions (Myles *et al.*, 2010). Recently, with the advent of next generation sequencing (NGS) technologies (Metzker, 2010), several sequence-based strategies allowing the simultaneous SNP discovery and genotyping have emerged as alternatives to array-based approaches (Poland and Rife, 2012). Among them, the GBS (Genotyping by Sequencing) is a cost-effective protocol optimized specifically for high diversity species (Elshire *et al.*, 2011). The GBS approach is based on the multiplex-NGS of reduced representation libraries (RRL) generated by random DNA shearing using methylation sensitive

restriction enzymes. Multiplex sequencing is achieved by tagging randomly sheared DNA fragments from different samples with unique short DNA sequences (barcodes) and pooling samples into a single sequencing channel. The GBS strategy is based on two main steps: (1) DNA library preparation and sequencing (Elshire *et al.*, 2011); (2) bioinformatic analyses using the GBS pipeline for SNP calling (Glaubitz *et al.*, 2014). The two steps are briefly explained in Fig. 1 and Fig. 2 respectively. GBS has been used to genotype large samples of plants as well as animals (Poland *et al.*, 2012; De Donato *et al.*, 2013; Romay *et al.*, 2013; Ward *et al.*, 2013; Jarquín *et al.*, 2014; Liu *et al.*, 2014; Rabbi *et al.*, 2014; Schilling *et al.*, 2014), but in grapevine there is only a single report, in which GBS was used to genotype an F₁ mapping population of 85 individuals (Barba *et al.*, 2014).

Figure 1. GBS protocol for library preparation and sequencing (reproduced from Myles, 2013)

In **step 1**, genome complexity is reduced by digesting each sample with the restriction enzyme ApeKI to generate a reduced representation library (RRL). Because only digested fragments within a certain size range (approximately 100–400 bp) will be successfully sequenced using Illumina sequencing technology, this results in DNA sequence reads being generated from only a small fraction of the genome. In **step 2**, sample-specific barcode adaptors are ligated to the sticky ends produced by ApeKI. The Illumina standard forward and reverse adaptors (sequencing primers + flowcell binding oligonucleotides) are also ligated during this step. **Step 3** involves pooling all of the samples together into a single tube. After pooling, the remaining steps of the Illumina standard library preparation procedure are performed and DNA sequencing takes place on one or multiple channels of a single lane of an Illumina next-generation sequencing (NGS) machine.

Figure 2. GBS pipeline work flow (reproduced from Glaubitz *et al.*, 2014)

(A) Barcoded sequence reads are processed and collapsed into a set of unique sequence tags, with one TagCounts file produced per input FASTQ file. The separate TagCounts files are then merged to form a “master” TagCounts file, which retains only those tags present at or above an experiment-wide minimum count. This master tag list is then aligned to the reference genome and a TagsOnPhysicalMap (TOPM) file is generated, containing the genomic position of each tag with a unique, best alignment. (B) The barcode information in the original FASTQ files is then used to determine the number of times each tag in the master tag list is observed in each sample (“taxon”) and these counts are stored in a TagsByTaxa (TBT) file. (C) The information recorded in the TOPM and TBT is then used to discover SNPs at each “TagLocus” (set of tags with the same genomic position), and optionally filter the SNPs based on a set of criteria such as the proportion of taxa covered by the TagLocus and minor allele frequency.

1.7. STUDY OF THE GENETIC DETERMINISM OF PHENOTYPIC TRAITS

Finding the genetic determinism of phenotypic variation is one of the principal goals of many plant geneticists. Location of specific genes and loci on a genetic map can greatly accelerate the breeding process by allowing the use of markers to screen and select individuals in early growth stages. This has been usually achieved using the key tool known as linkage mapping. In order to circumvent some limitations of linkage mapping, a new approach known as association mapping was introduced (Kruglyak, 1999). Association mapping has been used extensively to dissect human diseases (Slatkin, 2008), and in the last few years it has received an increased attention of plant geneticists (Nordborg and Weigel, 2008; Huang and Han, 2014), triggered principally by the advent of cheaper high throughput genotyping technologies and powerful computational methods and a high interest to exploit the natural diversity (Zhu *et al.*, 2008).

1.7.1. Linkage mapping

Traditionally, the first step toward characterizing and cloning genes was the construction of a genetic linkage map derived from a family segregating for the trait of interest. Linkage maps have been utilized for more than a decade, for identifying chromosomal regions that contain genes controlling simple traits (controlled by a single gene) and quantitative traits using QTL analysis. These maps can be generated using different types of molecular marker (RAPD, SSR, AFLP, SNPs...). Genes or markers that are close to each other or tightly-linked will be transmitted together from parents to progeny more frequently than genes or markers that are located further apart. In a segregating population, there is a mixture of parental and recombinant genotypes. The frequency of recombinant genotypes can be used to calculate recombination fractions, which may be used to infer the genetic distance between markers. By analyzing the segregation of markers, the relative order and distances between markers can be determined: the lower the frequency of recombination between two markers, the closer they are situated on a chromosome; and, conversely, the higher the frequency of recombination between two markers, the further away they are situated on a chromosome.

The most commonly used populations for molecular mapping such backcross, F_2 , recombinant inbred lines or double haploids are usually derived from crosses between pure inbred lines. Grapevine, however, is highly heterozygous, hindering the generation of pure parental lines because of a strong inbreeding depression. This, together with the long time between generations, hampers the use of linkage mapping approaches that are based on classical mapping populations. Instead, a special approach adapted for highly heterozygous species, named “two-way pseudo-testcross strategy”, is usually used (Grattapaglia *et al.*, 1995). This strategy is based

on segregant F_1 progenies derived from controlled crosses between two individuals with contrasting phenotypes for the trait of interest. The progeny is characterized using a set of molecular markers, and the resulting molecular dataset is used to construct a specific genetic map for each parent that can be then integrated into a consensus map. Subsequently, associations between the phenotypic variation of the trait and the allelic variation of the mapped markers are searched using different statistical methods (Doerge, 2002).

The two-way pseudo-testcross strategy however suffers from two major limitations: i) the F_1 progeny only undergo one generation of recombination events, limiting hence the mapping resolution of the study; ii) the only QTLs that can be detected in the mapping population are those that are in heterozygosis in the parents of the cross (Korte and Farlow, 2013; Huang and Han, 2014). Additionally, depending on the targeted trait, several years are generally required before the progeny enters in the production phase and can be phenotyped accurately.

The two-way pseudo-testcross strategy has been widely used through the last two decades to study the genetic determinism of several traits of interest in grapevine, including: pathogen resistance (Fischer *et al.*, 2004; Krivanek *et al.*, 2006; Welter *et al.*, 2007; Xu *et al.*, 2008; Marguerit *et al.*, 2009; Zhang *et al.*, 2009; Barba *et al.*, 2014), berry color and anthocyanin variation (Lijavetzky *et al.*, 2006; Fournier-Level *et al.*, 2009; Ban *et al.*, 2014; Costantini *et al.*, 2015), flower sex and cluster morphology (Dalbó *et al.*, 2000; Lowe and Walker, 2006; Riaz *et al.*, 2006; Marguerit *et al.*, 2009; Fechter *et al.*, 2012; Battilana *et al.*, 2013; Correa *et al.*, 2014), aroma and monoterpene content (Doligez *et al.*, 2006; Battilana *et al.*, 2009; Duchêne *et al.*, 2009), berry size and seedlessness (Doligez *et al.*, 2002; Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Doligez *et al.*, 2013), phenology and fruitfulness (Fanizza *et al.*, 2005; Costantini *et al.*, 2008; Doligez *et al.*, 2010; Duchêne *et al.*, 2012; Grzeskowiak *et al.*, 2013) and berry flesh firmness (Carreño *et al.*, 2014).

1.7.2. Association mapping

Association mapping (AM), also known as linkage disequilibrium (LD) mapping relies on LD to study the relationship between phenotypic variation and genetic polymorphisms across a set of diverse germplasm (Flint-Garcia *et al.*, 2003). The terms LD and AM have often been used interchangeably in the literature. However, they present subtle differences. According to Gupta *et al.* (2005), AM refers to the significant association of a marker locus with a phenotypic trait while LD refers to the non-random association between two markers. Thus, AM is actually an application of LD.

By exploiting broader genetic diversity, AM offers three main advantages over linkage mapping: i) AM exploits all the recombination events occurred during the evolutionary history of the

population under study, increasing hence the mapping resolution; ii) it can potentially examine the whole allelic diversity of the species at a time; iii) there is no need to make crosses and establish new populations to carry out the study, and particularly in the context of fruit trees, the plant material conserved at germplasm collections can be, in most cases, immediately phenotyped (Flint-Garcia *et al.*, 2003; Myles *et al.*, 2009).

1.7.2.1. Linkage Disequilibrium

Linkage disequilibrium is defined as the non-random association between alleles at different loci (Flint-Garcia *et al.*, 2003). LD is the outcome of a complex interplay of several evolutionary forces such as recombination, mutation, selection, genetic drift, mating system and population subdivision and admixture throughout the evolutionary history of a species (Ardlie *et al.*, 2002; Rafalski and Morgante, 2004). Therefore, it is expected to be highly variable, as is very often observed in simulation and empirical studies: molecular markers that are immediately adjacent to each other on a chromosome might be statistically independent, whereas those that are far apart might be highly correlated (Kruglyak, 1999; Jorde, 2000; Nordborg and Tavaré, 2002; Weiss and Clark, 2002; Zondervan and Cardon, 2004). For this reason, and in order to provide a high chance of identifying markers in high LD with the functional variant of a given trait, we need to understand the behavior of LD and how much it varies across the genome and among different populations (Kruglyak, 1999; Pritchard and Przeworski, 2001; Ardlie *et al.*, 2002; Flint-Garcia *et al.*, 2003). The extent of LD is often described using the statistics D' (Lewontin, 1964) and r^2 (Hill and Robertson, 1968). Pairwise D' and r^2 between two loci A and B with two alleles (A , a and B , b) at each locus are defined as follows:

$$D' = (P_{AB} - P_A \times P_B) / \max(P_{AB} - P_A \times P_B)$$

$$r^2 = (P_{AB} - P_A \times P_B)^2 / P_A \times P_B \times P_a \times P_b$$

Where P_{AB} and $P_A \times P_B$ are respectively the observed and expected frequencies of the haplotype AB , max refers to the maximal value and P_A , P_a , P_B and P_b are respectively the frequencies of the alleles A , a , B and b . Both D' and r^2 range between 0 (absence of LD) and 1 (complete LD).

In the context of association mapping, r^2 is usually preferable because it summarizes both recombinational and mutational history, it is indicative of marker-trait association (Flint-Garcia *et al.*, 2003; Carlson *et al.*, 2004) and it is less sensitive to small samples and easier to interpret than D' (Ardlie *et al.*, 2002; Gaut and Long, 2003).

Genome-wide distribution of LD has been studied in several plant species (Kim *et al.*, 2007; Stich *et al.*, 2008; Yan *et al.*, 2009; Comadran *et al.*, 2011; Xu *et al.*, 2012; Wang *et al.*, 2013; Würschum *et al.*, 2013). In grapevine however, there is still little information on this matter.

Most of LD studies reported in grapevine are based on the use of few markers and/or small sample sizes (Barnaud *et al.*, 2006; Lijavetzky *et al.*, 2007; This *et al.*, 2007; Fournier-Level *et al.*, 2009; Barnaud *et al.*, 2010; Emanuelli *et al.*, 2010; Myles *et al.*, 2010; Vargas *et al.*, 2013a; Vargas *et al.*, 2013b). In addition, most of them were focused to small genomic regions used for candidate gene studies or few chromosomes. In order to investigate the domestication history of grape, Myles *et al.*, (2011) compared the distribution of LD between *V. vinifera* and *V. silvestris*, using a set of 3,349 genome-wide distributed SNPs. They found that LD decayed very rapidly to $r^2 = 0.2$ within a distance lower than 1 kb and to $r^2 = 0.1$ within approximately 10 kb and appears unchanged between the wild ancestor and the domesticated grape. However, Barnaud *et al.* (2006) found that, LD assessed using SSR markers extends further than SNP-based LD, decreasing to $r^2 = 0.1$ within ca. 5 cM/650–1,080 kb (using genotypic data) or 10 cM/1,300–2,160 kb (using haplotypic data). These results suggested that a sparse set of SSR markers distributed homogenously over the grapevine genome would be sufficient to perform genome wide association mapping with reasonable statistical power (Barnaud *et al.*, 2006).

1.7.2.2. Methodology of Association Mapping

Two kinds of association mapping studies can be distinguished according to their scale and focus: i) candidate-gene association mapping, which focus on the study of marker-trait associations in particular candidate genes that are supposed (on the basis of prior information) to be involved in controlling the phenotypic variation for a given trait; ii) genome-wide association mapping (GWAS), which study the genetic polymorphism across the whole genome in order to find signals of association for a trait (or traits) of interest. The recent advances in genomic technology have made affordable to genotype tens to hundreds of thousands of genome-wide distributed SNPs across a large number of samples. As a result, plant geneticists are now moving toward comprehensive genome-wide studies (Zhu *et al.*, 2008).

In general, AM methodology can be subdivided into four principal steps regardless of whether candidate gene or GWAS is used:

1) Defining the association mapping population

This initial step is critical to the success of association analysis, since the genetic diversity, the extent of LD and the level of relatedness within the population determine the mapping resolution, markers density, methods of analysis, and statistical power (Balding, 2006; McCarthy *et al.*, 2008). The definition of an association mapping population consists usually in selecting a subset of the available germplasm for the species in question, which can exist as germplasm bank collections, synthetic populations and/or elite germplasm (Bressegello and Sorrells, 2006b). The natural diversity of fruit trees such as grapevine is generally conserved as

germplasm bank collections. In this case, core collections should be suitable for association studies, since they can represent most of the genetic variability with a manageable number of accessions.

2) Genotyping the population

In association studies, two steps of genotyping are usually carried out:

- A preliminary genotyping of the mapping population, using a sparse set of unlinked genome-wide distributed “background markers” to characterize the genetic composition of the population (Zhu *et al.*, 2008). The genotypic data generated in this step are usually used to assess the genetic diversity of the mapping population, and to study the population structure and estimate pairwise relatedness coefficients that are incorporated later into the association analysis.
- The final genotyping step, whose scale and strategy vary according to the association mapping method: For candidate gene studies, it is necessary to identify and genotype molecular markers located within or close to the gene of interest. This is usually achieved by resequencing DNA amplicons from the individuals of the association population. And for GWAS, ideally the whole genome should be covered with molecular markers (usually SNPs) at a density that reflects genome-wide LD structure. SNPs for whole-genome scan are usually genotyped using high throughput technologies including array-based techniques and genotyping by sequencing.

3) Phenotyping the population

The collection of accurate phenotypic data is as critical as genotyping to the success of the association study. Since this approach often involves a relatively large number of diverse accessions, phenotypic data collection with adequate replications across multiple years and multiple locations is challenging. Special attention should be given to the field design particularly if the field conditions are not homogenous. In the case of fruit trees, *in situ* conserved germplasm collections offer the opportunity to collect immediately phenotypic data.

4) Analyzing the data

In association testing, the strength of a potential relationship between a SNP and a phenotype is quantified by a statistical model. The basic statistical methods for association analysis, under an ideal situation, would be chi-square test, *t* test analysis of variance, or linear regression. However, since population structure and cryptic relatedness can generate spurious associations (Lander and Schork, 1994; Pritchard *et al.*, 2000b; Korte and Farlow, 2013), different statistical approaches have been designed to deal with these confounding factors. Particularly in crop systems, germplasm accessions can share recent common ancestry, which if not taken into account can also lead to spurious associations. For this reason, the inclusion of a kinship matrix

into the GWAS model is often more important and powerful than the inclusion of fixed-effect population structure covariates for reducing false positives. Usually, different linear regression models are tested for marker-trait association:

- Simple linear models or Linear Models (LM): these include i) the naïve model (GLM; General Linear Model) that performs association analysis using a least squares fixed effects linear model, ii) the Q model that is a GLM that uses the population membership estimates (Q matrix) as covariates to control for population structure (Yu *et al.*, 2006) and iii) the P model that is a GLM with PCs (Principal Components) as correction for population structure (Price *et al.*, 2006).
- Mixed Linear Models (MLM): a mixed model includes both fixed and random effects. Including random effects gives MLM the ability to incorporate information about pairwise relationships among individuals (known as kinship or K matrix). In these approaches, the Q matrix is fitted as fixed effect, whereas the K matrix is incorporated as the variance-covariance matrix of random effect for the individuals (Yu *et al.*, 2006). The MLM can be performed correcting only for pairwise relatedness (K model), or for both K and Q (Q+K model)
- Advanced Mixed Models: recently, more advanced versions of MLM have been developed, such as the Multi Locus Mixed Model (MLMM) that has been used for the identification of particular cases of synthetic associations and/or allelic heterogeneity (Segura *et al.*, 2012) and the Multi Trait Mixed Model (MTMM) that has been used for traits that are correlated due to pleiotropy (shared genetics) or linkage between causative polymorphisms (Korte *et al.*, 2012).

Interpretation of the results of GWAS is not trivial due to the multiple-testing problem. Statistical tests are generally called significant and the null hypothesis is rejected if the p-value falls below a predefined alpha value, which is usually set to 0.05. This means that 5% of the time, the null hypothesis is rejected when in fact it is true and we detect a false positive. This probability is related to a single statistical test; in the case of GWAS, hundreds of thousands to millions of tests can be conducted, each one with its own false positive probability. Therefore, testing a large number of markers requires a balance between detection power and the chance of making false discoveries. One of the most used approaches to correct for multiple testing is the Bonferroni correction. The Bonferroni correction adjusts the alpha value from $\alpha=0.05$ to $\alpha=(0.05/k)$ where k is the number of statistical tests conducted. For a typical GWAS using 50,000 SNPs, statistical significance of a SNP association would be set at 1E-06. It should be noted that this correction could be very conservative, as it assumes that each association test of

the 50,000 is independent of all other tests – an assumption that is generally untrue due to linkage disequilibrium among markers.

1.7.2.3. Current status of association mapping in plants

The first association mapping study conducted in plants have been published by Thornsberry *et al.*, (2001). It was focused on the *Dwarf8* candidate gene for flowering time in maize. Since then a series of AM studies have been published in several plant species, including maize, *Arabidopsis*, sorghum, wheat, barley, potato, rice, *Pinus taeda*, sugarcane, and *Eucalyptus* (reviewed in Zhu *et al.*, 2008). Most of these studies are focused to the study of particular traits using limited resources regarding sample size and number of markers through the candidate gene approach. More recently, the advent of cheaper high throughput genotyping technologies encouraged the plant geneticists to conduct genome-wide association studies. So far, GWAS have been conducted with promising results in the model species *Arabidopsis thaliana* (Atwell *et al.*, 2010) as well as a number of crop plants such as maize (Tian *et al.*, 2011; Riedelsheimer *et al.*, 2012; Li *et al.*, 2013), rice (Huang *et al.*, 2010; Zhao *et al.*, 2011), sorghum (Morris *et al.*, 2013), sunflower (Mandel *et al.*, 2013), tomato (Sauvage *et al.*, 2014) and chickpea (Kujur *et al.*, 2015). Based on the magnitude of resources employed, rice and maize are the two major models for plant GWAS and both have panels of thousands of genotyped lines (with up to a million SNPs) and multiple environment trials conducted for several traits (Huang and Han, 2014). On the other hand, modest GWAS based on a low marker density (either SNPs or SSRs) and/or small mapping populations have been also conducted in some species including barley (Pasam *et al.*, 2012), *Brassica napus* (Harper *et al.*, 2012), wheat (Brescghello and Sorrells, 2006a; Peng *et al.*, 2009; Liu *et al.*, 2010), rice (Agrama *et al.*, 2007), sorghum (El Mannai *et al.*, 2011), cotton (Zhao *et al.*, 2014) and watermelon (Reddy *et al.*, 2014).

In grapevine, all the association studies published so far are focused on the study of particular candidate genes including: *Myb*-genes (This *et al.*, 2007; Fournier-Level *et al.*, 2009), *VvDXS* (Emanuelli *et al.*, 2010), *VvPel* (Vargas *et al.*, 2013b), *VvGAI* (Vargas *et al.*, 2013a), *VvTFL1A* (Ward *et al.*, 2013; Fernandez *et al.*, 2014), and 183 candidate genes related with cluster architecture (Tello *et al.*, 2015). However, GWAS have not yet been tested in grapevine for the study of the genetic determinism of quantitatively inherited traits. On the other hand, population stratification has been reported in different grapevine germplasm collections studied using molecular markers (Aradhya *et al.*, 2003; Bacilieri *et al.*, 2013; Emanuelli *et al.*, 2013; Vargas *et al.*, 2013). Therefore, it is to expect that GWAS in grapevine will be prone to confounding effects. However, the extent of population structure confounding effect in the context of GWAS, and the efficiency of different statistical approaches in dealing with it are unknown.

Several reasons make of grapevine an excellent candidate for GWAS: first, grapevine has a medium sized genome (around 500 Mb) whose complete sequence is already available (Jaillon *et al.*, 2007; Velasco *et al.*, 2007); second, unlike selfing species characterized by extensive LD, grapevine is highly heterozygote resulting in a very rapid LD decay (Myles *et al.*, 2011), which would allow high resolution mapping of genetic associations; third, large to medium-size field collections of grapevine germplasm are available throughout the world (<http://www.vivc.de/>), thus, mapping collections can be genotyped once and directly phenotyped for different traits during several seasons without extra costs; additionally, thanks to the vegetative propagation, the same mapping collection can be reproduced in other environments and maintained indefinitely.

OBJECTIVES

2. OBJECTIVES

Grapevine is one of the oldest crops and from an economic point of view, the most important fruit species. However, our understanding of the genetic basis of relevant traits in grapevine is still very scarce. The identification and analysis of these determinants could provide information and tools useful to improve the management and breeding of this crop. The main objective of this work was the evaluation of a strategy not used until now in this species, GWAS, to gain more in-depth knowledge about the genetic determinants involved in the expression of most relevant agronomic and quality related traits in grapevine, and especially in table grape.

With this aim and taking advantage of the availability of an extended grapevine germplasm collection at the germplasm bank of El Encín (<http://www.madrid.org/coleccionvidencin/> Alcalá de Henares, Madrid, Spain) and the relatively affordable high throughput genotyping tools, we have performed different genome wide association studies using both, a high density SNP dataset and a low density SSR coverage.

To achieve this, we have set the following partial objectives:

1. The assessment of the efficiency of GBS for whole-genome SNP genotyping of a grapevine diversity panel.
2. The development and use of a panel of multiplex PCRs for microsatellite-based whole-genome analysis of a grapevine diversity panel.
3. The analysis and comparison of the extent and distribution of linkage disequilibrium with SNP and SSR markers.
4. The evaluation of SNP and SSR-based genome-wide association analysis for a set of table grape related traits.

MATERIAL AND METHODS

3. MATERIAL AND METHODS

3.1. PLANT MATERIAL

The collection of varieties used in this work (“the mapping collection”), consists of 274 grapevine accessions (Table A.1) that were selected following two steps. First, a core collection of 207 non-redundant grapevine genotypes with different uses and origins was constructed to represent the genetic and phenotypic diversity contained in the 1,852 *V. vinifera* accessions maintained at the germplasm bank of El Encín (Alcalá de Henares, Madrid, Spain; <http://www.madrid.org/coleccionvidencin/>). The selection process was based on the genotypic information of the 26 SSRs included in multiplex PCRs Mx01 and Mx02 (see section “Results”), referred to as “control loci” from now on. Later, a supplementary set of 67 accessions was added to further enlarge the phenotypic diversity and geographic representativeness of the mapping collection. This mapping collection included 14 known trio pedigrees (Table A.2) involving 31 out of the 274 accessions, which were used to check the consistency of genotypic scores.

3.2. PHENOTYPIC CHARACTERIZATION

3.2.1. Experimental conditions

Each accession in the mapping collection is represented by 4 to 16 vines (5 in the majority). All the plants are cultivated under the same conditions, within the vineyard of the germplasm bank of El Encín (Alcalá de Henares, Madrid, Spain). The region is under the semiarid Mediterranean climate. The plants are grafted onto 110 Richter, trained following the unilateral cordon trellis system and pruned to five spurs of two buds. The distance between vines is 0.9 m in rows and 2.5 m between rows. The vineyard is managed according to the organic agriculture practices as defined in the European Council regulations (CEE 2092/91). Supplemental irrigation is provided to compensate for water deficit during the dry season. The vines were seven years old in 2010 (the first year of phenotyping). In order to protect the grapes against animal attacks, the vineyard is net fenced; and during the veraison period of each growing season, seven clusters from each accession of the mapping collection were wrapped in paper bags open at the bottom to prevent bird attacks. Additionally, traps for wild boards and raccoons had been installed within the vineyard throughout the harvesting season.

3.2.2. Traits evaluation

The 207 cultivars included in the core collection were evaluated for 21 phenotypic traits related to fertility, phenology, cluster and berry quality (Table 1) during four growing seasons: 2010, 2011, 2012 and 2013; but the supplementary set of 67 accessions was evaluated only in 2013. In 2010, phenological traits (budbreak (Bb), flowering (Fl), veraison (Vr), and ripening (Rp) times) and fertility (number of inflorescences per shoot) were not assessed. In 2011, only Bb, Fl and fertility could be assessed because of a severe attack of downy mildew (*Plasmopara viticola*) causing the loss of almost all the production. Most of the traits were assessed according to the OIV descriptors (OIV, 2009) with minor modifications (Table 1). Phenotypic traits involving subjective decisions were assessed jointly by three trained judges (Table 1). Prior to be analyzed statistically, some traits were codified or transformed as follows: phenological data recorded as dates were transformed to the number of days since January 1st of each growing season; flower sex, berry skin color and muscat flavor were transformed to binary traits and codified as 0/1 for hermaphrodite/female, non-colored (green-yellow)/colored, and non-muscat/muscat, respectively. Additionally, as these three traits were qualitative and stable over years, their respective data recorded during the three years of study were merged in a single consensus dataset considering the mode value for each individual.

3.2.3. Statistical analysis of phenotypic data

Normality distribution of the assessed traits was checked using Kolmogorov-Smirnov's test. The null hypothesis of normal distribution was rejected when $P \leq 0.05$. As some traits were ordinal rather than quantitative, and some quantitative traits were not distributed normally, we used the non-parametric Spearman rank correlation coefficient (*Rho*) to assess correlation between traits and between years of study. All the analyses were performed using IBM SPSS Statistics 22.0.

Table 1. List of the evaluated traits with their abbreviated names, OIV codes and summarized descriptions

Trait	Abbreviation	OIV code	Description
Time of budbreak ^a	Bb	OIV 301	date when 50% of the buds are in green shoot tip stage
Time of flowering ^a	Fl	OIV 302	date when 50% of flowers are open.
Time of veraison ^a	Vr	OIV 303	date when 50% of the berries start softening
Time of ripening ^a	Rp	OIV 304	date when ripening was reached (when the must was 18 ° Brix)
Flower sex ^a	FSx	OIV 151	Flowers were scored at flowering period as hermaphrodite (erected stamens) or female (reflexed stamens)
Fertility	Frt	OIV 153	mean number of inflorescences per shoot evaluated in three vines and a maximum of 10 shoots per each vine
Mean cluster weight	MCW	OIV 502	mean weight of five clusters
Mean cluster length	MCL	OIV 202	mean length of five clusters measured as the distance from the uppermost to the lowest berry of the primary cluster
Mean cluster diameter	MCD	OIV 203	mean diameter of five clusters measured as the maximum distance between the lateral berries of the primary cluster
Mean peduncle length	MPL	OIV 206	mean peduncle length of five clusters measured as the distance from the insertion point on the shoot to the first ramification of primary cluster
Cluster compactness index according to the OIV descriptor ^b	CI-OIV	OIV 204	mode value of the compactness index of five clusters scored according to an ascending scale of five levels (1, 3, 5, 7 and 9) ranging from very loose to very compact
Cluster compactness index, according to Tello and Ibáñez (2014)	CI-12		CI-12 = MCW/(MCL) ²
Mean berry weight	MBW	OIV 503	mean weight of 100 berries taken from the middle part of five representative clusters
Mean berry length	MBL	OIV 220	mean value of 10 berries taken from the middle part of five representative clusters
Mean berry diameter	MBD	OIV 221	mean value of 10 berries taken from the middle part of five representative clusters
Berry shape index	ShI		ShI = MBL/MBD
Berry skin color ^b	Col	OIV 225	scored at maturity as green-yellow, rose, red, dark red-violet or bleu-black
Flesh firmness index ^b	FI	OIV 235	scored following a scale of five levels: 1, 3, 5, 7, and 9, going respectively from soft flesh to very firm flesh, by tasting at least five peeled berries
Muscat flavor ^b	Mst	OIV 236	scored by tasting mature berries as muscat or non-muscat (neutral or other particular flavors)
Mean number of seeds per berry	MNS		mean number of fully developed seeds in 10 berries taken from the middle part of five clusters
Mean seed dry weight per berry	MSDW		mean dry weight of all kind of seeds (fully developed and/or rudimentary) extracted from 10 berries taken from the middle part of five clusters

^a trait evaluated in the vineyard by three trained judges.

^b trait evaluated in the laboratory by at least three trained judges.

3.3. SNP GENOTYPING USING GENOTYPING BY SEQUENCING

3.3.1. DNA extraction

Frozen young leaves were ground to a fine powder using the Mixer-Mill-MM300 grinder (Retsch, Haan, Germany) and liquid nitrogen. DNA was purified using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. In order to remove polyphenols, a 3% PVPP (Sigma-Aldrich, Steinheim, Germany) was added to the lysis buffer. Extracted DNA was quantified using the ND-1000 spectrophotometer (NanoDrop, Delaware, USA) and checked for integrity (absence of smeared DNA band patterns) by gel electrophoresis against Lambda Phage DNA (Sigma-Aldrich, Madrid, Spain). To check for any significant error occurred during tissue sampling and/or DNA extraction, the identity of each extracted sample was confirmed using a subset of the 26 control markers as described below (see section 3.4.4.).

3.3.2. Library preparation and sequencing

The extracted DNA was sent to the Institute of Genomic Diversity at Cornell University (New York, USA) for genotyping according to the GBS protocol (Elshire *et al.*, 2011). DNA libraries were prepared using the *ApeKI* methyl-sensitive restriction enzyme (see section 1.5.2. for more details about library preparation). The 274 DNA libraries were pooled together and sequenced on three flow cell lanes (288-plex each) using the Illumina Genome Analyser III (Illumina, San Diego, CA) at Cornell University Life Sciences Core Laboratories. Quality check of the sequencing data revealed that read counts obtained for some samples were lower than expected (less than 10% of the mean). In order to improve read coverage, DNA libraries were prepared again for 85 accessions with lowest read coverage within the core collection (the additional collection was not considered). These libraries were prepared using newly extracted DNA following the same protocol as described above, and sequenced in a single flow cell lane as part of a 288-plex sequencing reaction.

3.3.3. SNP calling

In order to obtain the SNP genotypes for each individual, the Illumina sequencing data were processed in the Computational Biology Service Unit (CBSU, Cornell University) using TASSEL 3.0 GBS Discovery Pipeline (Glaubitz *et al.*, 2014), with some adaptations to heterozygous organisms. The bioinformatics processing of GBS sequences (see the section 1.5.2. for more details) consists briefly in: the raw 100 bp reads are first trimmed to remove barcodes then merged into a single tag count. The resulting tags are then aligned to the 12x.1 *V. vinifera* PN40024 reference genome (Canaguier *et al.*, 2014) using a Burrows–Wheeler Aligner (BWA)

(Li and Durbin, 2009). In parallel, the number of times each tag is observed in a given individual is calculated using the barcode information. SNP detection (SNP calling) is performed finally by aligning the individual tags located at the same unique position on the reference genome against each other. The final SNP data matrix was saved under the VCF format for subsequent filtering and analysis.

3.3.4. Data filtering and error checking

The raw SNP data were submitted to a stratified filtering process using the VCFtools v4.0 package (Danecek *et al.*, 2011) as illustrated in Fig. 3. The objective of data filtering was obtaining a high quality reliable SNP dataset by: (1) keeping only SNP genotypes with high quality scores ($GQ \geq 98$); (2) discarding individuals and SNPs with high proportion of missing values and (3) discarding SNPs with too low minor allele frequency ($MAF < 0.01$) to minimize the impact of sequencing errors. Finally, four different datasets of SNPs were generated with increased minimal value of minor allele frequency (DS4 to DS7) to be used for LD analysis and GWAS.

The accuracy of SNP calling was assessed by checking Mendelian inheritance in the 14 trio pedigrees included in the mapping collection using a nested function-based formula in Excel. A SNP calling error rate was estimated based on the detected Mendelian errors: assuming that each Mendelian error detected in a given pedigree is the consequence of only a single genotype calling error (since the probability that two or more genotyping errors occur in the same pedigree should be negligible), the SNP calling error rate was estimated as the number of Mendelian errors detected over the total number of individual genotypic scores, averaged over the 14 pedigrees:

$$\text{Error rate per pedigree} = \frac{\text{number of mendelian errors}}{\text{number of SNPs} \times 3}$$

The number of SNPs is multiplied by three because each pedigree has three members.

Figure 3. Schematic representation of the SNP filtering process

For each data set (DS) is shown the number of SNPs included, number of individuals (inds) and overall proportion of missing values (missing). The consecutive filters (from F1 to F7) are indicated by discontinuous lines. The filtering criteria used in each filter are from the top to the bottom: minimum genotype quality (GQ), minimum call rate per individual (MIND), minimum call rate per SNP (GENO), and minimum minor allele frequency (MAF), the filtering threshold values are given at the right side of each filtering criteria.

3.4. SSR GENOTYPING

3.4.1. DNA extraction

Frozen young leaves of each accession were ground to a fine powder using the Mixer Mill MM300 grinder (Retsch, Haan, Germany) and liquid nitrogen. Total DNA was purified using the BioSprint 96 workstation and the BioSprint 96 DNA Plant Kit (Qiagen, Hilden, Germany) according to the manufacturer instructions. Extracted DNA was quantified using the ND-1000 spectrophotometer (NanoDrop, Delaware, USA).

3.4.2. Markers selection

Initially, 249 SSR markers were selected from the literature and genomic databases (NCBI: <http://www.ncbi.nlm.nih.gov/probe> and GENOSCOPE: <http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/>; Table A.7) based on their physical position (inter-marker distance approx. < 3 Mb according to the 12x version of the *V. vinifera* genomic sequence (GENOSCOPE), and the available information about their degree of polymorphism. Alternative primers were designed for 37 of them (the new marker was coded adding "-2" at the end of its original name) because they presented amplification problems or evidence of existence of null alleles (Table A.7). The genomic sequences of the redesigned markers were retrieved from the 12x version of the *Vitis* genome sequence by "Blast search" using as probes the sequences of the original primers. Then, new primers were designed in the flanking regions of the microsatellite repeats using the software Primer3 v4.0 (Rozen and Skaletsky, 2000), and tested for PCR amplification. Additionally, primers for 22 new microsatellite markers were designed *in silico* in order to cover chromosomal regions for which no markers were available (Table A.7). Genomic sequences at intervals of approximately 2 Mb within the chromosomal gaps were retrieved from the 12x grapevine genome sequence and investigated using the WebSat software (Martins *et al.*, 2009) to identify microsatellites and design primer pairs for their amplification. The specificity of the candidate primer pairs was checked by aligning them against the genome sequence using Primer-BLAST (Ye *et al.*, 2012). Finally, the selected primer pairs were submitted to amplification test as described in section 3.4.3.1.

3.4.3. Multiplex PCRs design and genotyping

Allelic information for many selected SSRs was available only for the accession in which the marker was identified or few accessions more. For that reason, an expected fragment-size range was established for each selected locus by adding 30 bp to both sides of the allelic range determined according to the bibliographic and database information. The markers were then organized in groups of one to four with non-overlapping fragments to be labeled with the same fluorescent dye: FAM, NED, PET or VIC (Applied Biosystems, Foster city, CA, USA). After that, quadruplets combining four different dye-groups were constructed to constitute the multiplex sets.

3.4.3.1. Amplification tests

Labeled primer pairs for each locus were tested by genotyping a set of seven grapevine cultivars (Airén, Cabernet Sauvignon, Cardinal, Chardonnay, Crimson Seedless, Flame Seedless and Italia), using single locus PCR in order to check their functionality and to confirm/correct allelic

ranges. Single locus PCRs were performed in a total volume of 21 µl containing 5 ng of DNA template, 1x PCR buffer, 2 mM MgCl₂, 0.2 mM of each dNTP, 0.2 µM of each primer and 1 U of Taq DNA Polymerase (Biotools, Madrid, Spain). Amplifications were carried out using the following thermocycling conditions: one cycle at 95°C for 5 min; followed by 35 cycles at 94°C for 30 s, 52°C for 1 min (suitable for most of the loci and modified for the rest) and 72°C for 1 min; and a final step of 72°C for 30 min. PCR products were separated by capillary electrophoresis in an ABI3130 sequencer (Applied Biosystems, Foster city, CA, USA). The loading mixture contained 1 µl of PCR product diluted 50 to 160 times (depending on the intensity of the amplified fragments on 2% agarose gels), 0.1 µl of internal size standard (GeneScan-500LIZ; Applied Biosystem, Foster city, CA, USA) and 14 µl of Hi-Di Formamide (Applied Biosystem, Foster city, CA, USA). Prior to be loaded into the sequencer, the samples were denatured at 95°C for 5 min. Raw data obtained by capillary electrophoresis were transformed into allelic sizes using the GeneMapper v4.1 software (Applied Biosystem, Foster city, CA, USA).

3.4.3.2. Multiplex PCRs optimization

Based on the results of the amplification tests, the initially proposed multiplex sets were confirmed or corrected to avoid marker overlapping, then subjected to multiplex PCR assays using the same set of tester cultivars. Initially, all the multiplex PCRs were performed in a final volume of 15 µl containing: 1x Multiplex PCR Master Mix (Qiagen, Hilden, Germany), 5 ng of template DNA and an equimolar amount of 0.2 µM of each primer pair. Subsequently, each multiplex PCR was subjected to several cycles of optimization, modifying consecutively primer concentrations according to the signal intensity (peak height) of the amplification products (decreasing for the stronger ones and increasing for the weaker ones), until an acceptable equilibrium between the combined primer pairs was reached.

Four touchdown thermocycling programs (Don *et al.*, 1991), coded as Tp-A, Tp-B, Tp-C and Tp-D (Table A.6) were used for PCR amplification. All of them included a 90 min. at 72°C final extension step in order to address the “Plus-A” artifacts in PCR products (Smith *et al.*, 1995). Each multiplex was initially tested using the programs Tp-A (standard program) or Tp-B (suitable for primers with lower annealing temperatures). When amplification deficiencies were detected for any locus, the other programs were tested to identify the best performing one. Finally, the designed and optimized multiplexes were used to genotype the mapping collection. The supplementary set of 67 accessions was genotyped only for 109 SSRs (see Table A.7).

3.4.4. Data checking

In order to minimize possible mistakes, several control points were included throughout the study: (1) Sampling errors and possible contaminations during DNA extraction were checked by confirming the identity of the sampled accessions using the 26 control loci, which had been previously used to genotype the complete germplasm bank of El Encín (Alcalá de Henares, Madrid, Spain); (2) To detect possible errors during DNA manipulation or electrophoresis related problems, one control locus was included in each multiplex PCR (Table A.5); (3) Contaminations during the amplification were monitored using negative controls in each PCR; (4) To verify assay reproducibility, 20 accessions were genotyped twice; (5) PCR run and/or sequencer loading were repeated for any sample that failed to generate detectable amplicons; (6) Both allele definition (binning process) and genotyping (peaks selection) in GeneMapper v4.1 (Applied Biosystem, Foster city, CA, USA) were manually reviewed by two persons prior to generate the final genotypic table; (7) Finally, allelic inheritance was checked for each locus using the 14 trio pedigrees included in the mapping collection. When incompatibilities were found, allele calling was reviewed for all the individuals to discard mistakes in peak selection or data typing. If the incompatibility persisted, the amplification of the problematic marker was repeated using single locus PCRs with different DNA polymerases (Standard DNA polymerase or *pfu* DNA polymerase (Biotools, Madrid, Spain)) and/or with different thermocycling parameters. When the compatibility could not be recovered, the locus in question was suspected to present null alleles or unspecific amplification and new primer pairs were designed for it.

3.4.5. SSR data analysis

The genotypic dataset obtained for the core collection was used to estimate the following parameters for each analyzed locus: number of detected alleles per locus (K), observed heterozygosity (H_{obs}), expected heterozygosity (H_{exp}), polymorphic information content (PIC), probability of identity of unrelated individuals (PI) and estimated frequency of null alleles (F). These parameters were computed employing Cervus 3.0 software (Kalinowski *et al.*, 2007).

In order to evaluate the quality of the amplification profiles of the studied SSRs, we established a panel of seven descriptors related to the stuttering patterns, “+A” effect, low heterozygote peak ratio, presence of artifacts, separation between adjacent fragments, multilocus patterns and a global evaluation of the easiness of scoring of the marker. These descriptors were evaluated by visual examination of the electrophoretic profiles as displayed in GeneMapper v4.1 (Applied Biosystem, Foster city, CA, USA).

In order to examine SSR distribution in the genome with respect to coding regions, gene density distribution over the grapevine chromosomes was estimated using the information of grapevine

genome annotation available at the GENOSCOPE database (<http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/>). Gene density along each chromosome was represented as number of genes in consecutive chromosomal bins of 500 kb.

3.5. POPULATION STRUCTURE AND RELATEDNESS ANALYSIS

Given the higher resolution and power of SSRs in detecting genetic diversity and population structure (Yu *et al.*, 2009; Yang *et al.*, 2011; Emanuelli *et al.*, 2013), the population structure of the studied collection was analyzed using the genotypic data of 19 SSR markers (one from each chromosome, Table A.7). Population structure analysis was performed using the STRUCTURE 2.3.4 software (Pritchard *et al.*, 2000a). Ten independent runs for K values ranging from 1 to 10 were performed with a burn-in period length of 200,000, followed by 500,000 MCMC (Markov Chain Monte Carlo) iterations. The “admixture ancestry” and “correlated allele frequency” models were assumed, and no prior population information was used. The optimal subpopulations number (K) was chosen by applying the Evanno criterion (ΔK) (Evanno *et al.*, 2005). Individuals with probability of membership greater than 0.75 were assigned to subpopulations, while those with lower probabilities were assigned to the admixed group.

In order to assess the patterns of variation of phenotypic traits among the subpopulations determined by STRUCTURE, we performed a one-way ANOVA for each quantitative trait, with subpopulations as the independent factor, followed by means comparison. Additionally, we plotted the distribution of qualitative and ordinal traits over the subpopulations. ANOVA was performed using IBM SPSS Statistics 22.0, considering $P = 0.05$ as the significance threshold.

Pairwise kinship coefficients were calculated on the basis of the same molecular data used of structure analysis (19 microsatellite markers) according to the formula of Loiselle *et al.*, (1995) implemented in the software SPAGeDi (Hardy and Vekemans, 2002). Prior to be used for association analysis, negative kinship coefficients between individuals were set to zero as this indicates that they are less related than random individuals (Yu *et al.*, 2006).

3.6. LINKAGE DISEQUILIBRIUM ANALYSIS

3.6.1. SNP-based LD

LD was studied using the r^2 parameter (Hill and Robertson, 1968) computed via the software Haploview 4.2 (Barrett *et al.*, 2005). The analysis was performed using the dataset DS5 (25,961 SNPs at 0.05 MAF over 242 individuals; See Fig. 3) according to the following scheme:

1. Background LD (LD between SNPs located on different chromosomes) was studied using random samples of 100 SNPs chosen from each chromosome (a total of 1,900 SNPs).

Pairwise r^2 values between all unlinked SNP pairs were used to estimate the extent of background LD.

2. LD within chromosomes was estimated using all pairwise combinations of SNPs on each chromosome. The decay of LD over physical distance was investigated by plotting pairwise r^2 values against the physical distance between SNPs. Data-points were fitted using logarithmic regression function as implemented in the software Excel (Vargas *et al.*, 2013a). The overall distance of LD decay in each chromosome was estimated by determining the abscise value (physical distance) that gives solution to the regression function for the critical r^2 threshold of 0.2 (Zhu *et al.*, 2008). The number of SNPs needed for grapevine genome coverage was estimated by dividing the chromosome size by the average distance of LD decay.
3. Additionally in a sample of three chromosomes (chr1, chr2 and chr3) was studied:
 - i. The effect of minor allele frequency (MAF) on LD by calculating r^2 separately for SNPs with minimum MAF greater than 0.01, 0.05, 0.10, and 0.20 contained in the datasets DS4, DS5, DS6, and DS7 respectively (Fig. 3).
 - ii. The effect of sample size on LD by calculating r^2 in random samples of 15, 30, 60, and 120 drawn from the total collection of 242 individuals of the DS5.
 - iii. The effect of population structure on LD, using r^2 estimates from subsets of SNP data drawn from DS5 for each of the three subpopulations inferred by STRUCTURE.

3.6.2. SSR-based LD

SSR markers that were monomorphic in the core collection, that were difficult to score, and/or showing strong evidence of presence of null alleles were discarded. Rare alleles (alleles with frequency < 1%) were set as missing. The resultant genotypic data matrix composed of 250 SSR markers over 207 accessions (Table A.7) was used in this analysis. LD patterns were studied by calculating locus-pair multi-allelic r^2 using the software PowerMarker v3.25 (Liu and Muse, 2005). This statistic is implemented in PowerMarker for multi-allelic marker pairs as a weighted mean of the absolute allele-pair linkage disequilibria:

$$LD = \sum_u \sum_v P_u P_v |r_{uv}^2|$$

Where p_u and p_v are allele frequencies and r_{uv}^2 is the correlation coefficient between the two alleles. All the calculations were carried out without assuming Hardy-Weinberg equilibrium. Background LD was assessed as pairwise r^2 between all unlinked pairs of SSR markers (located on different chromosomes). LD within chromosomes was estimated using all pairwise combinations of SSRs on each chromosome. The decay of LD over physical distance was

investigated by plotting pairwise r^2 values against the physical distance between SSRs. Average distances of LD decay below the critical values of 0.2 (Zhu *et al.*, 2008) and the number of SSRs needed to cover the genome were estimated following the same procedure as described for SNP markers.

3.7. GENOME-WIDE ASSOCIATION ANALYSIS

3.7.1. SNP-based association analysis and model testing

Association analyses were performed separately for each year of study, using the SNP dataset DS4 (Fig. 3) that consists of 36,332 SNPs over 242 individuals, with a 0.01 MAF, and a mean rate of missing values of 12%. These analyses were based on a sample size of 242 accessions in 2013, and 207 accessions (only the core collection) in 2010 and 2012, due to the lack of phenotypic data for the supplementary set in these two years. GWAS were conducted using the software package TASSEL 4.0 (Bradbury *et al.*, 2007). The genome-wide significance threshold was determined using the Bonferroni correction for multiple testing for $P < 0.05$, with the total number of SNPs as correction factor.

Prior to perform association analyses for all the traits, four statistical models of increasing complexity were tested (Yu *et al.*, 2006): 1) The naïve General Linear Model (GLM); 2) GLM corrected for population structure (Q model); 3) Mixed Linear Model (MLM) that accounts only for pairwise relatedness between individuals (K model); and 4) MLM that accounts for both pairwise relatedness and population structure (Q+K model). The respective equations of these models are as follows:

$$y = X\alpha + \varepsilon \quad (1)$$

$$y = X\alpha + Q\beta + \varepsilon \quad (2)$$

$$y = X\alpha + K\mu + \varepsilon \quad (3)$$

$$y = X\alpha + Q\beta + K\mu + \varepsilon \quad (4)$$

Where y is the phenotype, X is the genotype, Q is the matrix of sub-population membership probabilities obtained from STRUCTURE (Pritchard *et al.*, 2000), K is the kinship matrix calculated using SPAGeDi (Hardy and Vekemans, 2002), α and β are unknown vectors of fixed effects, μ is unknown vector of random effects and ε is unknown vector of random residual.

These models were tested against four “model traits” whose moderate to major effect genes/QTLs were extensively investigated and confirmed in several studies:

- i. Berry skin color, controlled by *Myb* genes on chr2 (Kobayashi *et al.*, 2002; Kobayashi *et al.*, 2004; Lijavetzky *et al.*, 2006; This *et al.*, 2007; Walker *et al.*, 2007; Fournier-Level *et al.*, 2009; Azuma *et al.*, 2011; Carrasco *et al.*, 2014; Costantini *et al.*, 2015).

- ii. Flower sex, controlled by a locus on chr2 near the SSR markers VVMD3 and VVIB23 (Dalbó *et al.*, 2000; Lowe and Walker, 2006; Riaz *et al.*, 2006; Marguerit *et al.*, 2009; Fechter *et al.*, 2012; Battilana *et al.*, 2013).
- iii. Muscat flavor, with a major effect QTL mapped on chr5, around the *VvDXS* gene (Doligez *et al.*, 2006; Battilana *et al.*, 2009; Duchêne *et al.*, 2009; Emanuelli *et al.*, 2010; Battilana *et al.*, 2011).
- iv. Seed content, with a major effect QTL on chr18, near the SSR marker VMC7F2 and the candidate gene *VvAGL11* (Doligez *et al.*, 2002; Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Mejía *et al.*, 2011; Doligez *et al.*, 2013).

The evaluation of the four statistical models was based firstly on their ability to detect the previously reported major loci controlling these traits, and secondly on their ability to reduce the rate of false discoveries that was inferred through the examination of quantil-quantil (QQ) plots (Balding, 2006; McCarthy *et al.*, 2008). Basically, the deviation of the observed *P*-values from the expectation under the null hypothesis across the distribution is indicative of an excess of false positives in the results.

3.7.2. SSR-based association analyses

SSR-based association analyses were performed using a dataset composed of 264 SSR markers (Table A.7) and 21 phenotypic traits over 274 grapevine accessions in 2013, and 207 accessions in 2010 and 2012. It is to remember here that for the “supplementary set”, genotypic data were available only for 109 SSRs (Table A.7). Prior to the analysis, alleles with less than 1% frequency were set as missing.

Association analysis was conducted for each season separately using the GLM and Q+K models implemented in the software package TASSEL 4.0 (Bradbury *et al.*, 2007) as described for SNP-based GWAS above. The Bonferroni corrected genome-wide significance threshold was found highly conservative, so the nominal significance threshold $P = 0.01$ was also considered.

RESULTS

4. RESULTS

4.1. PHENOTYPIC DATA

Our initial objective was to phenotype the mapping collection during three growing seasons from 2010 to 2012. However, in 2011, a severe pathogenic attack of downy mildew (*Plasmopara viticola*) caused the loss of almost all the production, impeding hence the assessment of phenotypic traits beyond flowering. This was most likely the consequence of unusual climatic conditions characterized by the coincidence of mild temperatures (about 20°C) and uninterrupted strong precipitations (86.2 mm in five days) in May 2011 (Fig. 4). These conditions led to a dramatic proliferation of the pathogen throughout the vineyard in only a couple of days, destructing all the vegetative biomass as well as the developing clusters. As a consequence, in this year, only budbreak, flowering and fertility were assessed. For simplicity of data presentation and discussion, the three phenotypic traits evaluated in 2011 were arranged within the 2010 data as the equivalent ones in the latter were not evaluated. In order to compensate for the 2011 loss, traits assessment was extended to 2013. On the other hand, although the preventive measures such as net fencing, trapping and cluster wrapping helped to protect the grapes, we recorded some damages mainly on early ripening cultivars, destroying sometimes all the clusters produced by these cultivars. These damages were most of the time caused by birds, raccoons and wild boars. The major losses were recorded during the first year of study (2010), where 22 cultivars remained practically without grapes to characterize.

Figure 4. Monthly Cumulative rainfall (a) and average temperatures (b) recorded at the El Encín weather station during the four years of study

4.1.1. Traits distribution

The distribution of the evaluated traits across the three considered growing seasons (2010, 2012 and 2013) is shown on Fig. 5. The qualitative traits, flower sex, berry skin color and muscat flavor were stable over years, so their respective data recorded during the three years of study were merged in a single consensus dataset. As can be seen in the histograms and deduced from the normality test, several quantitative traits presented deviations from normal distribution. This was the case of the phenological traits, particularly budbreak and flowering times. Cluster weight also showed significant deviation from normality and was remarkably skewed towards small clusters. In the same sense, berry weight was skewed towards small berries and shape index was skewed towards round berries; whereas number of seeds per berry presented a seemingly bi-modal distribution because of the presence of seedless cultivars. These traits were square root and/or log transformed prior to association analysis. However, normal distribution was not always recovered under these transformations. The distribution of flower sex, flavor and seed content was clearly imbalanced: among all the accessions, only 4.98% were female and all the rest were hermaphrodite; only 8.71% were muscat flavored; and seedlessness was recorded for 7.05% of the accessions and all the rest were seeded. Obvious differences were observed among the three years of evaluation for fertility and phenological traits, especially budbreak and flowering times. Average fertility was relatively low in 2012 (1.08 ± 0.54 inflorescences/shoot), intermediate in 2010 (1.25 ± 0.51) and significantly higher in 2013 (1.77 ± 0.44). Regarding phenological traits, grape cultivars reached in average the budbreak stage earlier in 2010 (95 days since January 1st), followed by 2013 (100 days) then 2012 (111 days), whereas the order was 2010 (142 days since January 1st), 2012 (155 days), then 2013 (163 days) for flowering time.

Figure 5. Phenotypic traits distribution within the mapping collection across three growing seasons

Asterisks within figure's legends indicate the significance level of the Kolmogorov-Smirnov's normality test; ⁿ: non-significant, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

4.1.2. Phenotypic correlations between years and between traits

In order to assess whether the phenotypic data recorded in different years are consistent with each other, we analyzed the pairwise correlations between the three years of study for the quantitative and ordinal traits (Table 2). The most consistent traits were those related to berry size and shape (mean berry weight (MBW), mean berry length (MBL), mean berry diameter (MBD) and berry shape index (ShI)), which showed *Rho* values ranging from 0.79 to 0.89. Veraison and ripening times also showed high correlation between 2012 and 2013 (0.88 and 0.85 respectively). No information could be obtained about the behavior of these traits in 2010 due to the lack of the data in this year. However, the other two phenological traits (budbreak and flowering) were less consistent ($0.54 \leq Rho \leq 0.77$). Cluster compactness (CI-OIV) and flesh firmness, showed upper-intermediate correlations ranging from 0.50 to 0.64; better than the measurable traits: cluster diameter ($0.43 \leq Rho \leq 0.57$), peduncle length ($0.49 \leq Rho \leq 0.56$) or seed content traits (MNS and MSDW) that showed *Rho* values ranging from 0.44 to 0.69.

Table 2. Pairwise Spearman correlations between the three years of study for each trait

Trait	Abbreviation	Year pairs		
		2010-2012	2010-2013	2012-2013
Time of budbreak	Bb	0.67	0.54	0.63
Time of flowering	Fl	0.75	0.77	0.68
Time of veraison	Vr	-	-	0.88
Time of ripening	Rp	-	-	0.85
Fertility	Frt	0.64	0.62	0.71
Mean cluster weight	MCW	0.73	0.64	0.69
Mean cluster length	MCL	0.69	0.72	0.72
Mean cluster diameter	MCD	0.53	0.57	0.43
Mean peduncle length	MPL	0.49	0.56	0.51
Cluster compactness (OIV)	CI-OIV	0.56	0.50	0.62
Cluster compactness (Tello et al, 2014)	CI-12	0.55	0.60	0.54
Mean berry weight	MBW	0.87	0.89	0.87
Mean berry length	MBL	0.88	0.86	0.88
Mean berry diameter	MBD	0.84	0.83	0.84
Berry shape index	ShI	0.84	0.79	0.82
Flesh firmness index ^b	FI	0.55	0.59	0.64
Mean number of seeds per berry	MNS	0.50	0.52	0.56
Mean seed dry weight per berry	MSDW	0.44	0.49	0.69

Phenotypic correlations between traits were also examined (Fig. 6). The most important points to be highlighted from this correlation matrix are: 1) With the exception of the correlation between veraison and ripening ($Rho = 0.83$), the correlation between other phenological stages was weak ($0.24 \leq Rho \leq 0.44$). 2) Fertility was negatively correlated with phenological traits ($-0.18 \leq Rho \leq$

-0.41), and so was with cluster as well as berry size traits ($-0.22 \leq Rho \leq -0.41$). 3) Cluster weight was highly correlated with cluster dimensions (length and diameter), although the transversal axis ($Rho = 0.74$) seems to be more determinant than the longitudinal one ($Rho = 0.65$); additionally, cluster dimensions were correlated with each other ($Rho = 0.60$). 4) Cluster compactness as evaluated using the OIV subjective descriptor showed high correlation ($Rho = 0.71$) with the CI-12 compactness index calculated as a relationship between cluster weight and length (Tello and Ibáñez, 2014); moreover, the last one was negatively correlated with peduncle length ($Rho = -0.52$). 5) While berry weight was strongly and equally correlated with berry length and diameter ($Rho = 0.9$), berry shape was more correlated with the length of the longitudinal axis ($Rho = 0.59$) than the transversal one ($Rho = 0.25$). 6) The highest correlations showed by flesh firmness were with berry size traits (0.57 to 0.62). 7) Seed content traits, particularly, mean number of seeds per berry (MNS), were weakly correlated with berry size traits ($0.19 \leq Rho \leq 0.26$).

Figure 6. Pairwise Spearman correlations between the studied traits

The complete names of the traits are shown in Table 1. Correlations between traits reported here are from 2013 data. Correlations in other years are similar, so they are not presented.

The scale of color shading is given at the upper-right side of the figure. All the correlations are significant at $P = 0.05$, the non-significant values are not reported.

4.2. GENOTYPIC DATA

4.2.1. SNP genotyping

4.2.1.1. Library sequencing, SNP calling and filtering

The four sequencing lanes (three lanes for the whole collection and one lane for the 85 resequenced accessions) yielded a total of 661,089,485 bar-coded 100-bp sequence reads across the 274 grapevine accessions. No significant differences in the number of reads were observed among the different sequencing lanes of the same library. However, the number of reads was extremely variable among the sequenced accessions and almost continuously distributed (Fig. 7a and Table A.1). In the first sequencing lane for example, read counts ranged between 2,935 for Rkatsiteli cultivar and more than 4.5 million reads for Clairette de Gascogne, with a mean of 81,646.46 reads/accession and a standard deviation of 928,287.35 (coefficient of variation: 1,137%). In order to improve the uniformity of sequence coverage, the 85 underrepresented accessions were sequenced again using libraries constructed from newly extracted DNA.

A total of 358,454 biallelic SNPs (large structural variants, indels and multiallelic SNPs were not called) were identified for the 274 accessions, with an overall mean sequencing depth per SNP of 7.44 ± 7.25 . Approximately 73% of the called genotypes had a genotype quality score $GQ \geq 98$ (Fig. 7b). The SNP call rate per accession was found highly variable (Fig. 7c) and substantially proportional to the rate of read coverage (Fig. 7d). This figure (7d) shows also that SNP call rate increases exponentially with the increase of read coverage until approximately two million reads per accession, where the proportionality ratio tends to reach the plateau.

In order to generate a consistent and reliable SNP dataset to use for further analysis, the raw called SNP data were submitted to a stratified filtering process as illustrated in Fig. 3. Because of the highly heterozygote nature of the grapevine genome, SNP calling is error-prone due to heterozygote under-calling. However, the susceptibility to this kind of errors is significantly reduced for the genotypes called with a minimum GQ score of 98. Therefore, all genotype values in the dataset with GQ scores lower than 98 were filtered out. We observed that the two-dimensional data matrix (SNPs x individuals) was drastically reduced as the dataset was filtered based on missingness per SNPs (GENO) and/or per individuals (MIND) (Table 3). Given the relatively large number of identified SNPs (358,454) and the small sample size (274 accessions), we applied more stringent filters on the missingness per SNP (0.75) than on missingness per individual (0.1). These filters resulted in the elimination of 31 accessions and 300,858 SNPs (83.93% of the total).

Figure 7. Distribution of GBS read coverage, genotype quality scores and SNP call rates within the mapping collection

a. Distribution of read counts across the 274 sequenced accessions, **b.** Distribution of genotype quality scores (GQ) within a random sample of 748,209 genotype values (274 individuals x 2,832 SNPs), **c.** Distribution of SNP call rates across the 274 sequenced accessions, **d.** Relationship between read coverage and SNP call rate within the 274 sequenced accessions.

Table 3. Pattern of decline of the SNP dataset dimensions after filtering by combination of minimum call rate per individuals and by sites

		MIND thresholds					
		0.0	0.1	0.2	0.3	0.4	0.5
		273	242	198	122	76	34
GENO thresholds	0.0	358,454	358,454	358,454	358,454	358,454	358,454
	0.5	96,260	109,387	125,087	154,316	178,379	206,271
	0.6	74,521	87,845	104,654	134,032	158,546	185,780
	0.7	51,054	68,325	85,149	115,711	138,736	169,425
	0.8	25,606	46,689	64,867	95,175	120,727	145,634
	0.9	7,058	21,718	39,346	70,116	94,745	123,034
	0.99	61	3,826	6,949	24,998	41,465	80,566

Grey background: number of individuals. White background: number of SNPs. MIND: minimum call rate per individual. GENO: minimum call rate per site.

A large proportion of SNPs with low minor allele frequency (MAF) was found in the dataset (Fig. 3). For instance, 36.91% of the SNPs in DS3 had $MAF < 0.01$ and 54.92% with $MAF < 0.05$. Based on MAF filtering, four datasets, DS4, DS5, DS6 and DS7 containing SNPs with minimum MAF higher than 0.01, 0.05, 0.1 and 0.2 respectively were generated.

A comparison between the raw dataset (DS0) and the filtered dataset (DS5) is presented on Fig. 8. Although considerable data were lost by filtering, the quality of the remaining was considerably improved. For instance, while the average SNP mean depth was raised from 7.44 ± 7.25 in DS0 to 28.49 ± 12.04 in DS5, the overall proportion of missing values was reduced from $39\% \pm 12\%$ to $12\% \pm 10\%$. Additionally the mean minor allele frequency of SNPs was improved from 0.15 ± 0.12 in DS0 to 0.21 ± 0.11 in DS5. The accuracy of the filtered dataset was also confirmed by checking Mendelian inheritance in family pedigrees (Table A.3). The overall estimated SNP calling error rate (based on Mendelian errors) was 0.35%. However, since not all genotyping errors will be detectable as Mendelian errors, the reported error rate could be somewhat underestimated.

4.2.1.2. SNP density and genetic diversity

The dataset DS5 (Fig. 3) that consists of 25,961 SNPs across 242 individuals with an average call rate of $88.13\% \pm 6.26\%$ and minimum MAF of 0.05 was explored to visualize density distribution across the genome and estimate some diversity parameters. Among the 25,961 SNPs, only 433 were located on unmapped contigs. The remaining 25,528 SNPs were located on the 19 chromosomes of grapevine (Fig. 9 and Table A.4). Across all the genome, the smallest distance between two contiguous SNPs was 1 bp (in multiple sites and chromosomes) and the largest one was 1.98 Mb in chr18. The number of SNPs per chromosome was highly correlated with chromosome size (coefficient of correlation: 0.92; Fig. 10). Moreover, SNP density towards the chromosome ends was higher compared to the median regions (Fig. 9), reflecting substantially the distribution of gene density (Fig. A.1).

The average observed heterozygosity of the DS5-SNPs across the 242 accessions ranged between 0.00 and 0.99, with an average of 0.32 ± 0.15 , slightly higher (difference = 0.02) than the predicted heterozygosity that varied between 0.08 and 0.50 with an average of 0.30 ± 0.13 , confirming the high degree of heterozygosity in grapevine (Fig. 11 and Table A.4).

Figure 8. Comparison of depth of coverage, proportion of missing values and minor allele frequency between the raw (DS0, left) and the filtered (DS5, right) GBS datasets.

Figure 9. Distribution of GBS-SNPs across the 19 grapevine chromosomes

The SNPs presented here are from the dataset DS5 that consists of 25,961 SNPs with minimum MAF of 0.05 (see Fig. 3 for more details); chromosome identifiers are given at the left and number of SNPs per chromosome are given at the right side of the figure. Each bar in a given chromosome represents one SNP. SNP positions are given in Mb.

Figure 10. Relationship between number of GBS-SNPs per chromosome and chromosome size

Figure 11. Distribution of observed (H_{obs}) and expected (H_{exp}) heterozygosities in the SNP dataset DS5

4.2.2. SSR genotyping

4.2.2.1. Amplification tests

Six loci (FAM18, FAM68, FAM96, FAM104, VVIB54 and VVIV22) out of the 249 selected initially detected only one fragment in the testing collection, suggesting that they could be monomorphic. Even though, they were included in the designed multiplex sets since more alleles could be detected when the whole mapping collection will be genotyped. Three of them were confirmed as monomorphic (FAM18, FAM104 and VVIB54), whereas the rest (FAM68, FAM96 and VVIV22) detected only one allele more in the core collection. Even though these markers were maintained in the multiplexes because they can be polymorphic in genetic backgrounds of wider diversity. Eleven markers amplified more than two fragments per accession, they were hence classified as multilocus (Table A.7). Six of them (UDV-020, UDV-038, VMC3G7, VMCNG1E4-2, VMCNG1D3 and VMC4B7-2) were included in the designed multiplex PCRs and analyzed in the core collection, and an attempt was done to score each marker as two different loci. However, it was not possible to discriminate between the amplified fragments corresponding to each locus because they were located in many cases in the same narrow fragment size interval (Fig. 12). Even though, they were maintained in the multiplexes since they can be useful for genetic mapping or breeding purposes in biparental populations, where allele screening could be less complicated. The locus VVS3 was also discarded because it generated irreproducible amplification profiles (repeated PCRs produced fragments of different sizes).

Figure 12. Examples of electrophoretic profiles showing the possible amplification of multiple loci in a single genotype by the SSR markers VMC3G7 and UDV-020

Grey stripes represent the bin set (possible alleles in the core collection).

4.2.2.2. Multiplex PCRs design, optimization and validation

Two hundred and thirty six SSRs out of the 243 that passed the amplification tests were distributed in 34 multiplex sets (multiplexes from Mx01 to Mx34 in Table 4, and Table A.5). The seven remaining SSRs could not be assigned to any multiplex because they lead to amplicons overlapping. PCR conditions were optimized by adjusting primer concentrations and thermocycling parameters. For each multiplex, two to eight (3.8 in average) optimization rounds were needed to reach a balanced amplification between the primer pairs combined in the same reaction. Final protocols involved the use of primer concentrations ranging from 0.04 to 0.60 μ M (Table A.5) and four touchdown based thermocycling programs (Tp-A, Tp-B, Tp-C and Tp-D; Table A.6). Once optimized, these multiplex PCRs were used to genotype the core collection. Inheritance analysis using the 14 trio pedigrees included in the core collection revealed inconsistencies in 37 loci. Thirty five of them showed incompatible genotypes in at least one of the pedigrees (Table A.7) that were likely due to the presence of null alleles in the progenitors (i.e. observed “bb” offspring derived from “aa x bb” or “aa x bc” crosses). The unique exceptions were VMC5G7 and B003 that showed genotype inconsistencies in only one pedigree, but not fitting the presence of null allele pattern. These two cases may be the result of somaclonal mutations that are not rare in grapevine (Pelsy, 2010). In an attempt to avoid null alleles and recover monolocus segregation, alternative primers were designed for 37 problematic markers (32 suspected of carrying null alleles, 4 multilocus (UDV-021, FAM62, VMC1E11 and UDV-134) and VVS3). Successful single locus amplifications were obtained for 25 of them (Table A.7).

Table 4. Microsatellite markers included in each multiplex PCR

Multiplex	marker 1	marker 2	marker 3	marker 4	marker 5	marker 6	marker 7	marker 8	marker 9	marker 10	marker 11	marker 12	marker 13	marker 14	marker 15
Mx01	VVIP60	VVIB01	VRZAG83	VRZAG79	VVMD7	VRZAG62	VMC1B11	VVIQ52	VVMD25	VVS2	VVIH54	VRZAG112	VVMD24	VVIN73	VVIP31
Mx02	VRZAG29	VVMD28	VVMD32	VVMD27	VVMD21	VRZAG67	VVIV37	VMC4F3-1	VVIV67	VVMD5	VVIN16				
Mx03	VMC4F8	VMC5G7	VVMD27	VVIM63	VMC3E5	VMC2B1.1	UDV-005	VMC7f2-2							
Mx04	VVIQ57	VVIN61	VChr1b	VChr5c	VVIT68	UDV-053	VMC9B5	VVIN40	UDV-117	VVIN16					
Mx05	VMCNG2g7	UDV-109	VVC6	VRZAG79	VVIN33	VMC4C6	VMC1E8	VVIP05							
Mx06	VVIB01	VMCNG1D3	VChr5b	VMC3C7	UDV-077	VVIN62	VVIV08	VVIQ44							
Mx07	VVIP12	VMC9D3	VChr2c	VMC7g3	VMCNG2E1	VVMD21	VVS16	VMC2A3	VVIN74						
Mx08	VChr1a	VMC2C10-1	VV6718	VChr9a	VVS2	VVIP02	VMC4D9-2	VMC8G3							
Mx09	VRZAG29	VMC8F10	VMC2F10	VChr8a	VMC9H4_2	UDV-046	VChr19a								
Mx10	VMC2B3	Vchr4a	VVIM43	VMC1F10	VChr13c	VChr13b	VVMD5								
Mx11	VVIN54	VRZAG21	VVIQ67	VVIV37	GT183H07	VMCNG1E1	VVIS70	VVIQ61	VMC7B1						
Mx12	VMC4D4	VV5799	VVMD21	VVIP33	VChr16a	VMC4B7-2	VMC2H3	VChr18c	VMC5E9						
Mx13	VVIN75	VChr7a	VVIV04	VChr8b	VMC1B11	VMC2H10	VMC2H4	VMC3G7							
Mx14	VMC3B10	UDV-093	VRZAG83	VMC5G1-1	VVIS58	VVIM42A									
Mx15	UDV-022	VChr3a	VMC8B11	VMC16D4	VVMD7	VVIN70	VMC8B5	VChr18b							
Mx16	VVIQ66	UDV-040	VVIH01	VVIV67	VVIM93	VMC8F4.2	VChr18a								
Mx17	VMC9F2	VChr5a	VVIN31	VVIB66	VMC1E12	UDV-062	VVMD32								
Mx18	VMC1G7	VVIB22	VMC2E11	VMC2D9	VVMD25	UDV-100	VMC3D12	VVIP22	VVIM03	UDV-020					
Mx19	VVIP60	VVIP77	VChr7c	VMC5H2	VMC3E12	VVIB19	VVIM11	VVIN52	B003	VVIV33					
Mx20	VVIO61	VVIP72	VVIN56	VMC7h2	VMC1C10	VChr16b	VVIN73	VVIV22							
Mx21	VMC1A5	VVMD28	VMC1A12	VMC2A10	UDV-124	VMC8E6	VMC2H12	VMCNG1E4-2							
Mx22	VVIB72	VRZAG62	UDV-059	VMC3C11-1	VRZAG15										
Mx23	VVIB01	VMC8C2	UDV-063	VMC2A12	VMC6G1	VMC8G9	VVIB54								
Mx24	VVIO02	VMC6g10	VVIQ52	VVIM01	UDV-095	VVIN94	VMC5G8	VChr17a							
Mx25	VVMD27	UDV-090	UDV-085	VMC4G6	VMC8D11	VMC2B11	UDV-038								

Multiplex	marker 1	marker 2	marker 3	marker 4	marker 5	marker 6	marker 7	marker 8	marker 9	marker 10	marker 11	marker 12	marker 13	marker 14	marker 15
Mx26	VMC7h3	VVIN78	VVIP10	UDV-050	VVMD24	UDV-032									
Mx27	VVIF52	VMCNG2H2.2	UDV-026	VVMD25	UDV-058	UDV-116									
Mx28	VMC2H9	VMC7A4	UDV-075	VChr12a	VMC4F3-1	VVC34	A105								
Mx29	VMC6d12	VMC5C1	VChr11b	VMC8G6	TT251F02	VVIV67									
Mx30	FAM140	FAM110	FAM96	VVIQ52	FAM63	VMC1G3-2	VChr15a	UDV-104							
Mx31	FAM141	FAM91	FAM12	FAM45	VRZAG112	FAM18									
Mx32	FAM68	VVIQ25	FAM122	FAM116	FAM70	VChr17b	VVIP31	VChr19b							
Mx33	FAM24	VVIU20	VRZAG89	FAM42	VVIQ52	VVMD8	FAM135	p3_VvAGL11	UDV-023						
Mx34	UDV-110	VVIC50	VVIV37	FAM92											
Mx35	VChr1a-2	FAM140-2	VVMD25	TT251F02-2	VChr15a-2										
Mx36	VVIV67	VMC8G3-2	UDV-009	VChr18a-2	UDV-062-2										
Mx37	VRZAG29	VMC2D9-2	VChr17b-2	VChr19a-2	VVIM03-2										
Mx38	VMC16D4-2	VVIP72-2	VRZAG62	VVIS58-2											
Mx39	VVS3-2	VVIN75-2	VChr8a-2	VVIH54	FAM70-2										
Mx40	VVIQ25-2	VVIB22-2	VVIV37	VMC4H2	VVIS70-2										
Mx41	VVIM25	RTAChr3-17a	RTAChr3-19b	RTAChr4-9a	RTAChr13-20a	RTAChr15-4a	VVIV67	VVIP08	RTAChr19-11b						
Mx42	RTAChr1-14a	RTAChr3-15a	FAM65	VVIQ52	FAM62-2	RTAChr15-3b	RTAChr16-6b	RTAChr17-15a							
Mx43	RTAChr2-11a	UDV-021-2	VRZAG67	RTAChr13-13b	RTAChr15-1b	RTAChr15-7a	VMC1E11-2	RTAChr19-16b							
Mx44	VVIV36	RTAChr9-13a	VRZAG67	RTAChr13-16a	RTAChr17-15b	UDV-134-2									
Mx45	VVMD27	RTAChr7-9	FAM104	RTAChr11-11b	RTAChr13-23b										

SSR markers were unavailable in the literature for several chromosomal regions of the grapevine genome, leaving uncovered gaps of up to 10 Mb according to the 12x.2 genome sequence (Fig. 13; Canaguier *et al.* 2014; <https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences>). Although many microsatellite motifs were detected in these genomic sequences, it was difficult to find specific primers for their amplification. As an example, in a segment of 1 Mb from chromosomes 3 (between 13.5 and 14.5 Mb) and chromosome 15 (between 2.5 and 3.5 Mb) 91 and 153 microsatellites were identified, respectively. Among them, only 54% yielded candidate amplification primers, and their alignment against the grapevine genome sequence using Primer-BLAST revealed that most of them (90 and 81% respectively) had more than one matching site (i.e. susceptible to unspecific amplification). Finally, a total of 22 new SSR markers could be developed in those gaps and successfully amplified in single locus PCRs (Fig. 13 and Tables A.3 and A.5).

The 25 redesigned SSRs, the 22 newly developed and the 7 remaining unassigned after the construction of the first set of multiplex PCRs were organized in a second set of 11 multiplexes (from Mx35 to Mx45 in Table 4 and Table A.5). The examination of genotypic data obtained with these multiplexes in the trio pedigrees revealed that allelic inheritance was incompatible in three out of the 25 redesigned markers and 10 out of the 22 new ones, being the incompatibility pattern indicative of the presence of null alleles (Table A.7).

In summary, 290 SSR markers (243 selected initially, 25 redesigned and 22 developed in this study) were successfully organized in 45 multiplex PCRs with an average of 7.31 primer-pairs per reaction, ranging from 4 to 15-plex (Table 4; Fig. 14 and 15a). We should remember here that these 290 markers correspond to 270 unique loci, since the targeted by 20 out of the 25 redesigned ones were represented twice in the multiplexes, with the original primers in the first set (Mx01 to Mx34) and with the new primers in the second set (Mx35 to Mx45). All these loci except the six multilocus ones (see section 4.2.2.1.) were successfully genotyped in the mapping collection (Fig. 13).

In order to control for manipulation errors, all the multiplex PCRs from Mx03 to Mx45 included a “control locus” from Mx01 or Mx02 (Table A.5). Additionally, in order to verify the consistency of allele genotyping (“binning” process and allele calling), 20 accessions were genotyped twice by each marker. The control loci included in the multiplexes from Mx03 to Mx45 produced the same genotypes as the initial ones obtained using Mx01 or Mx02. In the same way, identical genotypes were observed for each pair of the duplicated samples at all the loci. New alleles not found neither in the testing panel nor in previous studies were detected for most of the loci, even though allele size overlapping was detected only between the markers VRZAG112 (allelic range 228-260 bp) and VVIN73 (254-267 bp) in Mx01. However, only one rare allele from VRZAG112 (260 bp with 0.029 allele frequency) was implicated, and its electrophoretic profile was clearly distinguishable from the VVIN73 one.

Figure 13 (previous page). Distribution of the 264 genotyped SSR loci along the grapevine genome

The six multilocus SSRs that were not genotyped in the mapping collection are not presented in this map. The chromosomal positions (in Mb) were determined by Blast-search against the 12x.2 *Vitis* genome assembly (Canaguier *et al.* 2014 <https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences>) using the primer sequences as probes. Vertical bars represent the 19 grapevine chromosomes. Red segments indicate the chromosomal gaps that were investigated for the development of new markers therein located (in blue). It should be noted here that the selection of markers and the development of new SSRs to cover the chromosomal gaps were based on the previously available version of the grapevine genome sequence (12x; <http://www.genoscope.cns.fr/>) from which inter-loci distances and ordering have undergone several important changes.

Figure 14. Example of PCR profile obtained with the multiplex PCR Mx01

FAM-, NED-, PET- and VIC-fluorescence labeled SSR markers are indicated in blue, black, red and green lines respectively. Length of the rectangles below the electropherogram represents allelic ranges of each locus.

4.2.2.3. Genetic diversity and quality of the studied SSR markers

The 264 unique loci (Fig. 13) identified in the core collection a total of 2,760 alleles, ranging between 1 and 31 per locus (10.45 on average), and most of them (92%) detected more than four alleles (Fig. 15b and Table A.7). Allelic sizes ranged between 56 and 456 bp (Fig. 15c and Table A.7). The difference between the smallest and the largest fragment detected at each polymorphic locus ranged between 2 bp (VVIV22, VVIB72 and VVIN78) and 162 bp (VMC5G1-1), with an average of 36.60 bp. Exceptionally, the locus Vchr8b amplified fragments ranging from 100 to 455 bp (difference of 355 bp). The allele sizes obtained for all the genotyped SSRs on a subset of 25 well-known and accessible accessions, which could be used as reference samples in other studies, can be found in Online Resource 1: Table S5 in Zarouri *et al.*, (2015). The estimated null allele frequency (F) for the 264 SSRs varied between -0.145 and 0.696, and 87% of the loci showed an $F < 0.1$ (Fig. 15d and Table A.7). The mean values of expected (H_{exp}) and observed (H_{obs}) heterozygosity were 0.699 and 0.658, ranging from 0.000 to 0.931 and from 0.000 to 0.928, respectively (Fig. 15e and Table A.7). The average polymorphic information content

(PIC) was 0.665 (0.000 - 0.925), and the average probability of identity for unrelated individuals was 0.15 (0.009 - 1.000) (Fig. 15e and Table A.7). Beside its genetic diversity, the usefulness of a molecular marker is also determined by the pattern of its amplification profile that determines in turn the easiness of allele scoring. The characteristics of the amplification patterns showed by each SSR (stuttering patterns, “+A” effect, low heterozygote peak ratio, presence of artifacts) and a global evaluation of the easiness of scoring of the marker are shown in Table A.7.

Figure 15. Distribution of multiplexing level and genetic diversity parameters of the 264 genotyped SSR loci

a: Number of loci per multiplex PCR. **b:** Number of alleles per locus. **c:** Size of the 2,760 detected fragments. **d:** Estimated frequency of null alleles. **e:** Observed (H_{obs}) and expected (H_{exp}) heterozygosity, polymorphic information content (PIC) and probability of identity (PI).

4.3. POPULATION STRUCTURE AND RELATEDNESS

Population structure analysis was carried out in order to fulfil three objectives: 1) to know whether the mapping collection used here present population stratification and how this stratification is linked with cultivar utilization and geographical origin; 2) to investigate how the population structure affects the distribution of the studied traits; and 3) to use the inferred sub-population membership probability matrix (Q matrix) in the association analysis models to correct for population structure confounding effect. The method implemented in the software used to study the structure of the mapping collection (STRUCTURE v 2.3.4; Pritchard *et al.*, 2000) allows only an ad hoc approximation to the number of populations (K). Furthermore, the biological interpretation of K might not be straightforward. So the best strategy is to focus on the K value that captures the majority of the structure present in the data but that is also biologically meaningful. Evanno *et al.* (2005) described a method for selecting the appropriate K value from the data, based on the second order rate of change of the likelihood function with respect to K. The Evanno criterion (ΔK) resulted in $K = 3$ as the most plausible number of subpopulations (Fig. 16). Among the total of 274 accessions of the mapping collection, 149 presented a probability of membership $Q \geq 0.75$ and they were assigned to one of the three subpopulations (q1, q2 and q3; Table A.1). The rest (125 accessions, with $Q < 0.75$) were assigned to the admixed group. The three reconstructed subpopulations were:

- **q1** (48 accessions): this cluster was composed almost exclusively of wine cultivars from Western Europe, mostly France (17 cultivars) and Spain (17 cultivars). Only two table cultivars from France (Precoc de Malingre and Roi des Precoces) pertained to this cluster.
- **q2** (56 accessions): this group was composed 35 cultivars from Spain represented by 12 table cultivars, 19 wine cultivars and 4 mixt-use cultivars; 8 seedless cultivars and other table cultivars mostly from the oriental countries.
- **q3** (45 accessions): this cluster was dominated by a group of 18 table cultivars from the Italian Peninsula, and wine cultivars mostly from the Balkan and Eastern Europe. It is to highlight that many cultivars in this group are new breeds.

The analysis of variance (ANOVA) and the comparative distribution of qualitative and ordinal traits revealed significant variation of the assessed phenotypic traits between the subpopulations (Fig. 17). Importantly, cultivars from q1 were more fertile, early flowering and ripening and produced small compact clusters with small and soft-flesh berries, compared to q2 and q3 cultivars that were relatively less fertile, late ripening, and have larger clusters with larger and firmer berries. Moreover, muscat flavored cultivars were concentrated in the q3 subgroup.

Figure 16. Population structure of the mapping collection used in this study

a. Variation of the Evanno criterion (ΔK) among the assayed clustering levels (K). **b.** Grouping of the individuals according to the most likely clustering level ($K = 3$); each individual is represented by a thin vertical line partitioned into three colored segments whose length is proportional to the probability of membership to each cluster.

Fig. 1. Quantitative and ordinal traits for the three subpopulations determined by STRUCTURE
 The plots show quantitative and ordinal traits for three subpopulations (q1, q2, q3). Quantitative traits are represented using boxplots, and ordinal traits are represented using stacked bar-plots. The grey line within the boxplots represents the grand mean. The significance level of ANOVA tests are given next to the title of each plot. "": non-significant, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Subpopulations not connected by the same letter are significantly different according to Tukey-Kramer HSD test at $p < 0.05$.

In addition to population structure, a finer scale relatedness matrix (kinship or K matrix) among all accession pairs within the mapping collection (Fig 18) was calculated in order to be used to correct for cryptic relatedness confounding in the association analysis models. Among the total pairwise kinship coefficients between the 274 individuals of the mapping collection, 87.89% were lower than 0.05 (of which 63.89% < 0), indicating absence of relationship among the involved individuals. The rest ranged from 0.05 to 0.5 with continuous decreasing frequencies

indicating the complex pairwise relationships existing among the implicated grapevine cultivars. However, the proportion of kinship values of 0.25 or higher did not exceed 0.25%, meaning that first degree relationships (parent-offspring and full siblings or similar) are rare among the individuals used in this study.

Figure 18. Distribution of kinship coefficients among the 274 individuals of the mapping collection

4.4. LINKAGE DISEQUILIBRIUM ANALYSIS

4.4.1. LD analysis using SNP markers

4.4.1.1. Background and linked LD distribution

The extent of background LD was estimated from a random sample of 994,181 unlinked SNP pairs (from different chromosomes) using Haploview 4.2. The values of r^2 between unlinked SNPs ranged between 0.000 and 0.339 with an average of 0.012 ± 0.011 (Fig. 19). Only 1.21×10^{-3} per cent of the unlinked SNP pairs presented r^2 values ≥ 0.2 .

LD between linked SNPs was estimated for each pairwise combination of SNPs on each chromosome (12,677,030 SNP pairs in total). More than 2% of linked SNP pairs across the whole genome presented r^2 values ≥ 0.2 (Table 5). Estimated distances of LD decay below the r^2 threshold of 0.2 for each chromosome and for the whole genome are shown in Table 5. LD decayed rapidly with increasing physical distances between SNPs (Fig. 20a and b, and Fig. A.2). The mean r^2 values across the genome started from a maximum of approx. 0.6 for SNPs less than 10 bp apart, and declined below a critical value of 0.2 within a distance of 0.94 kb. The average r^2 at distances higher than 10 Mb ($r^2 = 0.013$) was very close to the average r^2 among non-linked SNPs (0.012). Significant variation in the extent of LD was observed among the 19 grapevine chromosomes (Table 5), independently of SNP density (considered as the average SNPs inter-distance, Table 5). As an example, on chromosome 14, r^2 decayed below 0.2 within a distance of

0.03 kb, and on chromosomes 5 and 12 within 0.06 kb. However, on chromosomes 2 and 17 respectively it extended up to 4.29 and 6.79 kb. This variation was particularly apparent at closer distances (< 10 kb; Fig. 20a). Additionally, average r^2 values across the chromosomes showed negative correlation with chromosome size (Coefficient of correlation: - 0.77; Fig. 20c).

Figure 19. Distribution of background LD as estimated using a sample of 994,181 unlinked SNP pairs over 242 genotypes.

Table 5. Extent of SNP-based LD across the grapevine genome and estimated number of SNPs for whole-genome covering.

Chr.	Chr. size (Mb)	Number of SNPs (DS5)	Average inter-SNP distance (Kb)	r^2 threshold = 0.2		
				% $r^2 \geq 0.2$	Distance of LD decay (kb)	Number of SNPs for GWAS
chr1	24.23	1,396	17.36	1.96	0.28	86,535.71
chr2	18.89	855	22.10	3.26	4.29	4,403.26
chr3	20.70	1,159	17.86	1.99	0.21	98,571.43
chr4	24.71	1,314	18.81	2.22	0.52	47,519.23
chr5	25.65	1,764	14.54	1.38	0.06	427,500
chr6	22.65	1,318	17.18	1.83	0.77	29,415.58
chr7	27.36	1,545	17.71	1.67	0.19	144,000
chr8	22.55	1,415	15.94	2.10	1.45	15,551.72
chr9	23.01	1,161	19.82	1.78	0.13	177,000
chr10	23.50	1,084	21.68	1.94	0.25	94,000
chr11	20.12	1,099	18.31	2.85	1.59	12,654.09
chr12	24.27	1,390	17.46	2.18	0.06	404,500
chr13	29.08	1,613	18.03	1.68	0.18	161,555.56
chr14	30.27	1,781	17.00	1.25	0.03	1,009,000
chr15	20.30	1,112	18.26	2.06	0.38	53,421.05
chr16	23.57	1,075	21.93	2.15	0.13	181,307.69
chr17	18.69	970	19.27	3.13	6.79	2,752.58
chr18	34.57	2,136	16.18	1.31	0.14	246,928.57
chr19	24.70	1,341	18.42	2.20	0.36	68,611.11
All chr.	458.82 ^a	25,528 ^a	18.31 ^b	2.05 b	0.94 ^b	3,265,227.58 ^a

The number of SNPs required to cover each chromosome was estimated by dividing the chromosomal length (according to the 12x.2 version of the reference genome sequence) by the LD decay distance. ^a Column sum.

^b Column average.

Figure 20. Distribution of LD within chromosomes as estimated using a set of 25,528 SNPs over 242 genotypes.

a: Average r^2 values at different categorical physical distances across the genome. **b:** LD decay plot within chromosome 1 (the rest are shown in Fig. A.2), the red line represents the logarithmic trend line. **c:** Relationship between average LD per chromosome and chromosome size, the orange line depicts the tendency line.

4.4.1.2. Effect of MAF, sample size and population structure on the extent of LD

The effect of MAF on LD was investigated using the datasets DS4, DS5, DS6 and DS7 that contains SNPs with minimum MAF of 0.01, 0.05, 0.1 and 0.2, respectively (Fig. 3). MAF affected strongly LD estimates (Fig. 21a): the mean r^2 values increased proportionally with the increasing minimum MAF. This trend was especially apparent on short to medium distances (<1Mb). As an example, the mean r^2 value for SNPs less than 10 bp apart was 0.48 when estimated using the dataset DS4 ($\text{MAF} \geq 0.01$), but it was as high as 0.83 when estimated using DS7 ($\text{MAF} \geq 0.2$).

The effect of sample size on the accuracy of LD estimation was examined by comparing the mean r^2 values calculated using randomly derived samples of decreasing sizes (120, 60, 30 and 15 accessions) to the ones estimated using the whole collection (242 accessions). As shown in Fig. 21b, mean r^2 values were biased upward as the sample size was reduced. This tendency was especially evident for SNP inter-distances higher than 1 Mb. Small differences of mean r^2 estimates were obtained with sample sizes greater than 30 and marker inter-distances lower

than 1 Mb. With sample size of 15 accessions, however, significant differences about LD estimates were noticeable even at SNP inter-distances of 10 to 100 bp.

In the same way, the effect of population structure on LD was examined by comparing LD patterns in the whole collection *vs* in each of the three subpopulations determined by STCTURE (Fig. 21c). As can be seen in that Figure, the mean r^2 values across categorical inter-marker distances estimated separately in each subpopulation (q1, q2 and q3) were almost identical. However, at large inter-marker distances (> 10 kb), mean r^2 values in the whole collection were lower than mean r^2 values in the subpopulations.

Figure 21. Effect of MAF (a), sample size (b) and population structure (c) on SNP-based LD estimates.

Mean r^2 values were pooled over the first three chromosomes of grapevine (chr1, 2 and 3). “q1”, “q2”, “q3” refers to the three populations identified by STRUCTURE and “Total” refers to the whole collection.

4.4.2. LD analysis using SSR markers

The values of r^2 between SSRs from different chromosomes (background LD) ranged between 3.480×10^{-6} and 0.221 with an average of 0.011 ± 0.008 (Fig. 22a). Only 3.38×10^{-3} per cent of the unlinked SSR pairs presented r^2 values ≥ 0.2 .

The distribution of r^2 values between linked SSRs (linked LD) within each chromosome is presented in Fig. 22b and Fig. A.3. Only 0.49% of the linked SSR pairs across the genome presented r^2 values higher than 0.2. Overall, there was a clear relationship between LD decay and the increment of physical distance between the SSR markers (Fig 22c). Estimated distances of LD decay below the critical value of $r^2 = 0.2$ for individual chromosomes and across the whole genome are shown in Table 6. In the chromosomes 16, 17, and 19 the obtained r^2 values did not allow to estimate the distances of LD decay using the logarithmic function because the logarithmic tendency line didn't cross the baseline of the considered r^2 threshold. SSR-based LD across the whole genome declined below $r^2 = 0.2$ within an average distance of 0.24 kb, but with significant variation between the chromosomes. Chr2 showed by far the most extensive LD in the genome, decaying below 0.2 within a distance of 0.18 Mb, whereas the next extensive distance was 0.07 Mb in chr18.

Figure 22. Distribution of background and linked LD estimated using a set of 250 SSRs over 207 genotypes.

a: Distribution of background LD. **b:** LD decay plot on chr1, the rest are presented in Fig. A.3, the red line depicts the logarithmic trend line **c:** Average r^2 values at different categorical physical distances across the genome.

Table 6. Estimated distances of SSR-based LD decay across the grapevine genome and estimated number of SSRs for whole-genome covering.

Chr.	Chr. size (Mb)	Number of SSRs	Average inter-SSR distance (Mb)	r^2 threshold = 0.2	
				Distance of LD decay (kb)	Number of SSRs for GWAS
chr1	24.23	18	1.35	0.116	208,879
chr2	18.89	14	1.35	186.547	101
chr3	20.70	11	1.88	0.167	123,952
chr4	24.71	14	1.77	3.028	8,161
chr5	25.65	20	1.28	0.003	8,550,000
chr6	22.65	13	1.74	0.179	126,536
chr7	27.36	15	1.82	0.003	9,120,000
chr8	22.55	12	1.88	0.101	223,267
chr9	23.01	12	1.92	0.779	29,538
chr10	23.50	9	2.61	0.018	1,305,556
chr11	20.12	9	2.24	0.002	10,060,000
chr12	24.27	10	2.43	0.038	638,684
chr13	29.08	16	1.82	0.004	7,270,000
chr14	30.27	14	2.16	0.001	30,270,000
chr15	20.30	13	1.56	1.400	14,500
chr16	23.57	11	2.14	-	-
chr17	18.69	8	2.34	-	-
chr18	34.57	17	2.03	7.416	4,662
chr19	24.70	14	1.76	-	-
All chr.	458.82 ^a	250 ^a	1.90	0.241	67,953,836 ^a

The number of SSRs required to cover each chromosome was estimated by dividing the chromosomal length (according to the 12x.2 version of the reference genome sequence) by the LD decay distance. For chr16, 17 and 19 it was not possible to estimate the distance of LD decay because the logarithmic tendency line didn't cross the baseline of the considered r^2 threshold. ^a Column sum. ^b Column average.

4.5. GENOME-WIDE ASSOCIATION ANALYSIS

4.5.1. GWAS using SNP markers

4.5.1.1. Comparative assessment of association analysis methods

The presence of a population substructure within the mapping collection combined to the existence of complex familial relationships (kinship values > 0.05) were expected to affect the results of association analysis. In order to assess how different statistical models deal with this problem, we tested four methods of analysis with increasing complexity: the naïve General Linear Model (GLM, without any correction), GLM corrected for population structure (Q), a Mixed Linear Model (MLM) accounting only for pairwise relatedness (K), and a MLM accounting for both population structure and pairwise relatedness (Q+K). These models were tested for association analysis of four “model traits”, two of simple genetic control: berry skin color and flower sex, and two with more complex genetic architecture: muscat flavor and seed content.

The four statistical models succeeded in detecting significant associations with berry skin color and flower sex controlling loci on chr2, as well as the major effect QTLs described for muscat flavor on chr5 and for seed content on Chr18, with the Bonferroni-corrected genome-wide significance threshold of $P < 1.4E-06$ (Fig. 23 and 24 and Fig. A.4). However, the QQ-plots showed significant differences among the models about the deviation of the observed P -values regarding the expectation under the null hypothesis, with variable magnitudes among the “model traits” (Fig. 25). Generally, the degree of deviation from the expectation was inversely proportional to the degree of complexity of the tested models. The most evident difference was observed between the GLM models (naïve GLM and Q model) and the MLM models (K and Q+K models) across all the traits and years of study; and within each group there was a more or less difference between the models according to the considered trait. The Q model showed less deviation than the naïve GLM model when used to analyze muscat flavor, MNS-2010, MSDW-2010 and MNS-2012, whereas no difference was observed between them in the rest of the traits and years. Regarding the MLM models, the Q+K model performed only slightly better than the K model in almost all the traits and years.

On the other hand, the decrease of the number of spurious associations (less deviation from the expectation) was usually accompanied by a reduction in the number of SNPs significantly associated with the major genes/QTLs of the considered traits, particularly for muscat flavor (from 73 SNPs under GLM to 10 SNPs under Q+K) and seed content related traits (Table 7). And more extremely, the association signals in chr18 near the seedlessness locus were completely lost when the K or Q+K models were used to analyze the association with MSDW-2012.

Berry color

Muscat flavor

Mean number of seeds-2013

Figure 23. Genome-wide association profiles (Manhattan plots) for five traits obtained using four analysis models.

Trait names are given at the top of each panel. The panels corresponding to other years for seed content traits are presented in Fig. 1 in A.5. The identifier of the analysis model is shown at the upper-right side of each plot. The 19 grapevine chromosomes are identified in the legend at the bottom of each panel. The SNPs assigned to the last column of each plot (identified as “chr20”) correspond to the 433 SNPs that were located in unmapped contigs. The genome-wide significance threshold is $P < 1.4 \times 10^{-6}$ ($-\log_{10}(P) = 5.856$).

Figure 24. Chromosome-wide association patterns showing the colocalization of association peaks with major loci for berry color, flower sex, muscat flavor and seed content.

The shown associations are those identified using the Q+K model. Red triangles indicate the physical positions of the previously found QTLs and/or genes for each trait (their position relative to the vertical axis is arbitrary). The genome-wide significance threshold is $P < 1.4 \times 10^{-6}$ ($-\log_{10}(P) = 5.856$). **a)** Association pattern of berry skin color with the position of *Myb* genes on chr2; **b)** Association pattern of flower sex with the position of the closest SSR markers to the sex controlling locus on chr2; **c)** Association pattern of muscat flavor and the position of the candidate gene *VvDXS* on chr5; **d)** Association pattern of mean number of seeds with the position of the two QTLs found on chr18 near *VVIN83* SSR and near *VvAGL11* candidate gene or *VMC7F2* SSR.

Figure 25. Quantile-Quantil (QQ) plots for four analysis models across four traits.

MNS: mean number of seeds, MSDW: mean seed dry weight, analysis models: ●: GLM, ●: Q, ●: K, ●: Q+K_s

Table 7. Variation of the number of significantly associated SNPs according to the analysis model and to their position regarding the major loci governing the trait.

Trait	Localization	Statistical model			
		GLM	Q	K	Q+K
Berry color	Genome-wide	78	80	63	61
	Chr2	73	76	58	57
Flower sex	Genome-wide	131	112	51	48
	Chr 2	11	10	10	9
Muscat flavor	Genome-wide	673	221	41	26
	Chr 5	74	39	14	10
MNS-2010	Genome-wide	153	78	18	10
	Chr 18	12	8	6	6
MNS-2012	Genome-wide	315	237	29	20
	Chr 18	21	13	5	4
MNS-2013	Genome-wide	317	244	39	36
	Chr 18	35	29	10	10
MSDW-2010	Genome-wide	124	95	21	13
	Chr 18	8	8	5	3
MSDW-2012	Genome-wide	121	113	35	19
	Chr 18	8	11	0	0
MSDW-2013	Genome-wide	112	128	4	6
	Chr 18	18	15	2	3

We assumed that all the SNPs with significant association in the chromosomes C2, C5 and C18 are linked to the major loci controlling color and flower sex, muscat flavor and seedlessness respectively. Seed content was analyzed as mean number of seeds per berry (MNS) and as mean seed dry weight (MSDW) and the results are shown for three seasons (2010, 2012 and 2013). The genome-wide significance threshold is $P < 1.4 \times 10^{-6}$ ($-\log_{10}(P) = 5.856$).

4.5.1.2. GWAS based on the mixed model Q+K.

Based on the comparative assessment presented in the prior section, the rest of the traits were analyzed using the Q+K model (Fig. 26 and Fig A.5). However, since true positives with weak association signals can be lost under the Q+K model as shown above, the traits were also analyzed using the naïve GLM model for comparison (Fig A.6). Table 8 summarizes the main features of the stable (detected using the Q+K model for at least two seasons) and putative (detected only in one season using the Q+K model, but supported by additional evidences such as: clear peaks of association using the GLM model in other years, colocalization with previously detected QTLs or colocalization with associations of phenotypically correlated traits) associations detected. GWAS for the binary traits, berry skin color, flower sex and muscat flavor was performed using merged consensus data among the three years of study, and the consistency of the detected associations for these traits was judged based on their accordance with previously reported QTLs in several studies focusing on them. Most of the detected associations in this

study were concentrated on chr1 (seed content), chr2 (berry color, flower sex and muscat flavor), chr5 (muscat flavor, cluster weight, berry size and flesh firmness), chr8 (cluster compactness, berry size and shape), and chr17 and 18 (seed content and berry size). The detailed features of these associations including the physical position and P values of the significantly associated SNPs to a given trait in each year of study are presented in chromosome-wise tables (Tables A.8 to A.13) in the Annex.

Berry skin color: a total of 57 SNPs on chr2 were found significantly associated with this trait (Table A.9). Fifty five of them were located under the main association peak that colocalized with the known color locus (Fig. 24 and Table 8) and spanned a distance of 5.33 Mb on the second half of chr2 (from the 13,264,449 bp to 18,592,574 bp positions), with a maximum P -value of $1.93\text{E-}28$ at 15,014,566 bp. The two remaining SNPs were located separately on the first half of that chromosome (7,597,669 bp and 9,022,191 bp) with P values of $1.52\text{E-}07$ and $9.70\text{E-}07$ respectively. Additionally, single SNPs on chr3, chr12 and chr13 were significantly associated with this trait with P values of $1.03\text{E-}06$, $9.96\text{E-}21$ and $9.36\text{E-}08$ respectively (Fig. 23).

Flower sex: the highest association peak (maximum P -value = $1.91\text{E-}16$ at 4,848,223 bp; Table A.9) for this trait was found on the first half of chr2 colocalizing with the known flower sex locus (Fig. 24 and Table 8). This peak comprised nine SNPs and spanned a distance of 0.74 Mb (from 4,109,392 bp to 4,848,223 bp). Moreover, one to eight SNPs per chromosome from all the chromosomes but chr3, 10, 11 and 14 were significantly associated with this trait with P values ranging from $8.74\text{E-}07$ to $2.65\text{E-}12$ (Fig. 23).

Muscat flavor: Two main association peaks were identified for muscat flavor (Fig. 23). The highest one (maximum P -value = $7.07\text{E-}10$ at 3,068,001 bp; Table A.10) on chr5 near the previously detected major effect QTL (Fig. 24 and Table 8) included 10 SNPs starting at 735,377 bp and ending at 4,431,496 bp (peaks range = 3,70 Mb). The second was identified on chr2 near a minor effect QTL previously reported for this trait (Table 8). This one contained eight SNPs, ranging from 2,711,805 bp to 3,629,148 bp (0.92 Mb; Table A.9), with the culminant point (maximum P -value = $4.43\text{E-}09$) at 2,974,643 bp. In addition, two SNPs on chr7 and single SNPs on chromosomes 4, 8, 9, 15, 16 and 18 with P values ranging from $1.11\text{E-}06$ to $1.03\text{E-}08$ were also significantly associated with the trait (Fig. 23).

Seed content: Stable associations over years were detected for both variables (mean number of seeds (MNS) and mean seed dry weight (MSDW)) on chr1, chr10 and chr18 colocalizing with

small (chr1, chr10 and chr18) or large (chr18) effect QTLs reported in previous studies (Table 8). Moreover, a new stable association was detected for the first time on chr17 (Fig. 23).

- On chr1 (Table A.8), five SNPs ranging from 12,796,528 bp to 19,854,230 bp and five SNPs ranging from 20,997,589 bp to 23,238,804 bp were repeatedly associated for at least two years with MNS and MSDW respectively, with P values ranging between $1.22\text{E-}06$ and $4.06\text{E-}09$.
- On chr10, a single SNP at 19,547,987 bp was repeatedly associated with MNS for three years and with MSDW for two years, with P values ranging between $1.35\text{E-}06$ and $2.93\text{E-}09$.
- The new association signal detected on chr17 consisted of 15 SNPs ranging from 6,256,364 bp to 10,741,556 bp that were significantly associated to MNS and/or MSDW for at least one year (Table A.12). Among them, four SNPs ranging between 9,201,088 bp and 9,739,079 bp were repeatedly associated with MNS for two years (2012 and 2013) and two of them were also associated with MSDW in 2012; and behind those (between 6,256,381 bp and 6,256,518 bp) four SNPs were significantly associated with MNS in 2013 and with MSDW in 2012. P values of all the associations ranged between $1.14\text{E-}08$ and $1.29\text{E-}06$.
- On chr18 (Table A.13), two clearly individualized association peaks were identified. The largest one coincided with the major seedlessness QTL and encompassed four SNPs ranging between 32,027,523 bp and 32,712,280 bp that were associated with MNS for the three years of study, and one of them (32,125,496 bp) was associated with MSDW for two years. The P values of these associations ranged between $1.47\text{E-}07$ and $1.00\text{E-}11$. The second peak overlapped with the previously reported minor QTL for seedlessness in this linkage group (Table 8), and consisted of two SNPs at the positions 10,065,456 bp and 10,177,341 bp with P values varying between $1.28\text{E-}06$ and $1.45\text{E-}09$ according to the trait and year.

Berry size related traits: Berry size is related to its weight and to its dimensions (length and diameter). A new association stable over at least two years was identified for MBW on chr5 (Table A.10). Within this association signal, two SNPs located at 8,734,944 bp and 14,046,799 bp were significantly associated with MBW in 2012 and 2013 for the former, and in 2010 and 2013 for the latter, with P values ranging between $4.73\text{E-}07$ and $2.26\text{E-}08$. Additionally, two instable associations were detected for this trait on chr8 (Table A.11) and chr17 (Table A.12). The association signal on chr17 comprised 8 SNPs ranging from 5,148,640 bp to 6,366,721 bp

with P values ranging between $5.5964\text{E-}07$ and $7.2833\text{E-}08$, and overlapped with a major effect QTL identified recently in this region (Table 8). On chr8, MBW was associated with a single SNP at 20,789,025 bp with a P value of $2.97\text{E-}07$, coinciding with a previously identified QTL on this linkage group (Table 8). Unstable associations for MBL were identified in not regions not previously reported: in 2010 on chr5 (two SNPs: 16,010,197 bp and 16,010,204 with P values of $1.19\text{E-}06$ and $1.18\text{E-}06$ respectively; Table A.10), chr8 (two SNPs: 3,405,149 bp and 21,184,639 bp with P values of $7.05\text{E-}07$ and $1.30\text{E-}06$ respectively; Table A.11), and chr10 (a single SNP at 21,960,325 with a P value of $2.78\text{E-}07$; Fig. 26). A new stable association was identified for MBD on the first half of chr18 (Table A.13), far from the known major QTL for berry size that has been found in several studies overlapping with the major seedlessness QTL at the bottom part of chr18 (around 32 Mb; Table 8). This new association involved four SNPs ranging between 4,914,521 bp and 6,697,371 bp that were associated to the trait for at least two years with P values ranging between $2.07\text{E-}07$ and $1.19\text{E-}06$. Additionally, two instable associations were identified on chr17 (Table A.12) and chr11 (Fig. 26). The association on chr17 coincided with the MBW association and involved six SNPs ranging between 6,256,364 bp and 6,366,721 bp with P values ranging between $2.99\text{E-}07$ and $1.38\text{E-}08$. The association on chr11 (one SNP at 4,447,258 bp, P value of $1.15\text{E-}06$) was found near a QTL previously reported on the same chromosome (Table 8).

Berry shape: Putative associations were identified for this trait on chr8 (a single SNP at 9,738,860 bp, P value of $1.16\text{E-}06$; Table A.11) and chr11 (a single SNP at 10,371,958 bp, P value of $3.62\text{E-}07$; Fig. 26).

Flesh firmness: A putative association for this trait was detected on chr5 (Table A.10) overlapping with the MBW association indicated above and with a flesh firmness QTL reported in a previous study (Table 8). This association consisted of two SNPs located at 14,046,799 bp and 14,327,183 bp with P values of $9.11\text{E-}07$ and $4.95\text{E-}07$ respectively.

Cluster related traits: Putative associations were detected for two cluster related traits: cluster weight association detected in 2013 on chr5 (a single SNP at 20,051,857 bp with a P value of $1.29\text{E-}06$; Table A.10), overlapping with a previously reported QTL, and a new association with cluster compactness (CI-12) in 2012 on chr8 (a single SNP at 3,452,355 bp with a P value of $5.88\text{E-}07$; Table A.11).

Ripening time: only unstable association were detected for this trait: in 2012 on chr12 (Fig. 26), involving four SNPs ranging from 18,627,991 bp to 18,629,307 bp, with P values ranging

between 4.00E-07 and 3.11E-07. This association coincided with a previously reported QTL for this trait on linkage group 12 (Table 8).

Fertility: Even if any P value reached the significance threshold ($1.4\text{E-}6$) under the Q+K model, there are three arguments that suggest the presence of a real association for this trait on chr5: i) Outstanding association peaks on chr5 are clearly recognizable on the Manhattan plots corresponding to this trait over the three years of study (Fig. 26), with the maximum P values of $1.39\text{E-}04$ at 15,323,767 bp in 2010, $1.28\text{E-}04$ at 3,943,352 bp in 2011 and $2.76\text{E-}05$ at 25,459,851 bp in 2013. ii) Under the GLM model, highly significant associations were identified in the same region across the three years of study (Fig. A.6). iii) In a previous study, Doligez *et al.* (2010), reported a stable and large effect QTL on the same linkage group.

No significant associations were detected for three phenological traits (Bb, Fl and Vr), cluster length and diameter and peduncle length and cluster compactness CI-OIV (Fig. A.5), and there are not additional arguments to suggest putative associations for these traits.

Figure 26. Genome-wide association profiles obtained using the Q+K model for nine grape traits across three growing seasons

Only the traits with significant associations under the Q+K model are presented here, see Fig. A.5 for the rest. Each trait is represented in a single line of three plots that correspond respectively from the left to the right to the years 2010, 2012 and 2013. The 19 grapevine chromosomes are identified in the legends. The SNPs assigned to the last column of each plot (identified as “chr20”) correspond to the 433 SNPs that were located in unmapped contigs. The genome-wide significance threshold is $P < 1.40\text{E-}06$ ($-\log_{10}(P) = 5.856$).

Table 8. Main features of genetic associations identified for 13 grape traits across three growing seasons using a mapping collection of 242 accessions and a set of 36,332 SNPs.

Trait	Chr/ Region	Max <i>P</i> value position (bp)	Max <i>P</i> value	Year of detection	Studies reporting QTLs in the same region
Rp	12	18,629,307	3.11E-07	12	(Costantini <i>et al.</i> , 2008)
FSx	2	4,761,375	1.91E-16	Consensus [§]	(Dalbó <i>et al.</i> , 2000; Lowe and Walker, 2006; Riaz <i>et al.</i> , 2006; Marguerit <i>et al.</i> , 2009; Fechter <i>et al.</i> , 2012; Battilana <i>et al.</i> , 2013)
MCW	5	20,051,857	1.29E-06	13	(Fanizza <i>et al.</i> , 2005)
CI-12	8	3,452,355	5.88E-07	12	(Correa <i>et al.</i> , 2014)
MBW	5	8,734,944	2.26E-08	10, 12, 13	(Fischer <i>et al.</i> , 2004) [#]
	8	20,789,025	2.97E-07	12	(Mejía <i>et al.</i> , 2007; Doligez <i>et al.</i> , 2013)
	17	5,148,640	5.60E-07	13	(Doligez <i>et al.</i> , 2013)
MBL	5	16,010,204	1.18E-06	10	non
	8	3,405,149	7.05E-07	10	non
MBD	11	4,447,258	1.15E-06	12	(Doligez <i>et al.</i> , 2013) but MBW
	17	6,256,518	4.60E-08	13	(Doligez <i>et al.</i> , 2013) but MBW
	18	4,914,521	2.74E-07	10, 12, 13	non
ShI	8	10,371,958	3.62E-07	13	non
	11	9,738,860	1.16E-06	10	non
Col	2	15,014,566	1.93E-28	Consensus [§]	(Doligez <i>et al.</i> , 2002; Fischer <i>et al.</i> , 2004; Lijavetzky <i>et al.</i> , 2006; Fournier-Level <i>et al.</i> , 2009; Ban <i>et al.</i> , 2014; Costantini <i>et al.</i> , 2015)
FI	5	14,327,183	4.95E-07	12	(Carreño <i>et al.</i> , 2014)
Mst	2	2,974,643	4.43E-09	Consensus [§]	(Doligez <i>et al.</i> , 2006; Battilana <i>et al.</i> , 2009; Duchêne <i>et al.</i> , 2009)
	5	3,068,001	7.07E-10	Consensus [§]	(Doligez <i>et al.</i> , 2006; Battilana <i>et al.</i> , 2009; Duchêne <i>et al.</i> , 2009)
MNS	1	20,997,589	2.65E-10	10, 12, 13	(Cabezas <i>et al.</i> , 2006; Mejía <i>et al.</i> , 2007; Doligez <i>et al.</i> , 2013)
	10	19,547,987	2.93E-09	10, 12, 13	(Cabezas <i>et al.</i> , 2006; Costantini <i>et al.</i> , 2008)
	17	9,294,776	1.14E-08	12, 13	non
	18 (10)*	10,065,456	1.45E-09	10, 13	(Doligez <i>et al.</i> , 2002; Cabezas <i>et al.</i> , 2006; Mejía <i>et al.</i> , 2007; Costantini <i>et al.</i> , 2008; Doligez <i>et al.</i> , 2013)
	18 (32)*	32,027,523	1.80E-11	10, 12, 13	(Doligez <i>et al.</i> , 2002; Cabezas <i>et al.</i> , 2006; Mejía <i>et al.</i> , 2007; Costantini <i>et al.</i> , 2008; Doligez <i>et al.</i> , 2013)
MSDW	1	19,854,230	2.75E-07	10, 12, 13	(Cabezas <i>et al.</i> , 2006; Mejía <i>et al.</i> , 2007; Doligez <i>et al.</i> , 2013)
	10	19,547,987	3.28E-07	12, 13	(Doligez <i>et al.</i> , 2002; Cabezas <i>et al.</i> , 2006; Doligez <i>et al.</i> , 2013)
	17	6,256,364	4.99E-08	10, 12	non
	18 (10)*	10,065,456	1.90E-07	10, 13	(Doligez <i>et al.</i> , 2002; Cabezas <i>et al.</i> , 2006; Mejía <i>et al.</i> , 2007; Costantini <i>et al.</i> , 2008; Doligez <i>et al.</i> , 2013)
	18 (32)*	32,125,496	3.64E-08	10, 13	(Doligez <i>et al.</i> , 2002; Cabezas <i>et al.</i> , 2006; Mejía <i>et al.</i> , 2007; Costantini <i>et al.</i> , 2008; Doligez <i>et al.</i> , 2013)

See Table 1 for the description of the traits. The traits for which no significant associations were detected using the Q+K model are not listed in this table. *: two distinct association peaks with MNS and MSDW around 10 and 32 Mb on 18. [§]: GWAS was performed using the consensus phenotypic data among the three years of study. [#]: the QTL was found in a complex hybrid cross including non-*vinifera* background (up to seven American species). The genome-wide significance threshold is $P < 1.40E-06$ ($-\log_{10}(P) = 5.856$).

4.5.2. Association analysis using SSR markers

Like SNP-based GWAS, analyses of SSRs-traits associations were carried out using Q+K model (Table 9 and Table A.14), but GLM analysis was also developed for comparison purposes (results not reported). The genome-wide Bonferroni correction for the significance threshold ($P < 1.90E-04$) was found highly conservative when applied to the association results obtained using the Q+K model. At this threshold, among all the previously known major QTLs, only berry skin color and flower sex loci on chr2 were detected, although new associations (not reported before) were also found for time of flowering, time of veraison, mean cluster length and muscat flavor (Table 9). However,

P values for the SSR markers that were previously found to be linked to major effect QTLs underlying for example muscat flavor on chr5 (VVC6 with $P = 3.80E-4$) or seedlessness on chr18 (p3_VvAGL11 with $P = 3.50E-4$) remained below the Bonferroni corrected significance threshold. Therefore, in addition to the Bonferroni correction, a moderate nominal significance threshold of 0.01 was also considered. The consistency of the detected associations was judged according to the repeatability over years and colocalization with SNP-based associations identified in the present study and/or with previously reported QTLs or with genetic loci underlying phenotypically correlated traits. The significant SSR associations at the nominal level of 0.01 that were detected for more than one year and/or agreed with genetic associations found in the present study using the SNP markers and/or with QTLs reported in the literature are detailed in Table 9. The rest (non-replicated over years, neither coincided with previous discoveries) are listed in Table A.14 in the Annex.

Berry skin color: Eight SSRs spanning a distance of 10.38 Mb in chr2 (from VChr2c at 7,998,935 bp to VMC7g3 at 18,382,886 bp) were significantly associated to berry color, with P values ranging between $8.91E-03$ for VChr2c and $5.01E-07$ for VMC7g3. This association overlapped with the known grapevine color locus in this chromosome.

Flower sex: Two SSRs on chr2 (UDV-109 and VMC3B) were significantly associated with this trait, coinciding with the known flower sex locus. Besides this, flower sex was significantly associated with other SSR markers distributed almost over the whole genome (all the chromosomes but 7, 11 and 12) with P values ranging between $9.97E-03$ and $1.68E-08$.

Muscat flavor: Three associations overlapping with previously reported QTLs for this trait were detected on chr5, chr2 and chr13. The association on chr5 colocalized with the largest effect

QTL for this trait and involved three SSRs (FAM91, VVC6 and VRZAG79) spanning a distance of 4.91 Mb, with P values ranging between $1.53\text{E-}03$ and $3.80\text{E-}04$. The associations on chr2 and chr13 coincided with minor effect QTLs and involved single markers (VVS3-2 on chr2 and VMC8E6 on chr13). Moreover, one SSR on chr10, four SSRs on chr14 and two SSRs on chr19 (one of them with a P value above the Bonferroni corrected threshold) were significantly associated with muscat flavor.

Seed content: Stable associations over at least two years were detected for the mean number of seeds per berry (MNS) in chr1, chr17 and chr18 with P values ranging between $8.09\text{E-}03$ and $3.72\text{E-}04$. The association on chr18 involved four SSRs: VChr18a-2 (8,512,550 bp) and UDV-134-2 (18,100,286 bp) closer to the minor effect QTL, and VVIN16 (26,361,305 bp) and p3_VvAGL11 (30,313,541 bp) closer to the major effect QTL for seedlessness. The associations on chr1 and chr17 involved single SSRs (VRZAG29 and VChr17b-2 respectively) and coincided respectively with a minor effect seedlessness QTL, and a stable association with seed content detected for the first time in the present study with the SNP dataset. The same markers (VRZAG29 and VChr17b-2) on chr1 and chr17 were also significantly associated with the mean seed dry weight (MSDW) but only in one year. Instable associations with MSDW were also detected on chr18 (with VChr18b) and chr4 (with Vchr4a and VVIN75-2) where a seed content-related QTL was previously identified.

Berry size related traits: Stable associations with mean berry weight (MBW) and unstable associations with mean berry length (MBL) and mean berry diameter (MBD) were detected next to each other on chr5 (MBW with VChr5b and VMC16D4-2, MBL with UDV-077 and VRZAG89 and MBD with VVIN40), overlapping with the new stable association identified using the SNP data set in the present study. Stable associations with berry size related traits were also detected on chr18 (MBW and MBL with VMC2B1.1, and MBD with VVIQ44 and VChr18a-2), in the same location as the stable association detected using the SNP dataset. And, a new association was detected for the first time on chr7, where MBL and MBD were repeatedly associated over at least two years with the same SSR, FAM116. Besides this, unstable single trait associations were also identified with MBW on chr4 (with Vchr4a) and chr11 (with RTAChr11-11b) overlapping with previously reported QTLs for berry size; and new stable associations with MBD were detected on chr2 (with FAM140-2) and chr8 (with UDV-075).

Berry shape: A stable association with berry shape index was detected on chr8 (with VMC7h2 and VMC5H2), in the same location as the newly discovered association for this trait with SNP markers in this study.

Cluster related traits: A stable association with cluster weight (MCW) was detected on chr18 (with VMC2B1.1) in the same region as the berry size related association indicated above. In addition, two unstable associations with this trait were also detected: one on chr5 (with VMC9B5), near the berry size association identified in the present study and a previously reported QTL in the same region; the other was identified on chr16 (with VChr16b), overlapping with another QTL identified in a previous work. An unstable association was also detected with cluster length on chr2 (with VMC2C10-1 and VV6718) coinciding with a previously reported QTL. Besides this, a stable association with cluster compactness (CI-12) was identified for the first time on chr9 with the SSR marker VVIQ66.

Phenological traits: Three simultaneous associations to more than a single phenology related trait were identified: i) an association with times of budbreak (Bb) and flowering (Fl) on chr5 (Bb with FAM122; and Fl with VVIT68 and VChr5b); ii) an association with Fl and time of veraison (Vr) on chr1 (Fl with VRZAG29, VVIO02 and FAM141; and Vr with VVIO02) and iii) an association with Fl and time of ripening on chr17 (both with VVIN73) that overlapped at the same time with the berry size and seed content association. Additionally, separate associations with individual traits were also identified including: a stable association with Fl on chr7 (with VVIQ67, VVIV04 and VVIN56) that overlapped with a major QTL reported in a previous study; new stable associations with Fl on chr13 (with FAM62-2), chr16 (with VVIN52) and chr19 (with UDV-062-2), new stable associations with Vr on chr4 (with VVIN75-2) and chr7 (with VVIS58-2); and an unstable association with Vr on chr2 (with VMC2C10-1 and RTAChr2-11a) but overlapping with a previously reported QTL for this trait.

Fertility: An unstable association with fertility was detected on chr5 (with VMC2C10-1) but overlapped with the putative association with this trait identified in the present study with SNP markers and with a major effect QTL for fertility reported in a previous study.

No consistent associations were identified for flesh firmness, cluster diameter and peduncle length.

Table 9. Genome-wide associated SSRs to 18 grape traits for three growing seasons in comparison with SNP association results and previous QTL studies (see the key below the table).

Trait	SSRs						SNPs		Previous QTLs
	Chr	marker	position	P value			position	P value	
				2010	2012	2013			
Bb	5	FAM122	15,723,310	3.60E-04	6.23E-03		15,937,061	1.37E-03	
Fl	1	VRZAG29	5,286,106	1.59E-03		2.91E-04	5,391,428	6.81E-05	
	1	VVIO02	9,813,247	9.44E-03			9,581,202	1.66E-03	1
	1	FAM141	17,732,854	2.83E-03	5.67E-03		17,934,996	2.22E-02	
	5	VVIT68	7,099,947	5.26E-05	5.40E-04	9.88E-03	6,920,108	3.60E-03	
	5	VChr5b	9,474,726	5.72E-03	5.63E-03	7.24E-04	9,109,604	3.71E-02	
	7	VVIQ67	22,899,285	2.75E-03			22,945,681	2.01E-03	2
	7	VVIV04	23,860,671			3.75E-03	23,884,675	2.83E-03	
	7	VVIN56	25,737,398	7.12E-03		4.93E-03	26,146,746	1.40E-03	
	13	FAM62-2	16,059,252	1.77E-03		2.80E-03	16,718,453	8.79E-03	
	16	VVIN52	619,678		9.07E-03	2.45E-03	103,040	1.41E-03	
	17	VVIN73	5,636,037		4.14E-03	2.55E-03	5,584,385	1.43E-03	
	19	UDV-062-2	21,594,816	5.61E-03		4.03E-03	21,176,594	1.14E-02	
Vr	1	VVIO02	9,813,247		2.93E-04	4.18E-05	9,073,595	2.07E-03	
	2	VMC2C10-1	6,975,274		7.74E-03		7,115,153	8.64E-04	3
	2	RTAChr2-11a	10,707,832		6.98E-03		10,797,251	2.67E-03	
	4	VVIN75-2	14,969,759		5.49E-03	5.34E-04	15,799,610	4.49E-03	
	7	VVIS58-2	5,335,417		2.65E-03	1.89E-03	5,052,568	3.75E-04	
Rp	17	VVIN73	5,636,037			7.48E-03	6,101,046	3.16E-06	4
FSx	2	UDV-109	132,291			9.32E-04	115,832	5.10E-03	
	2	VMC3B10	5,000,200			3.54E-16	4,761,375	1.91E-16	5
Frt	2	VMC2C10-1	6,975,274		6.65E-03				6
	5	VChr5c	4,080,604			9.41E-03	3,712,008	4.34E-05	7
	17	VRZAG15	6,588,595		4.32E-03	6.84E-03	6,309,363	2.55E-03	
MCW	5	VMC9B5	20,174,295			4.86E-03	20,051,857	1.29E-06	8
	16	VChr16b	17,432,683			6.80E-03	16,334,670	4.31E-04	9
	18	VMC2B1.1	3,752,073		9.97E-03	8.46E-03	3,644,906	9.17E-05	
MCL	2	VMC2C10-1	6,975,274			9.55E-03	6,881,814	3.53E-04	10
	2	VV6718	16,446,991		1.30E-04		17,019,686	4.18E-05	
CI-12	9	VVIQ66	10,808,981	4.20E-03		4.71E-04	10,077,142	5.43E-03	
MBW	4	Vchr4a	946,454	2.38E-03			1,217,271	2.82E-05	11
	5	VChr5b	9,474,726		9.29E-03		8,734,944	6.47E-08	
	5	VMC16D4-2	17,606,676		2.88E-03	7.60E-03	16,010,204	4.46E-05	12*
	11	RTAChr11-11b	11,365,896	7.86E-03			10,166,210	5.42E-04	13
	18	VMC2B1.1	3,752,073		6.34E-03		4,450,073	8.78E-05	
MBL	5	UDV-077	21,538,151			6.07E-03	21,900,673	1.93E-04	
	5	VRZAG89	22,919,299	8.20E-03			23,928,871	3.05E-05	
	7	FAM116	20,388,473	1.57E-03		4.82E-03	20,565,562	2.47E-04	
	18	VMC2B1.1	3,752,073		2.38E-03	5.63E-03	3,285,445	3.58E-04	
MBD	2	FAM140-2	9,166,787	8.23E-03	4.07E-03	3.91E-03	8,874,389	9.99E-03	
	5	VVIN40	24,429,944			7.98E-03	24,520,703	3.19E-05	
	7	FAM116	20,388,473	2.01E-03	7.37E-03	1.30E-03	20,024,096	1.28E-04	
	8	UDV-075	2,400,696	7.73E-03	1.44E-03	5.45E-04	2,682,261	8.60E-04	
	18	VVIQ44	5,309,353	6.70E-04	3.56E-04		4,914,521	2.07E-07	

Trait	SSRs						SNPs		Previous QTLs
	Chr	marker	position	P value			position	P value	
				2010	2012	2013			
ShI	18	VChr18a-2	8,512,550			3.30E-03	8,101,417	7.71E-05	
	8	VMC7h2	10,149,243		3.71E-03	4.22E-03	10,371,958	3.62E-07	
	8	VMC5H2	15,407,796		5.83E-03		14,887,678	1.18E-03	
Col	2	VChr2c	7,998,935			8.91E-03	7,597,669	1.52E-07	
	2	FAM140-2	9,166,787			1.09E-03	9,022,191	9.70E-07	
	2	RTAChr2-11a	10,707,832			9.51E-05	10,797,212	7.36E-05	
	2	FAM24	13,508,269			3.78E-04	14,428,045	6.88E-23	14
	2	VV6718	16,446,991			1.12E-11	16,994,368	1.90E-15	
	2	VVIU20	16,650,964			6.55E-05	16,994,368	1.90E-15	
	2	VMC8C2	17,626,961			3.51E-04	17,723,411	3.65E-15	
	2	VMC7g3	18,382,886			5.01E-07	17,723,411	3.65E-15	
Mst	2	VVS3-2	4,014,388			5.44E-03	3,508,212	1.24E-07	15
	5	FAM91	782,861			1.53E-03	735,377	7.32E-07	
	5	VVC6	1,470,620			3.80E-04	2,202,620	4.69E-09	16
	5	VRZAG79	5,692,460			1.18E-03	5,657,967	9.65E-06	
	13	VMC8E6	27,180,657			1.15E-03	27,035,789	6.83E-03	17
MNS	1	VRZAG29	5,286,106	4.95E-04		1.60E-03	6,289,513	4.61E-07	18
	17	VChr17b-2	5,017,280	1.14E-03	3.72E-04		6,256,364	3.91E-07	
	18	VChr18a-2	8,512,550			1.88E-03	8,777,929	4.35E-03	19
	18	UDV-134-2	18,100,286		4.41E-03		18,139,390	6.00E-05	
	18	VVIN16	26,361,305	2.45E-03	8.09E-03	3.11E-03	27,284,156	9.82E-04	
	18	p3_VvAGL11	30,313,541	3.79E-03		3.87E-03	30,234,060	2.13E-07	20
MSDW	1	VRZAG29	5,286,106	8.54E-04			5,789,434	9.64E-07	21
	4	Vchr4a	946,454	6.83E-03			956,184	2.56E-03	
	4	VVIN75-2	14,969,759			3.35E-03	14,907,801	4.54E-04	22
	17	VChr17b-2	5,017,280	3.50E-03			6,256,364	4.99E-08	
	18	VChr18b	14,028,762	2.94E-03			13,865,684	4.48E-04	

See Table 1 for trait names. We presented here only the significant SSR associations detected using the Q+K model at the significance threshold of 0.01 for more than one year and/or agreeing with genetic associations found in the present research using the SNP markers and/or with QTLs reported in the literature (see Table A.14 for the rest of associations). The SNP associations represent to the maximum *P* value found in a radius of 1 Mb around the position of the corresponding SSR association across the three years of study. Bold *P* values highlight the significant associations at the Bonferroni corrected threshold ($1.90E-03$ for the SSRs and $1.4E-06$ for the SNPs). Italic *P* values highlight the significant associations at 0.01 for the SSRs and at the equivalent one for the SNPs $7.30E-05$ (both of them are 53 times more relaxed than the respective Bonferroni corrections). *: the QTL was found in a complex hybrid cross including non-*vinifera* background (up to seven American species). **List of references:** **1** (Costantini *et al.*, 2008; Marguerit *et al.*, 2009), **2** (Duchêne *et al.*, 2012; Grzeskowiak *et al.*, 2013), **3** (Costantini *et al.*, 2008; Grzeskowiak *et al.*, 2013), **4** (Mejía *et al.*, 2007), **5** (Dalbó *et al.*, 2000; Lowe and Walker, 2006; Riaz *et al.*, 2006; Marguerit *et al.*, 2009; Fechter *et al.*, 2012; Battilana *et al.*, 2013), **6** (Marguerit *et al.*, 2009), **7** (Doligez *et al.*, 2010), **8** (Fanizza *et al.*, 2005), **9** (Fanizza *et al.*, 2005), **10** (Marguerit *et al.*, 2009), **11** (Fanizza *et al.*, 2005), **12** (Fischer *et al.*, 2004), **13** (Doligez *et al.*, 2013), **14** (Doligez *et al.*, 2002; Fischer *et al.*, 2004; Lijavetzky *et al.*, 2006; Fournier-Level *et al.*, 2009; Ban *et al.*, 2014; Costantini *et al.*, 2015), **15** (Doligez *et al.*, 2006; Battilana *et al.*, 2009; Duchêne *et al.*, 2009), **16** (Doligez *et al.*, 2006; Battilana *et al.*, 2009; Duchêne *et al.*, 2009), **17** (Doligez *et al.*, 2006; Duchêne *et al.*, 2009), **18** (Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Doligez *et al.*, 2013), **19** (Doligez *et al.*, 2002; Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Doligez *et al.*, 2013), **20** (Doligez *et al.*, 2002; Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Doligez *et al.*, 2013), **21** (Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Doligez *et al.*, 2013), **22** (Doligez *et al.*, 2013).

DISCUSSION

5. DISCUSSION

5.1. PHENOTYPIC DATA

In this study we evaluated 21 phenotypic traits that were selected to represent different breeding objectives currently targeted for table grape genetic improvement. In this sense we selected traits related to grape phenology that is the basis for developing early and late ripening cultivars; traits related to yield that is predominantly determined by the number of clusters per vine (fertility), but also by cluster size; traits related to different features of the cluster that determine its final size and architecture; and traits related to different aspects of fruit quality such as berry size, color, firmness, flavor, and seed content. These traits were also selected and evaluated in such a way that they represent different categories of phenotypic traits: quantitative, such as phenology, fertility and cluster and berry sizes; semi-quantitative or ordinal, such as cluster compactness and flesh firmness; and qualitative such as flower sex or berry color. Another important aspect that distinguishes between these traits is the process of evaluation itself: there are quantitatively measurable traits such as counts (number of clusters or seeds), weights (of clusters or berries), and dimensions (length and diameter), and subjective traits assessed on the basis of the evaluator observation (phenology and cluster compactness) or taste (flavor and firmness). In this ultimate context, the subjective traits were assessed by at least two persons in order to reduce the effect of the evaluator on the phenotypic variation of the trait. Our objective behind choosing this diversified set of traits was to test the ability of GWAS in finding the genetic determinism in each particular combination of trait properties. On the other side, even if our interest was particularly focused on table grapes, we included in the mapping collection a large proportion of cultivars that are traditionally used for wine production. This was driven by the perspective of finding allelic variants with potential interest for table grape production within the wine grape compartment. Such allelic variants would have been remained underexploited because along the history of grape breeding, wine and table grapes have been inter-mated and selected separately from each other (Myles *et al.*, 2011).

The examination of traits distributions over the three years of study showed that fertility and phenological traits, particularly budbreak and flowering were, in average, considerably variable among years. This is not surprising since these traits are known to be strongly sensitive to the variation of environmental factors, such as the patterns of seasonal weather (light, temperature, rainfall, and humidity) and cultural practices (pruning, canopy management, irrigation, nutrition, and pest and disease management) (Keller, 2015). For instance, the low fertility observed in

2012 is most likely the consequence of the severe loss of leaf area caused by downy mildew, which coincided with the period of inflorescence initiation and differentiation (around flowering) in the previous growing season (Bennett *et al.*, 2005; Poni *et al.*, 2006; Lebon *et al.*, 2008). This incident, probably, triggered the phenomenon known in perennial fruit crops as alternate (or biennial) bearing. This phenomenon is the most likely factor that can explain why a relatively higher fruitfulness was observed in the subsequent year (2013).

On the other hand, with the exception of berry size traits (weight, length and diameter), veraison and ripening, the rest of the traits were only moderately to weakly consistent among years. The lack of consistency between years or environments is usually attributed to the effect of genotype-environment interaction, which causes that different genotypes have different relative performances in different environments (years in our case), changing hence the rank order for genotypes between environments (Falconer and Mackay, 1996). However, the use of subjective descriptors for the assessment of some traits may also be involved in a lower concordance between years. For example, the use of overall visual inspection to judge that a set of vines have reached 50% budbreak or flowering, can introduce some bias in the order of merit between cultivars having close budbreak or flowering times from one year to another.

The study of patterns of phenotypic correlations between traits is useful in the way that it helps to understand the functional and developmental relationships among traits and in the interpretation of genetic correlations that can be discovered in a given genetic mapping study. The highest correlations in the present study were found between berry size traits (weight, length and diameter). Correlation coefficient values were identical for all combinations of berry size traits, indicating that berry size is determined equally by tissue growth through longitudinal and transversal axis; whereas berry shape seem to be more determined by length than diameter increase. The dependence of fruit shape on its length was also reported for melon fruits (Monforte *et al.*, 2005; Díaz *et al.*, 2014). The highest correlation between phenological traits was found between veraison and ripening; the rest of phenological trait pairs showed only a weak correlation. This agrees with previous studies (Jones *et al.*, 2005; Tomasi *et al.*, 2011) and makes a good sense, because while veraison and ripening are related to the same plant organ (fruit), budbreak and flowering correspond to different plant organs (buds and flowers respectively) and could be hence governed by distinct functional processes. Fertility was negatively correlated with cluster as well as berry size traits, agreeing with the principal of compensation between yield components (i.e. as the number of clusters per vine increases, the number of berries per cluster and the average weight of these berries decrease) (Keller, 2015).

Strong relationship between seed content and berry size have been reported in several studies based on seedless F₁ populations (Doligez *et al.*, 2002; Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Doligez *et al.*, 2013). This was usually attributed to gibberellins biosynthesis by the seeds that regulate cell division and expansion in the pericarp. In our study however, seed content traits, particularly, number of seeds were not strongly correlated with berry size traits ($0.19 \leq Rho \leq 0.26$). This is clearly exemplified in many wine grapes that are characterized by high seed content but small berries such as Chardonnay and Cabernet Sauvignon. The lack of interdependence between berry size and seed content in these cultivars suggest that other mechanisms with strong effect should be involved in the control of tissue growth in grape berries.

The evaluation of such a number of traits in such a diverse grapevine population as done in this study has not been reported before in the literature. Vargas *et al.*, (2009) evaluated a similar set of traits in a collection of 322 accessions (245 non-redundant genotypes) but exclusively of table grapes. The range of variation (here expressed as standard deviation, SD) of the studied traits in our work coincided globally with those reported by Vargas (Table 10). However, some differences were observed in the overall mean of some traits such as peduncle length that was relatively higher in our mapping collection, and berry size that was relatively higher in Vargas's collection. The difference in peduncle length is due to the manner in which the trait was evaluated in each study: while it was measured from the insertion point until the first ramification of the principal cluster in our study, the first ramification in Vargas's study was considered as such regardless of whether it corresponded to the principal cluster or to the first shoulder (when the latter was present, the peduncle length resulted shorter). The difference in berry size however is not surprising since table grape berries are usually larger than wine grape berries that did not exist in Vargas's collection. With respect to phenotypic correlations, this author found that berry size related traits were the most consistent over years, whereas fertility, cluster related traits, flesh firmness, and seed content did not show high correlation between years. This agrees substantially with our findings. Regarding correlations between traits within years, Vargas *et al.* (2009) reported that the highest correlations (> 0.6) were found between berry size traits (weight, length and diameter) and between cluster weight and its diameter, whereas the correlation between cluster weight and its length was not so high; moreover, she reported that among the phenological stages, the strongest correlations were observed between veraison and ripening. Altogether, these results are in complete accordance with our findings.

Table 10. Descriptive statistics of quantitative traits as evaluated in our mapping collection vs. in a collection of 322 table grape accessions (GLOBAL322) by Vargas *et al.* (2009).

	MC-2010		MC-2012		MC-2013		MC-mean		GLOBAL322	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fert	1.25	0.51	1.08	0.54	1.77	0.44	1.37	0.50	1.40	0.52
MCW	396.85	237.96	406.07	226.56	396.08	213.56	399.67	226.03	338.36	195.62
MCL	18.37	5.31	19.12	4.60	19.96	4.62	19.15	4.85	19.49	4.45
MCD	11.59	2.62	13.15	2.71	11.06	2.52	11.93	2.62	12.13	2.71
MPL	5.51	2.22	4.08	1.49	4.67	2.14	4.75	1.95	2.03	1.15
MBW	2.82	1.43	2.39	1.37	2.20	1.17	2.47	1.32	3.02	1.32
MBL	17.23	4.31	15.96	3.97	14.59	3.20	15.93	3.83	17.81	3.84
MBD	15.25	2.70	14.81	2.57	13.95	2.35	14.67	2.54	15.72	2.61
MNS	1.87	0.71	2.20	0.80	2.10	0.81	2.05	0.77	2.20	0.60

MC: mapping collection. SD: standard deviation. For traits abbreviations see Table 1.

5.2. GENOTYPING BY SEQUENCING: ADVANTAGES AND LIMITATIONS

Genome-wide association mapping (GWAS) and genomic selection (GS) have already been tried in some plant species and gave promising results (Nakaya and Isobe, 2012; Huang and Han, 2014). One of the major obstacles to their extensive use in plant genetics and breeding is the high cost of obtaining dense, genome-wide marker data for large samples. Ultimately next generation sequencing (NGS)-based SNP genotyping strategies, especially GBS (Elshire *et al.*, 2011) have emerged as a low cost tools for whole-genome genotyping. In this study we presented the first evaluation of the efficiency of GBS for grapevine genotyping with a perspective of GWAS. Using the GBS protocol at a multiplexing level of approximately 96-plex was possible to identify more than 300 thousand SNPs in a grapevine diversity panel of 274 accessions. After strict data filtering based on genotype quality, missingness, and MAF, we obtained a reliable dataset of 36,332 SNPs for 242 accessions with average missing values of 12% at a minimum MAF threshold of 0.01, and 25,961 SNPs at a minimum MAF threshold of 0.05. The SNP dataset was drastically reduced as filtered based on the proportion of missing values (Table 3). This is the major inconvenient of the GBS strategy, and is the result of low sequencing coverage (Davey *et al.*, 2011). Even if the GBS protocol was optimized for evenness of representation among samples (Elshire *et al.*, 2011), the variation in depth of coverage between individuals in our collection was extremely high (coefficient of variation > 1000%, far from the 43% reported in the original GBS protocol), resulting in large number of missing values for many individuals. Although numerical comparisons cannot be made because precise statistics were not always presented, such variations have been reported in practically all the studies that used GBS (Ward *et al.*, 2013; Barba *et al.*, 2014; Jarquín *et al.*, 2014). De Donato *et al.*, (2013) reported a coefficient of variation of read counts among individuals of 39%, but in that study only 47 cattle

reads had been sequenced in a 48-plex reaction (whereas we used 96-plex); in addition, for library preparation they used the *PstI* restriction enzyme that has a longer recognition site (CTGCAG), producing thus less fragments than *ApeKI* that has a recognition site of five bases with one degenerate site (GCWGC, where W can be A or T). According to Elshire *et al.*, (2011), the lack of accuracy in DNA quantification is the most likely source of sample-to-sample variation in sequence coverage. De Donato *et al.*, (2013) reported that this variation was associated with DNA quality expressed as 280/260 ratio. In our study, this variation was not significantly correlated with DNA concentration, neither with quality ratios. Although, we were not able to identify what is actually behind these large differences, broadly speaking, missing values in GBS can be attributed to both technical factors (library complexity and sequence coverage of the library) and biological (presence-absence variation, polymorphic restriction sites, and differential methylation) (Poland and Rife, 2012). Additionally, it seems that unevenness of representation among sequenced individuals tends to expand as the size and diversity of the sequenced population increase.

Two approaches can be used to deal with the problem of missing data in GBS:

- 1) The use of DNA with uniformly increased quality and purity, and the decrease of the multiplexing level (to 48-plex or less) should improve depth coverage and reduce the magnitude of variation between sequenced individuals. In fact, submitting the DNA samples to whole-genome amplification prior to library preparation increased significantly the depth of coverage (results not shown), and so did the decrease of multiplexing level from 288 to 96-plex. However, these measures cannot be adopted without a considerable increase in the overall genotyping costs.
- 2) The imputation of missing values using different methods. However, the accuracy of the imputed missing values and their contribution to the improvement of the power of statistical inference is not a matter of consensus (Poland *et al.*, 2012; Crossa *et al.*, 2013; Rutkoski *et al.*, 2013; Ward *et al.*, 2013; Jarquín *et al.*, 2014). Moreover, given that we are working with a collection of unrelated (or distantly related) individuals, displaying a drastic decay of LD, the imputation of missing genotypes at the obtained SNP density (roughly 300 thousand) will be practically a hard (if not impossible) task. Anyway, the imputation of missing genotypes is out of the scope of the current study.

Despite the high proportion of missing data, the GBS protocol adopted in this work followed by data filtering provided a reliable SNP dataset with marker density higher than what is currently offered by the available array-based techniques in grapevine. In addition, Datasets with high rates of missing values can be used without problem to study population structure and genetic diversity (Romay *et al.*, 2013), and Rutkoski *et al.* (2013) concluded that high levels of missing data in dense marker sets is not a major obstacle for genomic selection. By contrast, GWAS may

not tolerate such high proportions of missing values. So, the filtering process can be used as a flexible tool to select the desired SNP data volume and quality (mainly proportion of missing data and MAF) according to the targeted objectives.

More than half of the detected SNPs in the current study presented $MAF < 0.05$. Similar patterns of MAF distribution were reported in maize (Romay *et al.*, 2013) and cattle breeds (De Donato *et al.*, 2013). This overrepresentation of rare SNPs in GBS data seems to be reflecting the distribution of allele frequencies in natural populations because in the majority of species most alleles are rare (Myles *et al.*, 2009; Gravel *et al.*, 2011). Nevertheless, the contribution of sequencing errors to the inflation of the proportion of rare variants, especially at a frequency of $1/n$ (where n is the sample size) cannot be completely discarded. Array-based methods by contrast are designed to genotype common SNPs, but have limited potential to capture rare variants (McCarthy and Hirschhorn, 2008). This limitation of array-based techniques has been already reported in grapevine (Lijavetzky *et al.*, 2007). Although their role is not yet elucidated, rare variants are one of the focus points within the discussions about “missing heritability” from GWAS in humans (Manolio *et al.*, 2009). In grapevine, it is to expect that rare polymorphisms have an important role in the diversity of phenotypic traits as large proportion of it is the result of punctual somatic mutations in vegetatively propagated plants (Torregrosa *et al.*, 2011). However, the identification of such low-frequency causative polymorphisms through association studies may be challenging even when large sample sizes are used (Morrell *et al.*, 2011).

SNP density was found higher in gene-rich regions toward chromosome ends and lower in gene-poor regions towards the central regions (Fig. A.1). Interestingly, the distribution of gene-poor regions had been found highly correlated to the distribution of repetitive DNA-rich regions in the grapevine genome (Jaillon *et al.*, 2007). The same pattern of SNP density distribution along the chromosomes was reported in other species (De Donato *et al.*, 2013; Romay *et al.*, 2013; Wang *et al.*, 2013; Jarquín *et al.*, 2014; Schilling *et al.*, 2014). This pattern can be related to two factors: 1) The GBS protocol targets gene-rich regions (under-methylated) through the use of methylation sensitive restriction enzymes such as the *ApeKI*. This holds true even if DNA methylation is not specific to only non-coding regions and the methylation sensitivity of the enzyme is not strict. 2) The GBS pipeline discards any sequence tag aligned to more than one position, resulting in the reduction of the number of SNPs that can be mapped to repetitive DNA-rich regions. The sampling bias of methylation sensitive restriction enzymes towards gene-rich regions of the genome is a good reason to choose it as a genotyping tool for GS and GWAS, since most of the targeted causative polymorphisms are likely to be found there (Jorgenson and Witte, 2006).

5.3. MICROSATELLITE MULTIPLEX PCRs: A TOOL FOR WHOLE-GENOME GENOTYPING OF GRAPEVINE

Even if, SNP markers have been gaining popularity since the last decade due to their high frequency, ability to high throughput automated analysis and the decrease of their costs (Rafalski, 2002a; Davey *et al.*, 2011), microsatellite markers remain the markers of choice for many genetic studies due to their attractive features (high variability, codominance, transferability and reproducibility). This is particularly true for studies that require the use of relatively small number of markers and/or small samples. However, the use of large sets of SSR markers is limited due to the high cost and the amount of work required for their analysis, as well as the difficulties of automation. Those limitations could be partially overcome by adopting multiplexing strategies (Butler, 2005a; Butler, 2005b; Missiaggia and Grattapaglia, 2006; Blacket *et al.*, 2012) that would lead to significant savings in time, efforts and laboratory reagents (Elnifro *et al.*, 2000; Raabová *et al.*, 2010). Empirical results demonstrated that SSR-based multiplexing and multiloading reduced the cost of PCR reagents up to 50% and the cost of electrophoresis up to 85% when compared to genotyping based on single locus PCR (Masi *et al.*, 2003; Merdinoglu *et al.*, 2005). Guichoux *et al.* (2011) estimated that, even for a moderate number of samples, a 12-plex multiplexing could be eight times cheaper than simplex PCR. Additionally, genotyping errors due to the manipulator are less likely to occur, as laboratory manipulations are considerably reduced. Two approaches could be used for multiplexing: 1) the amplification of only one or a few markers, followed by pooling the products from the individual PCRs prior to electrophoresis (Hall *et al.*, 1996); 2) the joint amplification of all the loci to be loaded in the same electrophoresis run. In this study, we have used the second approach to address the development of an efficient tool that would allow grapevine genome-wide scanning using as fewer as possible PCR runs and electrophoresis tracks. Multiplex PCRs were developed following two main steps: 1) markers selection and organization in multiplex sets; and 2) the optimization of PCR conditions for the proposed multiplexes. In addition, one of the important issues in this work was the establishment of up to seven control points (see section 3.3.4) to avoid possible mistakes that could be very difficult to detect in large scale genotyping projects. The approach adopted in this study should be applicable for the development of similar tools in any other species.

5.3.1. Critical points in the process of multiplex PCRs development

SSR markers were selected based on three criteria: genome-wide coverage, high polymorphism and diversity in the fragment size ranges. In spite of the numerous efforts devoted to the development of SSR markers for grapevine genotyping since the early nineties, the number of

available SSRs remains limited when compared to other genome sequenced crop species. For instance, whereas 4,109, 8,192 and even 70,732 entries for SSR probes were found at the NCBI for soybean, wheat and rice respectively, only 1,079 entries were found for grapevine (<http://www.ncbi.nlm.nih.gov/probe>). In addition, for most of them (62.5%) fragment size ranges are concentrated between 150 and 250 bp (Fig. 15c). This limits marker combination possibilities, compromising hence the development of multiplex PCRs with high multiplexing levels (Hill *et al.*, 2009). Moreover, the available information is usually scarce or incomplete, since many of these markers have been developed and/or tested using only a few individuals. For example, in our PCR conditions, five UDV (Di Gaspero *et al.*, 2005), four VMC (*Vitis* Microsatellite Consortium) and one FAM (Huang *et al.*, 2011) SSRs amplified in several accessions more than two fragments in a narrow size interval, resulting in unreliable electrophoretic profiles. This could be related to the presence of multiple binding sites for their respective primers in the grapevine genome. For instance, the design of alternative primers allowed recovering monolocus segregation for four of them. In the same way, three loci described as monomorphic in *V. vinifera* were selected because they were only studied in four (VVIB54; Merdinoglu *et al.* 2005) or even two accessions of this species (FAM18 and FAM104; Huang *et al.* 2011), suggesting that more alleles could be detected in a larger sample. Our results revealed that it is unlikely to detect more than one allele at these loci in *V. vinifera*. Nevertheless, those markers, as well as other 21 that detected only 2 or 3 alleles (Table A.7), could be more informative in genetic backgrounds from other *Vitis* species. For example, the above mentioned monomorphic markers detected more than one allele when other *Vitis* species were analyzed (Merdinoglu *et al.*, 2005; Huang *et al.*, 2011).

On the other side, we found chromosomal regions spanning up to 10 Mb for which SSR markers had never been described (Fig. 13). Attempting to develop new markers, we identified many microsatellite motifs (up to 153/Mb) in the investigated regions across the grapevine genome, but successful primers could be designed only for a few of them. Moreover, PCR failures and/or irregularities in allele inheritance related to the presence of null alleles have been found for nearly half of the 22 new markers developed in this study. Difficulties with SSR markers development have been already reported in plants (Tero *et al.*, 2006) as well as animal species (Zhang, 2004; McInerney *et al.*, 2011), and linked to the presence of repetitive DNA and/or transposable elements. This might be also the case in our study, since most of the SSR gaps were located in low gene density regions of the grapevine genome (Fig. 27), which had been found substantially complementary to high density of repetitive/transposable elements (Jaillon *et al.*, 2007).

Figure 27. Localization of SSR gaps regarding the distribution of gene density along the grapevine chromosomes.

Vertical bars represent the 19 grapevine chromosomes, on which blue sectors represent regions covered by SSR markers from the literature (blue dashes) and white sectors represent SSR gaps that were investigated for the development of new SSR marks (red dashes) with low success rates. Vertical bar-plots on the right represent gene density distribution along the chromosomes, each bar in the plots indicate the number of annotated genes (<http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/>) in chromosomal bins of 500 kb.

Chromosomal distances are given in Mb according to the 12x version of the genome sequence build (<http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/>).

The second critical step for the development of the multiplex PCRs consisted in the establishment of suitable thermocycling programs and the adjustment of primers concentrations to obtain a balanced amplification for all the markers included in the same reaction. Touchdown PCR (Don *et al.*, 1991) proved to be highly suitable for multiplex amplification. In fact, only two touchdown-based programs (Tp-A and Tp-B; Table A.6) were enough to run 90% of the designed multiplex PCRs, which included primers with a wide range of annealing temperatures. As an example, Mx05 includes primers with annealing temperatures ranging from 48 up to 67°C. Touchdown based thermocycling programs combined with the use of commercial kits optimized for multiplex PCR contribute to save time and labor in the optimization steps when compared to conventional protocols (Masi *et al.*, 2003). In this study, they allowed the successful amplification of up to 15 primer-pairs in the same PCR reaction. As far as we know, this is the highest multiplexing level reached for microsatellite based grapevine genotyping. Merdinoglu *et al.* (2005) considered the preferential amplification of small fragments over long fragments for multiplex design, which limited the multiplexing level that could be reached. Our results demonstrated that it is possible to amplify fragments with size differences up to 329 bp in the same reaction; as an example, Mx19 includes 10 loci with fragment sizes ranging from 76 to 401 bp. Nevertheless, an average of 3.8 (between two and eight) optimization assays were needed to achieve balanced amplifications.

The 45 developed multiplex PCRs were tested and optimized using seven grapevine accessions. Their validity for genotyping large diversity panels was tested by genotyping 207 accessions representing most of the cultivated grapevine genetic diversity. Indeed, this collection was selected from the 1,852 *V. vinifera* accessions maintained at the germplasm bank of El Encín (Alcalá de Henares, Madrid, Spain), using the genotypic dataset of the 26 control loci through the genetic diversity maximization strategy (Le Cunff *et al.*, 2008). Moreover, the main progenitors of the current grapevine cultivars (Lacombe *et al.*, 2013) are present in this collection. The representativeness of the collection was also verified by comparison with two large grapevine germplasm collections: 1) An Italian collection of 745 accessions (Cipriani *et al.*, 2010): A set of 22 “Vchr” markers identified an average of 9.8 (3 to 21) alleles/locus in that collection, almost the same as in the core collection (between 3 and 20 with an average of 9.7). 2) A collection of 2,323 *Vitis vinifera* subsp. *sativa* cultivars conserved at INRA grape repository at Vassal (France) (Lauco *et al.*, 2011): Twenty out of the 26 control loci detected a lower number of alleles (between 5 and 19, with an average of 11.95) in the core collection than in the Vassal collection, where an average of 16.9 (6 to 36) alleles/locus were identified. However these differences are mainly related to the presence of uncommon genetic material carrying rare alleles in the larger French collection (2,323 vs 207 accessions). For instance, 9 and 12 alleles not detected in our study by the markers VMC4F3 and VVIV67 respectively had been detected

in Laucou *et al.* (2011), but with frequencies lower than 0.005. These findings suggest that, although additional alleles may be detected when studying more distant genetic material, the allelic ranges obtained using the core collection should represent a good estimation of the actual diversity that can be found in cultivated grapevine.

5.3.2. Potential applications of the developed multiplex PCRs

As in many other woody plant species, genetic mapping in grapevine has been carried out using mainly the double pseudo test-cross strategy (Grattapaglia *et al.*, 1995), which is based on the study of allelic segregation of markers found in heterozygosis in one or both progenitors of an F₁ progeny. The development of new genotyping platforms and technologies has increased the number of SNPs that can be genotyped to hundreds of thousands allowing the construction of highly saturated genetic maps. However, with the currently available information, these markers are not suitable for a direct comparison with the previously published genetic maps and QTLs in grapevine, which are predominantly based on microsatellite markers. The number of markers that can be genotyped using the panel of multiplex PCRs developed in this study is large enough to allow that comparisons. For example, the analysis of the segregation types inferred from the SSR data set over the core collection suggests that, in a supposed project involving a mapping progeny derived from a cross between Cabernet Sauvignon and Pinot Blanc, this tool would allow to map 235 out of the 264 studied loci, with an average of 12.37 per chromosome (from seven in chromosome 10, to 17 in chromosome 5).

American and Asian *Vitis* species constitute a valuable source of resistance genes that can be introduced into *V. vinifera* varieties through interspecific breeding programs (Töpfer *et al.*, 2011). The limiting step in this kind of programs is the recovery of the genetic background of the recurrent parent (*vinifera*) by backcrossing. The assistance of the selection process by genome-wide distributed SSRs organized in ready to use multiplex PCRs, such as the developed in this study, has the potential to make considerable savings in time and costs. The efficiency of such a tool could be improved by organizing SSRs from each chromosome in a distinct multiplex PCR as suggested in a previous simulation study (Herzog *et al.*, 2013).

Multiplexes Mx01 and Mx02 allow the study of 26 unlinked markers, including at least one in each of the 19 *V. vinifera* chromosomes. They were designed re-organizing the multiplex PCRs S, A and B used by Ibanez *et al.*, (2009) to genotype 376 table grape accessions. These 26 microsatellites have been used to genotype the complete germplasm bank of El Encín (Alcalá de Henares, Madrid, Spain). Twenty of them have been also used through eight multiplex PCRs and three sequencing runs to characterize most of the INRA (France) grape repository (Laucou *et al.*, 2011). Additionally, smaller subsets of the control loci used in this study (including the six OIV and the nine GrapeGen06 SSRs) have been used to characterize grapevine genetic resources

from most of the viticultural regions in the world (This *et al.*, 2011). Given the large amount of information available for these markers, Mx01 and Mx02 can be used for a rapid and cost efficient characterization of unexplored grapevine germplasm and short range genetic studies, such as parentage analysis. In this work we have used Mx01 and Mx02 to certify the identity of each of the extracted DNAs. Moreover, the inclusion of one of the loci amplified with Mx01 and Mx02 as control markers in each multiplex PCR allowed to verify assay reproducibility and sample traceability throughout the study, factors that should be considered by any researcher aware of the consequences of genotyping errors (Pompanon *et al.*, 2005).

Although important cost and time savings might be obtained using the panel of multiplex PCRs presented in this study, a considerable investment in primers labeling will be still required. Economic labeling methods could be used to further decrease the costs. As an example, Blacket *et al.* (2012) proposed the use of four fluorescently labelled universal primers (one primer for each fluorescent dye) through the three primer-PCR approach. This kind of techniques would offer an inexpensive alternative to the commercial synthesizing of custom labeled primers for multiplex genotyping. However, it should be noted that additional optimization steps may be needed, because sometimes the three primers-approach seem to decrease PCR efficiency (de Arruda *et al.*, 2010).

5.3.3. The problem of Null alleles

The presence of null alleles at microsatellite loci can introduce important biases into genetic studies (Callen *et al.*, 1993; Pompanon *et al.*, 2005). Their existence in grapevine have been already demonstrated by sequence analysis (Sefc *et al.*, 1999). Another reliable approach for null allele detection is the analysis of allelic inheritance in family groups (Dakin and Avise, 2004). Using this approach, we identified the probable existence of null alleles in 49 out of the 284 analyzed primer pairs. Most of them (69.39%) showed estimated null allele frequencies > 0.10 . However, the absence of null alleles in the 14 studied trio pedigrees does not discard their presence in the rest of the genotypes. In fact, 22% of the loci with a moderate to high frequency of null alleles ($0.10 < F < 0.34$) did not show any incompatible genotype among the studied pedigrees. On the other hand, incompatible genotypes were detected for 67.5% of primer pairs that, after PCR repetition, failed to amplify at least one sample. These results suggest that most of the finally recorded PCR failures could be related to the presence of null alleles at homozygote state. Out of the 32 markers showing null alleles for which alternative primers were designed, 18 recovered the normal segregation of alleles, decreasing the estimated null allele frequency in most of the cases (from 0.170 to 0.002 on average), and no amplification failures were noticed for any redesigned primer pair.

5.4. POPULATION STRUCTURE

The diversity panel studied here is composed of cultivars from 34 countries around the world, representing most of the cultivated grapevine gene pool. STRUCTURE analysis revealed that the studied collection is composed of a large proportion of admixed genotypes (46% at membership probability threshold of 0.75), owing most likely to the selection and breeding history of this species, and to the vegetative propagation that lead to the overlapping of multiple levels of family relationships in the same gene pool. However this does not prevent the detection of a consistent structuring level at $K = 3$ reflecting very well the different geographical origins and human use of the clustered cultivars. In fact, while q1 includes, almost exclusively, wine grapes from the Western Europe (mainly, France and Spain); q2 includes table grapes mainly from Spain and oriental countries, but also 19 wine grape cultivars, of which 16 are from the Iberian Peninsula (among others Albillo Mayor, Bobal, Hebén and Tempranillo). According to Felix Cabello (personal communication), most of these wine cultivars should have been derived from table grapes introduced by Muslims into Al-Andalus during the Middle Ages. For this reason they are genetically more linked with table cultivars in q2 rather than wine cultivars in q1. The third group (q3) is also dominated by table grapes (35 out of 45), but mostly from the Italian Peninsula and to less extent from the Eastern Europe.

So far, the most comprehensive description of genetic structure in grapevine was reported by Bacilieri *et al.*, (2013). These authors analyzed the genetic structure within a collection of 2,096 grape cultivars from the French grapevine collection of Vassal (INRA, France), using a set of 20 microsatellite markers. They found that the most relevant structuring level identified three clusters: Wine-West and Central Europe, Wine-Balkans and East Europe, and Table-East. Moreover, they identified a second structuring level ($K = 5$) that revealed two additional groups: a geographic group from the Iberian Peninsula and Maghreb, and a group comprising table grapes from Italy and Central Europe. Although not exactly similar, our results agree substantially with the findings reported by these authors. Indeed, broadly speaking, while our clusters q1 and q2 correspond respectively to Wine-West and Central Europe and Table-East revealed at $K = 3$ in Bacilieri's study, our q3 group corresponds to Table-Italian and Central Europe that was identified at $K = 5$ (whereas it was assigned to the admixed group at $K = 3$). These marginal discrepancies should be the result of sampling bias: our collection includes a large number of European cultivars (mostly from Spain (81), Italy (51) and France (36)), but, the Oriental cultivars are somewhat underrepresented. On the other hand, the three-level clustering found in the current study is in agreement with the eco-geographic based classification proposed by Negrul (Negrul, 1946). Actually, our clusters q1, q2 and q3 are globally equivalent to Negrul's *proles*, *occidentalis*, *orientalis* and *pontica*, respectively. Furthermore our results

substantially agrees with other population structure studies in grapevine (Aradhya *et al.*, 2003; Emanuelli *et al.*, 2013).

5.5. PATTERNS OF LINKAGE DISEQUILIBRIUM IN GRAPEVINE

Linkage disequilibrium (LD) is the non-random association between alleles at different loci. A variety of statistical measures have been proposed to assess the variation of linkage disequilibrium and most of them are used to quantify the degree of association between marker pairs. Knowledge of the extent and structure of LD is important to assess the usefulness of association mapping strategies. The decay of LD over physical or genetic distance determines the resolution as well as the density of markers needed for association analysis (Kruglyak, 1999; Pritchard and Przeworski, 2001; Ardlie *et al.*, 2002; Flint-Garcia *et al.*, 2003). In the current study we performed an in-depth examination of LD over the grapevine genome using both SNP and SSR markers.

5.5.1. LD assessed using SNP markers

We presented here a detailed view of genome-wide LD in grapevine using a high density GBS-SNPs (25K) over a diversity panel of 242 accessions representing a large proportion of cultivated grapevine gene pool. Since LD is a property of marker pairs, it is difficult to obtain summary statistics of LD across the whole-genome or part of it (Gaut and Long, 2003). In this study, we used a combination of graphical representations (LD decay plots, average LD across bins of physical distance) and numerical estimations (distance of LD decay based on logarithmic regression, percent of LD beyond a given threshold) to investigate the distribution of LD across the genome and how it is affected by some factors. So far, this is the most comprehensive LD assessment carried out in grapevine.

The average genome-wide LD between linked SNPs (within chromosomes) decayed very rapidly to $r^2 = 0.2$ within a distance lower than 1 kb. Previous studies focused on short genomic regions (Lijavetzky *et al.*, 2007; This *et al.*, 2007; Vargas *et al.*, 2013a; Vargas *et al.*, 2013b) and a more comprehensive study based on 3,349 genome-wide distributed SNPs (Myles *et al.*, 2011) indicated that the levels of LD may be low in grapevine. Similar estimates of LD decay have been reported in the USA maize seed bank (2,815 accessions), using a panel of more than 600K GBS-SNPs (Romay *et al.*, 2013). However, more recently, Nicolas *et al.* (2016) reported that LD in grapevine reached $r^2 = 0.2$ within a distance ranging between 9 and 458 Kb. This discrepancy may be explained by two hypotheses: firstly, these findings are based on the set of 372 SNPs from four 2 Mb genomic regions (chromosomes 8, 19, 12, and 17), whereas our study is genome-wide; secondly, while Nicolas *et al.* (2016) used the Hill and Weir's model, we used the logarithmic function to estimate the extent of LD.

Our data indicated that LD varies significantly between chromosomes, with a slight negative correlation with chromosome size (shorter chromosomes exhibited higher average LD than longer ones). By contrast, Khatkar *et al.* (2008) reported a positive relationship between LD and chromosome size, hypothesizing that this could be related to a higher recombination frequency per unit of physical distance in long chromosomes than in shorter ones. However, LD is affected by several other factors besides recombination, which are likely implicated in creating this tendency. Indeed, artificial selection for example extends LD in the genomic regions associated to domestication traits. In this sense, the highest levels of LD in the present study were found in chr2 and chr17 that contain genetic loci highly involved in the domestication process of grapevine: flower sex and berry color in chr2 and berry size in chr17. The involvement of these two chromosomes in grapevine domestication was also demonstrated by Myles *et al.* (2011) who found large selection sweeps (5Mb) in them. But at the same time, they are the shortest chromosomes in the grapevine genome.

5.5.1.1. Effect of MAF, sample size and population structure on LD estimates

The patterns of LD can be affected by different factors, especially allelic frequencies and size and structure of the population used. The use of several ad-hoc generated datasets with specific filtering criteria has allowed us to understand more thoroughly the effect of these factors in our study.

The minimum minor allele frequency affected LD estimates especially at short distances: higher MAF thresholds induced an increase in the mean r^2 values, which is in agreement with previous studies (Khatkar *et al.*, 2008; Yan *et al.*, 2009; Qanbari *et al.*, 2010). This tendency can be explained by the mathematical properties of the r^2 statistic: LD between two markers estimated as r^2 cannot reach the unity unless allele frequencies of the markers in question are identical (Flint-Garcia *et al.*, 2003); and the decrease of minimum MAF in a SNP dataset is accompanied by an increase in the number of SNP pairs with divergent allele frequencies, inducing thus a reduction in r^2 . This means that estimates of LD can be biased by the used SNP dataset. Indeed, array-based techniques over-sample SNPs with medium to higher MAF, whereas GBS over-samples SNPs with low MAF. Thus, short range LD estimated using GBS-SNPs could be lower than the estimated using array based SNPs, under the same conditions. As an example, the distance of LD decay in maize estimated using an Illumina GoldenGate assay (5 to 10 kb; Yan *et al.*, 2009) was higher than that estimated using GBS-SNPs (1kb; Romay *et al.*, 2013).

On the other hand, our results showed that LD estimates are biased upward when small sample sizes are used. The same trend was observed in other species, including humans (Weiss and Clark, 2002), cattle (Khatkar *et al.*, 2008) and maize (Yan *et al.*, 2009). For unbiased and accurate assessment of LD in humans, Weiss and Clark (2002) think that the sample size of 100

individuals should be adequate; and Khatkar *et al.* (2008) estimated that a minimum sample size of 75 individuals is required in cattle. Our results suggest that in grapevine, the accuracy of LD estimates obtained using roughly less than 30 individuals may be significantly compromised, especially at inter-marker distances higher than 1 kb.

Regarding the effect of population structure, in a population that presents a clear substructure, high levels of LD should be expected between distantly located markers in the genome, and even between markers located at different chromosomes (Pritchard and Przeworski, 2001). The genetic structure of cultivated grapevine can be largely understood as a complex pedigree, due to the vast number of higher order familial relationships. Current grapevine populations shares a common evolutionary history starting from a domestication process with weak bottleneck effect and slight geographic isolation, followed by human breeding and vegetative propagation with probably a continuous introduction of genetic material from the wild ancestor (*V. vinifera silvestris*) (Myles *et al.*, 2011). In our study, even if a relevant population substructure was detected in the mapping collection, the level of LD detected between unlinked SNPs was very small. The most likely explanation is that the population structure is probably too weak to induce important spurious LD between unlinked SNPs. Moreover, unlike humans (Reich *et al.*, 2001; Ardlie *et al.*, 2002) or maize (Romay *et al.*, 2013), that displayed large differences in LD distribution among the different studied populations or germplasm groups, in our study, no significant variation in the extent of LD was found among the subpopulations identified by STRUCTURE.

5.5.2. LD assessed using SSR markers

Linkage disequilibrium in *V. vinifera* has been scarcely studied, especially with SSR markers. Barnaud *et al.* (2006) reported the first study of LD in cultivated grapevine using a core collection of 141 cultivars and 38 microsatellite loci located on five linkage groups. Significant genotypic LD was found only within linkage groups, with r^2 values declining to around 0.1 within 5 cM (650–1,080 kb; for genotypic data) or 10 cM (1,300–2,160 kb; for haplotypic data). Our study is the first one to provide a genome-wide analysis of LD in grapevine using a set of 250 SSRs over a diversity panel of 207 grapevine accessions, representing most of the cultivated grapevine gene pool. We found that LD between linked SSRs (i.e. within chromosomes) presented substantial variation among the chromosomes. The distances of LD decay under the critical value of 0.2 ranged from a maximum of 186 kb in chr2 to 1 pb in chr14, but it was not possible to estimate this distance for chr16, 17 and 19 because the obtained r^2 values in these chromosomes were so low that the logarithmic tendency line didn't cross the baseline of $r^2 = 0.2$ (Fig. A.3 and Table 6). These results suggest that the estimated distances of LD decay were probably biased downward in more chromosomes, particularly those showing extremely short

distances of LD decay. This is most likely the result of the lack of representativeness of short range LD (i.e. LD across distances < 1 Mb) due to insufficient marker density. In fact, only 66 out of 1,611 pairwise r^2 values corresponded to inter-marker distances shorter than 1Mb, being 0.27 kb the closest distance between two SSRs in our study. Anyway, at the whole-genome level, our estimates indicated that linked LD decayed below the critical value of $r^2 = 0.2$ within an average distance of 0.24 kb. However, the overall pattern of LD decay across the whole genome (without considering the logarithmic tendency line) as depicted by data-points in Fig. A.3 (“All chr”) is in accordance with the results reported by Barnaud *et al.* (2006) (see Fig. 4 in that study). Moreover, low pairwise r^2 values between SSR markers have been also reported in other outcrossing plant species, such as *Eucalyptus* (Arumugasundaram *et al.*, 2011), maize (Van Inghelandt *et al.*, 2011) and ryegrass (Li *et al.*, 2011).

5.5.3. Comparison of SNP-based vs. SSR-based LD

LD is known to be affected by several evolutionary forces such as mutation, recombination, selection and population demography (Flint-Garcia *et al.*, 2003; Gaut and Long, 2003; Gupta *et al.*, 2005). In addition, a large part of LD variation among genomic regions is still linked to unknown factors, probably related to local selection or to differences in recombination rate or genomic structure (Nicolas *et al.*, 2016). From a practical point of view, the type of molecular marker used to examine LD also influences its patterns (Mueller, 2004). This point has not received much interest in genetic studies, and in grapevine particularly, this has not been examined before. The present study allowed for a direct comparison between SNP- and SSR-based LD in the same research population. Our results indicated that LD in *V. vinifera* decline rapidly with physical distance regardless whether it is estimated using SNPs or SSRs. However, due to the underrepresentation of short-range SSR-LD mentioned above, an accurate comparison between SNP-LD and SSR-LD can be performed only until a minimum distance of 0.1 to 1 Mb (Table 11). Mean SSR-based r^2 values were only slightly higher than SNP-based r^2 values, according to the considered distance class. Moreover, the comparison between the distances of LD decay below the critical value of $r^2 = 0.2$ as estimated using SNP vs. SSR markers (Fig. 28) showed that at the whole genome level, the extent of SNP-based LD was slightly higher than SSR-based LD (0.94 versus 0.24 kb), although this tendency was not uniform across all the chromosomes: while in chromosomes 1, 6, 7, 8, 10, 11, 13, 16, 17, and 19, SNP-based LD was markedly more extended than SSR-based LD, this trend was reversed in chromosomes 2, 4, 9, 15 and 18, and there was almost no difference in chromosomes 3, 5, 12 and 14.

Table 11. Comparison between SNP and SSR-based LD at different categorical physical distances

Distance intervals	Number of SNP pairs	Number of SSR pairs	Mean r^2_{SNPs}	Mean r^2_{SSRs}
0-10bp	5,832	-	0.606	?
10-100bp	15,077	-	0.365	?
0.1-1kb	5,934	1	0.245	0.327
1-10kb	15,920	-	0.179	?
10-100kb	117,069	3	0.116	0.031
0.1-1Mb	1,033,605	62	0.069	0.073
1-10Mb	6,407,934	888	0.026	0.033
>10Mb	5,075,659	657	0.013	0.015

Figure 28. Distances of LD decay below $r^2 = 0.2$ estimated using SNP vs. SSR markers

In the genetics literature, it has been largely accepted that the rapidly evolving multiallelic SSRs due to high mutation rates (Gulcher, 2012) are more effective in capturing LD than do the older bi-allelic SNP markers (Slatkin, 1994; Chapman and Wijsman, 1998; Xiong and Jin, 1999; Ohashi and Tokunaga, 2003). This has been also confirmed by empirical findings in some organisms, such as humans (Abecasis *et al.*, 2001; Varilo *et al.*, 2003; Payseur *et al.*, 2008) and maize (Remington *et al.*, 2001). Our results are contradictory (at least in part) with these studies. The most plausible explanation of this contradiction is that the estimated distances of SSR-based LD decay in several chromosomes could be biased downward because of the low SSR density employed in this study as compared to SNP density (Fig. 9 and Fig. 13). On the other hand, the non-uniform discrepancies between the chromosomes could be related to the stochastic nature of LD (i.e. unpredictably, molecular markers that are immediately adjacent to each other on a chromosome might be statistically independent, whereas those that are far apart might be highly correlated; Weiss and Clark, 2002; Carlson *et al.*, 2004), therefore, higher r^2 values might be missed just because of the sampling bias. The most obvious example that illustrates the

deficiency of SSR density should be Chr17. In this chromosome, with only seven SSRs genotyped, being the closest ones 0.62 Mb apart, the highest SSR-based r^2 value detected was 0.088 for a distance of 2.85 Mb. As a consequence, when the logarithmic function was used to estimate LD decay below $r^2 = 0.2$, it returned a value of 0 kb, which does not make biological sense. By contrast, when assessed using the SNP dataset, this chromosome showed the most extended LD over the genome (6.79 kb for $r^2 = 0.2$), likely due to the presence of a large selection sweep around the 6 Mb region (Myles *et al.*, 2011). These results are in agreement with a similar work conducted by Van Inghelandt *et al.* (2011), studying the extent of LD between unlinked, linked and adjacent SSR and SNP markers in a collection of 1,537 commercial maize inbred lines. These authors found that unlike SNPs, SSRs detected low r^2 values almost equal between unlinked, linked and even adjacent marker pairs, concluding that this was due to insufficient density of SSR markers. Breseghello and Sorrells (2006) also attributed the lack of consistent LD between 33 SSRs in chromosome 2D of wheat (0 to 15 cM spaced) to insufficient marker density. Nevertheless, even if multiple arguments support the hypothesis that the SSR density used in our study is insufficient for capturing a complete and accurate picture of LD in grapevine, it need to be verified by genotyping a higher number of SSRs, at least in a reduced sample of chromosomes.

5.5.4. Implications for GWAS

The pattern of LD decay across a genome has often been analyzed to estimate the density of markers to use in whole genome association scans and subsequent marker assisted selection (MAS), and to determine the resolution and power of QTL detection. If LD in the genome extends to large distances, small number of markers will be needed to cover the genome but the resolution of GWAS will be poor. By contrast, if LD level is low, QTLs could be detected with higher resolution but excessive number of markers will be required in order to perform GWAS with sufficient statistical power. Similarly for MAS, if a particular gene is under high LD, marker-assisted selection can be efficiently performed with a single marker irrespective of its location in or near the gene. By contrast, if a gene is under low LD, several markers at various positions of the gene must be used to transfer a complete allele because recombination can occur within the gene.

In order to interpret LD in terms of power to detect associations and determine the required SNP density, the useful level of LD need to be defined (Ardlie *et al.*, 2002). In the literature, there is no consensus about what is the useful level of LD. Arbitrary values of r^2 ranging from 0.1 to 0.8 with a bunch of intermediate values (Reich *et al.*, 2001; Ardlie *et al.*, 2002; Barrett and Cardon, 2006; Zhu *et al.*, 2008; Yan *et al.*, 2009; Qanbari *et al.*, 2010; Van Inghelandt *et al.*, 2011) or r^2

thresholds based on more elaborated concepts, such as the 95th percentile (Q_{95}) of unlinked r^2 distribution (Brescaglio and Sorrells, 2006a) have been proposed and/or used as useful levels of LD. The lack of consensus in this matter can be linked to different factors, among others: the species in question, the size of the mapping population (inversely related to r^2 (Kruglyak, 1999)), the effect size of the targeted-QTL (high level of LD is needed to detect associations with small-effect QTLs) and the rate of LD decay itself (authors tend to choose higher r^2 values as a threshold when LD is extensive and *vice-versa*).

According to Ersoz *et al.* (2007), the r^2 threshold of 0.1 can be taken as the minimum value for low power GWAS. Our study indicated that to achieve such a level in grapevine, approximately 6,500 SNPs would be sufficient (results not shown). However, only doubling the targeted minimum r^2 threshold ($r^2 = 0.2$) implies multiplying the number of required SNPs by more than 500 times (more than 3 million SNPs). Assuming the actual nucleotide diversity in grapevine to be 1 SNP each 64 bp (Lijavetzky *et al.*, 2007), the total number of SNPs that can be found in the whole genome would be around seven millions. Therefore, to achieve an r^2 of 0.2, it would be needed to sequence half the genome (see Table 5 and Table 6 for SNP and SSR density estimations at chromosome and genome-wide levels), and to perform high powered GWAS (say $r^2 \geq 0.6$), whole-genome sequencing will be practically required. However, given that a large proportion of the grapevine genome consists of repetitive DNA/transposable elements (Jaillon *et al.*, 2007), and since most of the interesting polymorphisms would be located within or in the proximity of coding regions (Jorgenson and Witte, 2006), the required SNP density for a successful GWAS may be significantly lower if only SNPs from coding regions are to be genotyped. Here, the GBS approach will have a great advantage because as demonstrated above, it over-samples SNPs from the coding regions. Furthermore, these inferences are based on the assumption that pairwise LD reflects the physical distance between SNPs. However, this should be taken with much caution because although a clear relationship exists between r^2 values and physical distance, it is far from smooth as can be seen in LD decay plots (Fig. A.2): close SNPs exhibited smaller r^2 values, while distant SNPs exhibited larger r^2 values than one should expect and *vice versa*. This is a typical characteristic of LD owing to a complex interplay of several factors such as recombination, mutations, population demography and selection throughout the evolutionary history of the species (Ardlie *et al.*, 2002; Nordborg and Tavaré, 2002; Flint-Garcia *et al.*, 2003).

Based on the hypothesis that SSRs have increased power in detecting LD when compared to SNPs (Slatkin, 1994; Chapman and Wijsman, 1998; Xiong and Jin, 1999; Ohashi and Tokunaga, 2003), it should be possible to map genetic determinants of quantitatively inherited traits using a sparse set of SSRs covering all the genome through GWAS approaches. However, the SSR

density required to carry out an effective GWAS has never been clearly determined, particularly in grapevine, because of the lack of a complete understanding of LD patterns in this species. As stated before, our results suggest that SSR-based LD in grapevine would not extend much further than SNP-based LD. In fact, even if we consider only the LD decay distances from the chromosomes where SSR-based LD was higher than SNP-based LD (chromosomes 2, 4, 9, 15 and 18), the average distance of LD decay below $r^2 = 0.2$ will not exceed 50 kb. This way, to perform a GWAS with reasonable power, at least 10,000 SSR markers would be required. However, to carry out a high powered GWAS, the number would need to be increased several orders of magnitude further, which would be a prohibitive task if the genotyping is to be carried out using conventional approaches. Tamiya *et al.* (2005) estimated that SSR-based LD in humans extends up to 100 kb, concluding that about 27,000 SSRs would be needed to cover the euchromatin part of the human genome. In order to perform GWAS using such a large amount of microsatellites while avoiding the associated technical burden, human geneticists implemented a multistage strategy based on DNA pooling method (Sham *et al.*, 2002).

5.6. GENOME-WIDE ASSOCIATION ANALYSIS

5.6.1. GWAS using SNP markers

5.6.1.1. Association analysis models: efficiency in controlling for false positives

One of the most challenging constraints in association mapping is the occurrence of false associations due predominantly to the presence of cryptic population structure and familial relatedness within the mapping population. In order to deal with this problem, several statistical approaches have been developed (Devlin and Roeder, 1999; Pritchard *et al.*, 2000b; Price *et al.*, 2006; Yu *et al.*, 2006). We have compared four association analysis models with increasing complexity (GLM, Q, K and Q+K models respectively) in order to select the best one in reducing false positives. The naïve GLM model was found non suitable for association analysis because, it falsely increase the number of genetic associations throughout the genome, due to population structure and relatedness confounding effects. The mixed model that accounts for pairwise relatedness through the kinship matrix (K model), was more effective than the GLM model that incorporates the inferred population structure as cofactor (Q model). However, when the targeted phenotype is strongly correlated to population subdivision (such as muscat flavor that was almost only present in q3 subpopulation), the Q model reduced significantly the rate of false positives (Fig. 25). Finally, the combination of both population structure and pairwise relatedness in the same model (Q+K model) reduced further the rate of false positive

associations, independently of the degree of correlation of the phenotype to the population structure.

The superiority of K over Q model was to be expected because unlike the Q matrix that takes into account only a few axes of variation, the K matrix captures the relatedness between each possible pair of individuals in the mapping population (Myles *et al.*, 2009). In grapevine context this makes even more sense, because the currently cultivated grapes can be regarded as a large family network where different levels of ancestry coexist simultaneously rather than admixture of random mating subpopulations; and, the complex pattern of relatedness is better captured by K than Q matrix. Although, this scenario is not plausible in the case of maize and *Arabidopsis*, similar results have been obtained in these species (Yu *et al.*, 2006; Zhao *et al.*, 2007; Atwell *et al.*, 2010). By contrast, Del Carpio *et al.*, (2011) found the Q model more effective than the K model in deflating the significantly associated markers to *Brassica rapa* metabolites. These contrasting results indicate that the optimal association analysis model could be species-specific and should be determined on a case by case basis. However, the efficiency of mixed models in reducing confounding is limited, because spurious associations could be still detectable after correction for both, structure and relatedness, even for the simplest traits, such berry skin color and flower sex. Available information points towards these traits are regulated by single genetic loci located on chr2. However, significant associations with these traits were detected respectively on three and 14 chromosomes more, even when the most stringent model (Q+K) was used for the analysis. According to Platt *et al.*, (2010), spurious associations that cannot be eliminated by correcting for population structure may arise from three sources: correlation between causal polymorphisms and unlinked non causal markers; existence of more than a single causal polymorphism (especially when they are correlated); and epistasis. Additionally, the higher number of false discoveries for flower sex comparing to berry skin color may be attributed to the unbalanced distribution of this trait (Fig. 5) (Stram, 2014). In parallel, our results demonstrated, as previous studies (Zhao *et al.*, 2007), that reducing the rate of false positives is at the cost of eliminating true positives too, because they could be highly correlated with the underlying population structure or relatedness. This was the case of seed content association in 2012: while a significant association with seed dry weight was identified on chr18 near the major effect seedlessness QTL using the naïve GLM or Q model, the association signal was completely lost when the GWAS was performed using the K or Q+K models (Table 7).

5.6.1.2. Detection power and resolution of GWAS

Strong and/or stable associations were detected in this study for berry skin color, flower sex, muscat flavor, seed content traits, and berry size traits. The common features among these are

high heritability, low sensitivity to environmental factors and simple to moderately complex genetic inheritance (known to be controlled by major genes or QTLs that have been confirmed several times in different genetic backgrounds (Table 8)). One of the prominent properties of the identified associations is that many of them presented a complex diffuse shape (spanning up to several Mega-bases) hindering sometimes the determination of the center of the peak and/or to decide whether the pool of the associated SNPs correspond to one or multiple peaks (loci). This was the case of berry color association for example, which involved 57 SNPs spanning a total distance of 10.99 Mb on chr2 (Table A.9). Additionally, the associated SNPs and the strength of their association to a given trait were not always similar from one year to another. This hampers to know whether associations identified next to each other or overlapping in the same chromosome correspond to a single locus, or multiple different loci arranged together in the same region. The seed content association on chr1 for example involved a total of 21 SNPs spanning a distance of 17.44 Mb (from 5,789,433 bp to 23,238,804 bp) if we consider all the significantly associated SNPs regardless of the year of detection, seed content sub-trait (mean number of seeds (MNS) and mean seed dry weight (MSDW)). However, the number and location of the associated SNPs (and the association peak) are highly variable among the two sub-traits and between the years of study within each sub-trait, making it difficult to identify the center of the locus and whether there are one or multiple loci. Similar patterns of genome-wide associations have been reported previously in *Arabidopsis* (Atwell *et al.*, 2010) and in grapevine for berry skin color (Myles *et al.*, 2011). The latter was attributed to the extended LD in chr2, found also in our study. This complicates further the follow-up studies such as fine mapping or candidate gene testing in order to identify functional variants and questions one of the major advantages of GWAS over the classical linkage mapping, its higher resolution power. The involvement of extended LD in inducing such a pattern of association in chr2 is obvious; however, extensive LD alone cannot explain this phenomenon everywhere in the genome, because as we showed earlier in agreement with previous studies, LD decays very rapidly in grapevine.

On the other hand, only a few non-stable associations were detected for ripening time, cluster weight and compactness index, berry length and shape index and flesh firmness. Whereas, no significant associations (under the Q+K model) were detected for fertility, cluster length and diameter, peduncle length, and all phenology traits but ripening time. The lack of power in detecting genetic associations for these traits may be attributed to one or the interaction of several factors related to the genetic architecture of the trait, the mapping population (size and composition), the genotypic data and the statistical analysis.

GWAS use to have difficulties when used for traits with complex genetic architecture (Korte and Farlow, 2013). This kind of traits are usually characterized by low heritability, large environmental dependency and the involvement of a large number of genes of small effects in the genetic determinism of the trait. In *Arabidopsis*, meaningful results have been obtained for traits underpinned by only a few loci that explain a large proportion of the phenotypic variance using a mapping panel of less than 100 accessions (Atwell *et al.*, 2010). However, the situation is different for traits involving a large number of small effect loci, which use to require several thousands of individuals (Korte and Farlow, 2013). On the other side, phenotypic data were collected from a germplasm collection that was established for the purpose of germplasm conservation rather than conducting association studies. This material is more suitable for the study of highly heritable and domestication related traits (Brescaglio and Sorrells, 2006b), whereas traits with more complicated genetic control, such as yield, would require specific designs based on replicated trials.

Myles *et al.* (2011) suggested that the rapid decay of LD in grape would require millions of SNPs to perform high powered GWAS. Our study is based on a set of 36,332 genome-wide distributed SNPs. It is to expect that several genes involved in the expression of phenotypic traits, particularly of complex genetic control, may have been missed just because of insufficient coverage of the genome. On the other side, the power to detect an association depends on the allelic frequency of the functional variant in the sample. Even if a low frequency allele has an enormous effect on the phenotype, GWAS will have a low power to detect it (Manolio *et al.*, 2009; Myles *et al.*, 2009). Our SNP dataset is biased towards rare alleles (28% and 47% of the SNPs have minor allele frequency lower than 0.05 and 0.1 respectively), which could lead to the non-detection of associations if the frequency spectrum of functional alleles is similar to frequency spectrum of random SNPs.

Another important issue in association mapping is the selection of the significance threshold. The genome-wide significance threshold used in this study was determined using the Bonferroni correction assuming that all the SNP-trait association tests are independent from each other. This assumption is likely to be violated owing to the presence of marker pairs in extensive LD throughout the genome, resulting hence in a somewhat conservative threshold of significance (Balding, 2006). The very stringent corrections used in the association analyses to remove false positives (due to population structure, multiple testing issues, etc.) would have likely removed a substantial number of true-positives.

Fertility is an example of a low heritable ($H < 0.11$; Fanizza *et al.*, 2005) and highly environment-dependent trait (Fig. 5) (Sommer *et al.*, 2000; Sanchez and Dokoozlian, 2005). QTL mapping in biparental populations has allowed the detection of only medium effect QTLs

($R^2 < 20\%$), mostly unstable over years (Fanizza *et al.*, 2005; Marguerit *et al.*, 2009; Doligez *et al.*, 2010; Grzeskowiak *et al.*, 2013), and with large discrepancies in locations and effects, confirming the environmental dependency and the complex genetic architecture of this trait. We have found evidences about the involvement of population structure in hampering the detection of significant associations for this trait. Visual examination of Manhattan plots derived from GLM analysis (Fig. A.6) allows to see an association peak on chr5, in the same region as the stable QTL discovered by Doligez *et al.*, (2010). However, the inclusion of population structure and pairwise relatedness in the model (Q+K model) resulted in an overcorrection that maintained the association signal below the genome-wide significance threshold. This is most likely the consequence of the significant correlation of fertility to population structure, as revealed in Fig. 17.

Unlike fertility, phenology related traits are characterized by their very high heritability estimates (0.89 to 0.92; Duchêne *et al.*, 2012), suggesting that a rather large part of the phenotypic variation of these traits is of genetic origin. However, the non-detection of associations in our study and the lack of overlapping results among QTL studies (Fischer *et al.*, 2004; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Marguerit *et al.*, 2009; Duchêne *et al.*, 2012; Grzeskowiak *et al.*, 2013) may indicate that in grapevine developmental stages are governed by many genes of small effects, probably with complicated genetic interactions among them. Nevertheless, it should be noted also that heritability estimates are population and environment specific, hence, further assessments are needed to confirm the values reported by Duchêne *et al.*, (2012).

Cluster architecture is economically important; however, it has been addressed only in two QTL (Marguerit *et al.*, 2009; Correa *et al.*, 2014) and two candidate gene studies (Fernandez *et al.*, 2014; Tello *et al.*, 2015). The results points out that the parameters used to characterize this trait are moderately heritable, but there is a lack of consensus between the studies about the location of the identified QTLs. They suggest that these traits are controlled by complex dynamic genetic networks (Tello *et al.*, 2015). Another issue that may be involved in the low statistical power of GWAS for these traits is the lack of a precise method for the quantitative assessment of cluster architecture (Sepahi, 1980; Shavrukov *et al.*, 2004; Tello and Ibáñez, 2014). This limitation is shared with berry flesh firmness (Rolle *et al.*, 2011); although, a recent study has discovered several medium effect QTLs ($R^2 < 20\%$) for berry firmness in an F_1 progeny (Carreño *et al.*, 2014). One of them contained the *VvGAIL* candidate gene, which seem to be involved in the genetic control of grape texture parameters (Vargas *et al.*, 2013a).

Overall, the GWAS results obtained in this study for traits with highly complex genetic architecture are in agreement with what was predicted in the GWAS simulation conducted by Fodor *et al.*, (2014) in grapevine, as well as with empirical results from other species, such as

Arabidopsis (Atwell *et al.*, 2010). Moreover, the ability of GWAS to detect strong and stable associations for these traits was not significantly improved even when relatively larger mapping populations and higher SNP density were used in rice (Huang *et al.*, 2010; Zhao *et al.*, 2011; Huang *et al.*, 2012), foxtail millet (Jia *et al.*, 2013) or maize (Yang *et al.*, 2014).

5.6.1.3. New genetic associations

A total of 21 agronomic and phenotypic traits related mainly to table grape production were included in the present study. Some of them such as berry color, flower sex, muscat flavor and seed content were largely studied through QTL mapping (Dalbó *et al.*, 2000; Doligez *et al.*, 2002; Fischer *et al.*, 2004; Cabezas *et al.*, 2006; Doligez *et al.*, 2006; Lijavetzky *et al.*, 2006; Lowe and Walker, 2006; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Battilana *et al.*, 2009; Duchêne *et al.*, 2009; Marguerit *et al.*, 2009; Battilana *et al.*, 2013; Doligez *et al.*, 2013; Ban *et al.*, 2014), and the discovered major effect loci were usually followed-up through candidate gene testing (Kobayashi *et al.*, 2002; Kobayashi *et al.*, 2004; Azuma *et al.*, 2007; This *et al.*, 2007; Walker *et al.*, 2007; Fournier-Level *et al.*, 2009; Emanuelli *et al.*, 2010; Azuma *et al.*, 2011; Battilana *et al.*, 2011; Mejía *et al.*, 2011; Fechter *et al.*, 2012; Carrasco *et al.*, 2014; Costantini *et al.*, 2015). Other traits, such as phenological stages, fertility (also called fruitfulness), and cluster architecture, which are economically more or equally important, have received less attention (Fischer *et al.*, 2004; Fanizza *et al.*, 2005; Costantini *et al.*, 2008; Marguerit *et al.*, 2009; Doligez *et al.*, 2010; Duchêne *et al.*, 2012; Grzeskowiak *et al.*, 2013; Correa *et al.*, 2014; Fernandez *et al.*, 2014). This should be related with the complex quantitative nature of these traits, which lead in some cases to disappointing results, discouraging to continue or to undertake new studies. Our principal objective in this work was to test GWAS in grapevine using the QTLs detected in previous works to validate the results. But also to discover new genetic associations that would have been missed in QTL mapping studies, which are based on biparental populations with narrow genetic basis. In order to overcome this limitation, we used a mapping collection of broad genetic basis representing an important part of the cultivated grapevine gene pool. Our results demonstrated that GWAS was effective in identifying many of the previously reported QTLs for several traits (Table 8), but it allowed also detecting new genotype-phenotype associations that have not been reported before in the literature. The most consistent genetic associations discovered here should be the association with seed content related traits on chr17 and the associations with berry size related traits on chr5 and chr18 not linked to seedlessness QTLs.

Seed content in grape berries has been almost always studied in table grape using F₁ progenies involving seedless progenitors (Doligez *et al.*, 2002; Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Doligez *et al.*, 2013). These studies agreed that seedlessness is governed

by a major effect QTL located at the end of chr18 in addition to a couple of medium to minor effect QTLs distributed on other chromosomes, but to date none on chr17. The association detected on chr17 (Table A.12) involved a total of 15 SNPs ranging from 6,256,364 bp to 10,741,556 bp locations and associated with seed number (MNS) and/or seed dry weight (MSDW) for at least one year. However, it is difficult to identify the center of this association because the association peaks didn't overlap among all years and traits. If we consider the most repeatedly associated SNPs, then the center of the association would be around the two SNPs located at 9,294,776 bp and 9,739,079 bp (associated with MNS in 2012 and 2013 and with MSDW in 2012). However, if we consider the maximum association peaks, then the center would be around the four SNPs located between 6,256,381 bp and 6,256,518 bp (maximum associations with MNS in 2013 and with MSDW in 2012), which is also the center of an association with berry weight (MBW) and berry diameter (MBD). Moreover, the possibility of the presence of two different loci underlying seed content in this chromosome cannot be discarded.

Berry size studies have also been studied mainly on stenospermocarpic seedless backgrounds (Doligez *et al.*, 2002; Cabezas *et al.*, 2006; Mejía *et al.*, 2007; Costantini *et al.*, 2008; Doligez *et al.*, 2013), identifying a major effect QTL on linkage group 18, colocalizing with the major effect QTL for seedlessness, being *VvAGL11* (around the 32 Mb position) the most likely candidate gene for this locus (Mejía *et al.*, 2011). The major berry size QTL on chr18 has not been identified in the present study. Instead, a stable QTL for a berry size related trait (MBD: mean berry diameter) was found on this chromosome (around 5 Mb) but not colocalizing with seed content traits. Besides this, Fischer *et al.* (2004) reported two stable QTLs for berry size on linkage groups 5 and 13 in a complex hybrid cross including up to seven American *Vitis* species. In this study we have identified a genetic locus related with berry size on chr5 in a pure *vinifera* background (Table A.10). Two SNPs within this association peak (positioned at 14,046,799 bp and 14,327,183 bp) were simultaneously associated with flesh firmness, in agreement with Carreño *et al.*, (2014). This suggests that the phenotypic correlation between berry size and firmness ($Rho = 0.6$) could be, at least partly, to genetic factors. However, there is no biological evidence explaining why large grape berries are firmer than the small ones. On the other hand, Doligez *et al.* (2013) reported a stable and large effect QTL for berry size on chr17 that not colocalized with QTLs for seed traits by studying a cross between seeded grape cultivars, suggesting a potential use to select for seedlessness without reducing berry size. Our results confirmed the presence of a stable QTL for berry size on chr17, providing further arguments towards the involvement of this locus in the domestication process of grapevine (Myles *et al.*,

2011), but our results suggest that it may be overlapping with a QTL related with seed content (Table A.12).

Altogether, these findings confirm the superiority of LD-based QTL mapping over family-based QTL mapping when consider the amount of allelic diversity that can be discovered at a time: while the major QTLs for berry size and seedlessness on chr18 segregated only in seedless progenies and the berry size QTLs on chr5 and chr17 segregated separately in seeded backgrounds, in our mapping collection, all these QTLs were found simultaneously, in addition to the seed content QTL on chr17, which had never been reported before.

In addition to the stable associations discussed above, new putative associations were also identified with mean berry length on chr8 and chr5. The latter involved two SNPs positioned at 16,010,197 bp and 16,010,204 bp, near the stable association with berry weight (14,046,799 bp and 14,327,183 bp) (Table A.10). Given the high phenotypic correlation between the two traits ($Rho = 0.9$) that can be considered as different expressions of berry size, it should be expected that they are governed by the same genetic factors, as found in tomato (Lippman and Tanksley, 2001). However, in the lack of further arguments, it is not evident whether the two associations correspond to the same locus or two loci positioned next to each other.

Although fruit shape is an important characteristic in fruit breeding, it has received only a very scarce attention in grapevine (Wycislo *et al.*, 2008). So far, the present study is the first one that investigated the genetic determinism of this trait, discovering two putative associations on chr8 and chr11 that would need to be confirmed in further investigations.

Associations not previously reported were also detected with: muscat flavor, on chromosomes 4, 8, 7, 9, 15, 16, and 18; berry color, on chromosomes 3, 12, and 13; and flower sex, on all chromosomes but 3, 10, 11 and 14. Whereas the consistency of the associations with muscat flavor would need to be verified in further studies, the associations with berry color and flower sex are most likely false positives due to population structure or relatedness confounding because these two traits are known to be monogenic and governed by single loci located at chr2 (Dalbó *et al.*, 2000; Doligez *et al.*, 2002; Fischer *et al.*, 2004; Lowe and Walker, 2006; Riaz *et al.*, 2006; Marguerit *et al.*, 2009; Battilana *et al.*, 2013; Ban *et al.*, 2014). Especially for flower sex, false positives seem to be related with the unbalanced distribution of the trait.

5.6.2. SSR-based association analysis

Simulations as well as empirical studies carried out in several organisms suggested that SSR markers possess a high power in detecting genetic associations (Slatkin, 1994; Chapman and Wijisman, 1998; Xiong and Jin, 1999; Abecasis *et al.*, 2001; Remington *et al.*, 2001; Ohashi and Tokunaga, 2003; Varilo *et al.*, 2003; Payseur *et al.*, 2008). In cultivated grapevine, Barnaud *et al.*

(2006) found that, LD assessed using SSR markers extends to larger distances, suggesting that genome-wide scan would be effective using a small number of SSR markers. In order to test the effectiveness of this approach, we used the panel of 45 multiplex PCRs to genotype a set of 264 SSRs distributed throughout the genome against the mapping collection of 274 grapevine accessions. The obtained data were used to perform association analysis for 21 phenotypic and agronomic traits using the Q+K mixed linear model.

Due to the high number of statistical tests carried out in genome wide scans, the selection of an optimal significance threshold that deal effectively with the multiple testing problem without compromising the results through high stringency is an issue of primary importance. We found that Bonferroni correction (threshold $P < 1.90\text{E-}04$) was highly conservative, therefore we relaxed the significance threshold to the nominal level of $P < 0.01$. Our rational in this is that, relaxing the significance threshold until SSR markers known a priori to be linked to major effect QTLs became significantly associated to the respective traits.

The number of significantly associated SSRs at $P < 0.01$ was inflated for some traits, particularly the simple ones that are known to be controlled by major loci, such as flower sex (33 SSRs from 10 chromosomes) (Table 9 and Table A.14). This is in accordance with the SNP-based association analyses, and confirms the hypothesis that the increased number of associations for this trait when compared with berry color is the result of their highly imbalanced distribution (Fig. 5). Relative excess of associations after accounting for Q and K was also the case for some quantitative traits such as flowering (22 SSRs from 10 chromosomes) or berry length (15 SSRs from 10 chromosomes) (Table 9 and Table A.14). While many of the significant associations with flower sex are definitely spurious, since this trait is known to be controlled by a major locus in chr2, this would not be the situation for the quantitative traits. Part of the detected associations would be false positives introduced because of the use of a relaxed significance threshold (0.01), since such a large number of associations has not been found in the linkage mapping studies developed for these traits (for flowering time see for example Costantini *et al.*, 2008; Duchêne *et al.*, 2012; Grzeskowiak *et al.*, 2013). But this approach suffers from the limited number of QTLs that can be detected at a time; only QTLs that are segregating in the mapping population can be detected (Korte and Farlow, 2013; Huang and Han, 2014).

The SSR based approach developed in this study identified several associations that colocalized with previously reported QTLs and/or with associations detected using the SNP dataset in the present work. These include: the association with seed content on chr17; with berry weight and berry shape on chr5 and chr8, respectively, also detected with the SNP dataset; with fertility on chr5, agreeing with Doligez *et al.* (2010) and with the SNP-based association profile observed with GLM; and the association with flowering time on chr7, in the same region as a QTL

identified by Duchêne *et al.*, (2012). Additionally, the SSR-based association analysis identified for the first time new stable associations with: budbreak time on chr5 (FAM122); flowering time on chr5 (VVIT68 and VChr5b), chr13 (FAM62-2), chr16 (VVIN52), chr17 (VVIN73) and chr19 (UDV-062-2); veraison time on chr1 (VVIO02); fertility on chr17 (VRZAG15); cluster weight on chr18 (VMC2B1.1); cluster compactness on chr9 (VVIQ66); berry length and diameter on chr7 (FAM116) and berry diameter on chr8 (UDV-075). However, the reliability of the new associations detected exclusively using the microsatellites may be questionable because of the relaxed significance level as mentioned above, and need to be further studied to be confirmed.

To our knowledge, the most successful SSR-based GWAS were those conducted in humans by the research group of Inoko H. (Tamiya *et al.*, 2005; Bahram and Inoko, 2007) using 20,000 to 30,000 SSRs (approx. 1 SSR each 100 kb) and the DNA pooling strategy. Moreover, the most promising associations discovered by the SSRs have been always locally refined later using SNP markers (Kawashima *et al.*, 2006; Hui *et al.*, 2008; Nakabayashi *et al.*, 2009; Michikawa *et al.*, 2010; Shibata *et al.*, 2013; Saka *et al.*, 2015). However, the principal of this approach (designed specifically for the case-control designs) and its technical and cost requirements may limit its extension to crop plants. SSR-based association mapping studies have been carried out in a couple of crop species such as wheat (Breseghello and Sorrells, 2006a; Peng *et al.*, 2009; Liu *et al.*, 2010), rice (Agrama *et al.*, 2007), sorghum (El Mannai *et al.*, 2011), cotton (Zhao *et al.*, 2014) and watermelon (Reddy *et al.*, 2014). However, all these studies were based on a modest population size and arbitrary marker density (maximum of 212 SSRs and 212 accessions in (Zhao *et al.*, 2014)). In addition, population structure confounding has not usually controlled (Q+K model was used only in two studies). The number of significant associations per trait (based on different threshold definitions) varied from one to a maximum of 42 for *verticillium* resistance in cotton (Zhao *et al.*, 2014). But none of them were confirmed in further fine mapping studies.

The SSR-based association analysis presented here is the most comprehensive regarding the number of individuals and markers used, and it is the first of its kind conducted so far in grapevine and in any perennial fruit crop. As demonstrated above this analysis allowed to identify reliable associations overlapping with previous discoveries and others that need further investigations in order to be confirmed.

5.6.3. Comparison of high density SNPs vs. low density SSRs for GWAS

Currently, SNPs and SSRs are the molecular markers most commonly used for genetic studies. Several research works have been dedicated to the comparison of informativeness and capabilities of SSRs and SNPs for population genetics-related analysis such as genetic diversity,

relatedness and population structure (Rosenberg *et al.*, 2003; Liu *et al.*, 2005; Hamblin *et al.*, 2007; Jones *et al.*, 2007; Yu *et al.*, 2009; Yang *et al.*, 2011; Simko *et al.*, 2012; Gärke *et al.*, 2012; Singh *et al.*, 2013; Würschum *et al.*, 2013; Mesak *et al.*, 2014; this list was not intended to be exhaustive); but only few of them were focused on the comparison of the performances of SNP and SSR markers for LD-based mapping. In addition to the scarcity of such studies, they are usually based on simulated data and theoretical assumptions (Chapman and Wijtsman, 1998; Xiong and Jin, 1999; Ohashi and Tokunaga, 2003), that lack consistency with the complex evolutionary history of living organisms. Our study provided an opportunity to perform directly such a comparison using empirical data. The few studies published to date about the extent of LD in grapevine reported that SNP-based LD decayed very rapidly (Lijavetzky *et al.*, 2007; This *et al.*, 2007; Myles *et al.*, 2010; Myles *et al.*, 2011; Vargas *et al.*, 2013a; Vargas *et al.*, 2013b) whereas it extended further when using SSR markers (Barnaud *et al.*, 2006). These results suggested that a sparse set of SSR markers distributed homogeneously over the grapevine genome would be sufficient to perform GWAS with reasonable statistical power, equivalent to a moderate to high density SNP-based GWAS (Barnaud *et al.*, 2006). In this study, we have performed a GWAS for 21 table grape related traits using: a set of 36,332 SNPs (with MAF of 0.01 and mean missingness of 12%) over a diversity panel ranging from 207 (in 2010 and 2012) to 242 (in 2013) individuals; and a set of 264 SSRs over the same mapping collection, with a small difference in size and coverage in 2013 (274 individuals, but 67 of them genotyped only for 109 SSRs). The direct comparison between SSR- and SNP-based association results was not straightforward given the large differences in marker density and significance thresholds considered for each of them. Since our SNP coverage is several orders of magnitude higher than SSR coverage (practically, all the SSR loci are covered by the SNP dataset), the expectation was that any association detected by the SSRs should be detected also by SNPs, and not *vice versa*. In order to reveal similarities and differences between the two types of markers, we prospected the SNP association profiles in a radius of approximately 1Mb around the position corresponding to each associated SSR with a given trait, and we reported the SNP with the maximum *P* value found therein (Table 9 and Table A.14). The results have been compared at two roughly equivalent significance thresholds (associations to berry color and flower sex out of chr2 are not considered): 1) the genome-wide Bonferroni correction ($1.90\text{E-}04$ for SSRs and $1.40\text{E-}06$ for SNPs); 2) a relaxed significance threshold 53 times lower than the Bonferroni correction for both marker types (0.01 for the SSRs and $7.30\text{E-}05$ for SNPs; Table 9 and Table A.14).

i. At the genome-wide Bonferroni correction

At this threshold only skin color and flower sex loci on chr2 were detected simultaneously by SNPs and SSRs. However, an association to muscat flavor on chr19

(UDV-046), and unstable associations to flowering (VVIT68 on chr5), veraison (VVIO02 on chr2) and cluster length (VV6718 and UDV-046 on chr19) were detected with SSRs but not with SNPs. These associations are new since QTLs for the corresponding traits have never been declared in these regions. This discrepancy cannot be explained by the lack of coverage in the SNP data set because SNPs existed at the same positions but not significantly associated with those traits at the genome-wide Bonferroni corrected threshold. On the other side, stable and putative associations to several traits detected by the SNP set (Table 8) were not detected by the SSRs, including the major loci controlling muscat flavor and seedlessness on chr5 and chr18 respectively, although SSR markers known to be associated with those QTLs (VVMD27 and VRZAG79 for the former and VMC7F2 and VVAGL11 for the latter) were included in the analysis.

ii. At the relaxed significance threshold

At this threshold the number of overlapped associations between SNPs and SSRs increased, including associations that coincide with previously known QTLs (muscat flavor associations on chr2 and 5, seedlessness on chr1 and 18, ripening on chr17 and fertility, MCW and MBW on chr5); but also associations detected for the first time in our study, such as berry size (chr18, around 4 Mb location), berry shape index (chr8), and seed content (chr17). On the other side, some genetic associations overlapping with previous QTLs were detected by SSRs but not by their SNP counterpart. Those included: muscat flavor on chr13 (VMC8E6), seed content on chr4 (VVIN75-2), flowering on chr7 (VVIQ67), veraison on chr2 (VMC2C10-1), ripening on chr17 (VVIN73), cluster weight on chr16 (VChr16b), cluster length on chr2 (VMC2C10-1) and berry weight on chr11 (RTAChr11-11b). However, with the exception of flowering time on chr7, which coincide with the large effect QTL described by (Duchêne *et al.*, 2012) for this trait, all the rest were unstable associations that colocalize with minor to medium effect QTLs. Additionally, new associations (that have never been declared in the literature) were detected repeatedly in at least two years by the SSRs but not by the SNPs, such as budbreak and flowering on chr5 (FAM122 and VVIT68 respectively), veraison on chr1 (VVIO02), compactness index on chr9 (VVIQ66) and MBD on chr8 (UDV-075).

This comparison indicates that the statistical power of the SNP-based GWAS developed is higher than that of the SSRs-based, because only a few SSR-based associations reached the genome-wide multiple testing-corrected threshold. This is consistent with the low LD levels that we found between linked SSR pairs (see section 5.5.2.), suggesting that much more SSRs would need to be genotyped in order to achieve the same power as the 36,332 SNP dataset used here. In addition to low density, the problem of null alleles may also be implicated in decreasing the

power of SSRs. The SSR marker VMC7F2 for example has been found tightly linked with seedlessness in all the QTL mapping studies that used it. By contrast, in the current study this was not the case: the smallest P value obtained with this marker was 0.51 for all the seed content trait-year combinations (2 traits x 3 years). This marker indeed presented null alleles, and even if we used the redesigned version (VMC7F2-2) that gave consistent allele inheritance in 14 trio pedigrees, this does not guarantee that it is null allele-free in the rest of accessions.

Despite this difference in detection power, as reported above, some associations were detected with SSRs and not with SNPs, particularly when the less stringent significance threshold was considered. One of the most plausible factors that can explain this discrepancy is the non-hierarchical structure of LD (Carlson *et al.*, 2004). This phenomenon is better appreciated when we look closer to the association profiles of both SNPs and SSRs. Fig. 29 depicts the association profiles of SNPs and SSRs around some of the most extensively studied candidate genes in grapevine: the tandemly positioned *Myb* genes responsible for color variation (Kobayashi *et al.*, 2002; Kobayashi *et al.*, 2004; Lijavetzky *et al.*, 2006; This *et al.*, 2007; Walker *et al.*, 2007; Fournier-Level *et al.*, 2009; Azuma *et al.*, 2011; Carrasco *et al.*, 2014; Costantini *et al.*, 2015); and *VvDXS*, suggested as a candidate gene for muscat flavor expression (Battilana *et al.*, 2009; Duchêne *et al.*, 2009; Emanuelli *et al.*, 2010; Battilana *et al.*, 2011; Emanuelli *et al.*, 2014). This figure shows that markers situated relatively far from the genes are more significantly associated with the trait than closer ones and even more than markers located within the genes (the case of three SNPs within the sequence of *VvDXS* in Fig. 29c). This is not related to differences in diversity parameters between the associated and non-associated markers (missingness, MAF, heterozygosity and Hardy-Weinberg equilibrium; data not shown), indicating that other uncontrolled factors may be behind this behavior. In fact, these association profiles reflect the complex pattern of LD distribution along the genome that we observed with both SNPs and SSRs (see sections 5.5.1. and 5.5.2.), which is far from being a monotonic function of physical distance. This is because LD is the result of a complex interplay of several factors including mutation, recombination, selection, bottleneck effect, genetic drift, mating system, population structure, migration and admixture along the evolutionary history of the species (Ardlie *et al.*, 2002). By this process, it is not surprising thus to find two markers at the same distance from a gene, one in strong LD with it, and the other one not associated with it, regardless of marker type (SSR or SNP). All together, these findings indicate that marker density should be increased to a maximum in order to insure that some markers at the vicinity or within the gene are in strong LD with the functional variant; but it indicates also that much care should be taken at the time of interpreting association results by tight linkage between a marker and a functional variant.

Figure 29. Association profiles of SNPs (a, c) and SSRs (b, d) around the candidate genes responsible for berry color and muscat flavor

VvmybA1 and *VvmybA2* in chr2 are the candidate genes underlying berry color, and *VvDXS* in chr5 is the candidate gene underlying muscat flavor. Red bars represent the start and end points of each gene. For the SNPs, we presented approximately 2 Mb from both sides of the genes; For the SSRs we presented the whole chromosome.

CONCLUSIONS

6. CONCLUSIONS

1. The GBS protocol adopted in this work followed by data filtering has provided a reliable dataset with marker density higher than what is currently offered by the available array-based techniques in grapevine.
2. The most important limitation of the GBS strategy for association analyses is the large proportion of missing genotypes. However, with the accelerating development and prices reduction of sequencing technologies (allowing more sequencing depth) and the use of haplotypic information and imputation strategies, this limitation is likely to be overcome in the near future.
3. The developed panel of microsatellite multiplex PCRs allows the study of 264 genome-wide distributed microsatellite loci using only 45 PCRs and sequencing runs. This represents a significant increase in information gain and time and cost savings when compared to the currently available multiplexes.
4. The genotyping of the mapping collection using the panel of microsatellite multiplex PCRs generated a vast amount of reliable information related to allele size ranges, diversity parameters and presence of null alleles, as well as features of amplification profiles and ease of scoring of the studied microsatellite markers. This information can be used to identify the most suitable loci for future microsatellite based genetic studies in grapevine; but also to design *ad-hoc* multiplexes, combining the most informative loci for specific studies in only a few multiplex PCRs.
5. The presence of null alleles is common in grapevine microsatellite markers and should be considered in any genetic diversity or linkage disequilibrium based studies.
6. Large chromosomal regions for which SSR markers have never been developed still exist in the grapevine genome, and it seems that this is related to a high density of repetitive/transposable elements in these regions.
7. Linkage disequilibrium analysis pointed out that estimates based on small sample sizes (roughly less than 30 individuals) or generalizations based on only few chromosomes or genomic regions may be seriously biased. Additionally, MAF distribution in the dataset should be considered at the time of LD assessment.
8. SNP-based LD for medium to high power GWAS is unlikely to extend beyond an average distance of 1 kb, therefore, the number of required SNPs for such studies will be of the order of millions rather than thousands.
9. The set of SSR markers used in the present study did not allow to obtain reliable estimates of LD decay throughout the genome because of insufficient marker density. However, it does

indicate that SSR-based LD would not extend much more than SNP-based LD; therefore thousands of SSR markers would be needed to perform high powered GWAS.

10. For grapevine association analysis, accounting for pairwise relatedness (kinship matrix, K) in the mixed model is more effective in reducing spurious associations than accounting for population structure alone; and further improvements can be achieved by accounting for both of them simultaneously.

11. GWAS was effective in discovering genetic loci underlying simply to moderately inherited traits such as color, sex, muscat flavor, seed content and berry size. However, the statistical power was limited for traits with more complex genetic architecture. The study of these traits would need to increase the size of mapping collections and SNP density and to use more sophisticated experimental designs, such as replicated trials.

12. The high resolution SNP-based GWAS have provided new insights into the genetic determinism of some traits. The most relevant ones are: 1) chr17 harbors a strong association implicated in the genetic determinism of berry size and seed content. 2) the major loci controlling berry size seem to be located on chr5 and chr17 rather than chr18 as found previously in seedless F₁ populations; being the former (chr5) independent from seed content traits and colocalizing instead with a putative flesh firmness association. Such independent association could be useful for table grape breeding through marker assisted selection.

13. Low resolution GWAS based on SSR markers provided only a weak power of detection, as predicted through LD analyses. Nevertheless, it allowed to detect several previously discovered QTLs and SNP associations found in the present study, and provided some evidence of new associations.

REFERENCES

7. REFERENCES

- Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, et al (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* **68**: 191–197
- Agrama HA, Eizenga GC, Yan W (2007) Association mapping of yield and its components in rice cultivars. *Mol Breed* **19**: 341–356
- Alleweldt G, Possingham JV (1988) Progress in grapevine breeding. *Theor Appl Genet* **75**: 669–673
- Aradhya MK, Dangl GS, Prins BH, Boursiquot J-M, Walker MA, Meredith CP, Simon CJ (2003) Genetic structure and differentiation in cultivated grape, *Vitis vinifera* L. *Genet Res* **81**: 179–192
- Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**: 299–309
- Arroyo-García R, Martínez-Zapater JM (2004) Development and characterization of new microsatellite markers for grape. *Vitis* **43**: 175–178
- Arroyo-García R, Ruiz-García L, Bolling L, Ocete R, López M a, Arnold C, Ergul A, Söylemezoğlu G, Uzun HI, Cabello F, et al (2006) Multiple origins of cultivated grapevine (*Vitis vinifera* L. ssp. *sativa*) based on chloroplast DNA polymorphisms. *Mol Ecol* **15**: 3707–3714
- Arumugasundaram S, Ghosh M, Veerasamy S, Ramasamy Y (2011) Species discrimination, population structure and linkage disequilibrium in *Eucalyptus camaldulensis* and *Eucalyptus tereticornis* using SSR markers. *PLoS One* **6**: e28252
- Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**: 627–631
- Azuma A, Kobayashi S, Yakushui H, Yamada M, Mitani N, Sato A (2007) *VvmybA1* genotype determines grape skin color. *Vitis* **46**: 154–155
- Azuma A, Udo Y, Sato A, Mitani N, Kono A, Ban Y, Yakushiji H, Koshita Y, Kobayashi S (2011) Haplotype composition at the color locus is a major genetic determinant of skin color variation in *Vitis × labruscana* grapes. *Theor Appl Genet* **122**: 1427–1438
- 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

- Bahram S, Inoko H** (2007) Microsatellite markers for genome-wide association studies. *Nat Rev Genet*. doi: 10.1038/nrg1962-c1
- Balding DJ** (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**: 781–791
- Ban Y, Mitani N, Hayashi T, Sato A, Azuma A, Kono A, Kobayashi S** (2014) Exploring quantitative trait loci for anthocyanin content in interspecific hybrid grape (*Vitis labruscana* × *Vitis vinifera*). *Euphytica* **198**: 101–114
- Barba P, Cadle-Davidson L, Harriman J, Glaubitz JC, Brooks S, Hyma K, Reisch B** (2014) Grapevine powdery mildew resistance and susceptibility loci identified on a high-resolution SNP map. *Theor Appl Genet* **127**: 73–84
- Barnaud A, Lacombe T, Doligez A** (2006) Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L. *Theor Appl Genet* **112**: 708–716
- Barnaud A, Laucou V, This P, Lacombe T, Doligez A** (2010) Linkage disequilibrium in wild French grapevine, *Vitis vinifera* L. subsp. *silvestris*. *Heredity* **104**: 431–437
- Barrett JC, Cardon LR** (2006) Evaluating coverage of genome-wide association studies. *Nat Genet* **38**: 659–662
- Barrett JC, Fry B, Maller J, Daly MJ** (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**: 263–265
- Battilana J, Costantini L, Emanuelli F, Sevini F, Segala C, Moser S, Velasco R, Versini G, Grando MS** (2009) The 1-deoxy-D-xylulose 5-phosphate synthase gene co-localizes with a major QTL affecting monoterpene content in grapevine. *Theor Appl Genet* **118**: 653–669
- Battilana J, Emanuelli F, Gambino G, Gribaudo I, Gasperi F, Boss PK, Grando MS** (2011) Functional effect of grapevine 1-deoxy-D-xylulose 5-phosphate synthase substitution K284N on Muscat flavour formation. *J Exp Bot* **62**: 5497–5508
- Battilana J, Lorenzi S, Moreira FM, Moreno-Sanz P, Failla O, Emanuelli F, Grando MS** (2013) 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* **54**: 1031–1037
- Bennett J, Jarvis P, Creasy GL, Trought MCT** (2005) Influence of defoliation on overwintering carbohydrate reserves, return bloom, and yield of mature chardonnay grapevines. *Am J Enol Vitic* **56**: 386–393
- Bergamini C, Cardone MF, Anaclerio A, Perniola R, Pichierri A, Genghi R, Alba V, Forleo LR, Caputo AR, Montemurro C, et al** (2013) Validation assay of p3_VvAGL11 marker in a wide range of genetic background for early selection of stenospermocarpy in *Vitis vinifera* L. *Mol Biotechnol* **54**: 1021–1030

- Blacket MJ, Robin C, Good RT, Lee SF, Miller AD** (2012) Universal primers for fluorescent labelling of PCR fragments--an efficient and cost-effective approach to genotyping by fluorescence. *Mol Ecol Resour* **12**: 456–463
- Bouquet A** (2011) Grapevines and viticulture. *In* AF Adam-Blondon, JM Martínez-Zapater, C Kole, eds, Genetics, genomics, and breeding of grapes. CRC Press, pp 1–29
- Bowers JE, Dangl GS, Meredith CP** (1999) Development and characterization of additional microsatellite DNA markers for grape. *Am J Enol Vitic* **50**: 243–246
- Bowers JE, Dangl GS, Vignani R, Meredith CP** (1996) Isolation and characterization of new polymorphic simple sequence repeat loci in grape (*Vitis vinifera* L.). *Genome* **39**: 628–633
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES** (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635
- Breseghello F, Sorrells ME** (2006a) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* **172**: 1165–1177
- Breseghello F, Sorrells ME** (2006b) Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Sci* **46**: 1323–1330
- Bryda EC, Riley LK** (2008) Multiplex microsatellite marker panels for genetic monitoring of common rat strains. *J Am Assoc Lab Anim Sci* **47**: 37–41
- Butler JM** (2005a) Forensic DNA typing protocols. *Methods Mol Biol* **297**: 53–65
- Butler JM** (2005b) Constructing STR multiplex assays. *Methods Mol Biol* **297**: 53–66
- Cabezas JA, Cervera Goy MT, Ruiz-García L, Carreño J, Martínez-Zapater JM, Cervera M, Ruiz-Garcia L** (2006) A genetic analysis of seed and berry weight in grapevine. *Genome* **49**: 1572–1585
- Callen DF, Thompson a D, Shen Y, Phillips H a, Richards RI, Mulley JC, Sutherland GR** (1993) Incidence and origin of “null” alleles in the (AC)_n microsatellite markers. *Am J Hum Genet* **52**: 922–927
- Canaguier A, Scalabrin S, Le Paslier M-C, Duchene E, Mohellibi N, Bérard A, Chauveau A, Boursiquot J-M, Di Gaspero G, Hausmann L, et al** (2014) Improvement of the grapevine genome assembly. *Plant and Animal Genome Conference (PAGXXII)*, San Diego, CA
- Carlson CS, Eberle MA, Kruglyak L, Nickerson DA** (2004) Mapping complex disease loci in whole-genome association studies. *Nature* **429**: 446–452
- Del Carpio DP, Basnet RK, De Vos RCH, Maliepaard C, Paulo MJ, Bonnema G** (2011) Comparative methods for association studies: a case study on metabolite variation in a *Brassica rapa* core collection. *PLoS One* **6**: e19624

- Carrasco D, De Lorenzis G, Maghradze D, Revilla E, Bellido A, Failla O, Arroyo-García R** (2014) Allelic variation in the *VvMYBA1* and *VvMYBA2* domestication genes in natural grapevine populations (*Vitis vinifera* subsp. *sylvestris*). *Plant Syst Evol* **301**: 1613–1624
- Carreño I, Cabezas JA, Martínez-Mora C, Arroyo-García R, Cenis JL, Martínez-Zapater JM, Carreño J, Ruiz-García L** (2014) Quantitative genetic analysis of berry firmness in table grape (*Vitis vinifera* L.). *Tree Genet Genomes* **11**: 818–828
- Chapman NH, Wijsman EM** (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am J Hum Genet* **63**: 1872–1885
- Cipriani G, Di Gaspero G, Canaguier A, Jusseaume J, Tassin J, Lemainque A, Thareau V, Adam-Blondon A-F, Testolin R** (2011) Molecular linkage maps: Strategies, resources, and achievements. *In* A-F Adam-Blondon, J-M Martínez-Zapater, C Kole, eds, *Genetics, Genomics, and Breeding of Grapes*. CRC Press, pp 68–89
- Cipriani G, Marrazzo MT, Di Gaspero G, Pfeiffer A, Morgante M, Testolin R** (2008) A set of microsatellite markers with long core repeat optimized for grape (*Vitis* spp.) genotyping. *BMC Plant Biol* **8**: 127
- Cipriani G, Spadotto A, Jurman I, Di Gaspero G, Crespan M, Meneghetti S, Frare E, Vignani R, Cresti M, Morgante M, et al** (2010) The SSR-based molecular profile of 1005 grapevine (*Vitis vinifera* L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. *Theor Appl Genet* **121**: 1569–1585
- Comadran J, Ramsay L, MacKenzie K, Hayes P, Close TJ, Muehlbauer G, Stein N, Waugh R** (2011) Patterns of polymorphism and linkage disequilibrium in cultivated barley. *Theor Appl Genet* **122**: 523–531
- Correa J, Mamani M, Muñoz-Espinoza C, Laborie D, Muñoz C, Pinto M, Hinrichsen P** (2014) Heritability and identification of QTLs and underlying candidate genes associated with the architecture of the grapevine cluster (*Vitis vinifera* L.). *Theor Appl Genet* **127**: 1143–1162
- Costantini L, Battilana J, Lamaj F, Fanizza G, Grando MS** (2008) Berry and phenology-related traits in grapevine (*Vitis vinifera* L.): from quantitative trait loci to underlying genes. *BMC Plant Biol* **8**: 38
- Costantini L, Malacarne G, Lorenzi S, Troggio M, Mattivi F, Moser C, Grando MS** (2015) New candidate genes for the fine regulation of the colour of grapes. *J Exp Bot*. doi: 10.1093/jxb/erv159

- Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, et al** (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3* **3**: 1903–1926
- Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon A-F, Boursiquot J-M, This P** (2008) Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. *sativa*. *BMC Plant Biol* **12**: 31
- Dakin EE, Avise JC** (2004) Microsatellite null alleles in parentage analysis. *Heredity* **93**: 504–509
- Dalbó M a, Ye GN, Weeden NF, Steinkellner H, Sefc KM, Reisch BI** (2000) A gene controlling sex in grapevines placed on a molecular marker-based genetic map. *Genome* **43**: 333–340
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al** (2011) The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158
- Da Silva C, Zamperin G, Ferrarini A, Minio A, Dal Molin A, Venturini L, Buson G, Tononi P, Avanzato C, Zago E, et al** (2013) The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome. *Plant Cell* **25**: 4777–4788
- Davey JW, Hohenlohe P a, Etter PD, Boone JQ, Catchen JM, Blaxter ML** (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**: 499–510
- De Arruda MP, Gonçalves EC, Schneider MPC, da Silva AL da C, Morielle-Versute E** (2010) An alternative genotyping method using dye-labeled universal primer to reduce unspecific amplifications. *Mol Biol Rep* **37**: 2031–2036
- Decroocq V, Favé MG, Hagen L, Bordenave L, Decroocq S** (2003) Development and transferability of apricot and grape EST microsatellite markers across taxa. *Theor Appl Genet* **106**: 912–922
- Devlin B, Roeder K** (1999) Genomic control for association studies. *Biometrics* **55**: 997–1004
- Díaz A, Zarouri B, Fergany M, Eduardo I, Alvarez JM, Picó B, Monforte AJ** (2014) Mapping and introgression of qtl involved in fruit shape transgressive segregation into “Piel de Sapo” melon (*Cucumis melo* L.). *PLoS One* **9**: e104188
- Di Gaspero G, Cipriani G, Marrazzo MT, Andreetta D, Castro MJP, Peterlunger E, Testolin R** (2005) Isolation of (AC)_n-microsatellites in *Vitis vinifera* L. and analysis of genetic background in grapevines under marker assisted selection. *Mol Breed* **15**: 11–20

- Di Genova A, Almeida AM, Muñoz-Espinoza C, Vizoso P, Trivisani D, Moraga C, Pinto M, Hinrichsen P, Orellana A, Maass A** (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol* **14**: 7
- Doerge RW** (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* **3**: 43–52
- Doligez A., Audiot E, Baumes R, This P** (2006) QTLs for muscat flavor and monoterpenic odorant content in grapevine (*Vitis vinifera* L.). *Mol Breed* **18**: 109–125
- Doligez A., Bouquet a., Danglot Y, Lahogue F, Riaz S, Meredith P, Edwards J, This P** (2002) Genetic mapping of grapevine (*Vitis vinifera* L.) applied to the detection of QTLs for seedlessness and berry weight. *Theor Appl Genet* **105**: 780–795
- Doligez A, Bertrand Y, Dias S, Grolier M, Ballester J-F, Bouquet A, This P** (2010) QTLs for fertility in table grape (*Vitis vinifera* L.). *Tree Genet Genomes* **6**: 413–422
- Doligez A, Bertrand Y, Farnos M, Grolier M, Romieu C, Esnault F, Dias S, Berger G, François P, Pons T, et al** (2013) New stable QTLs for berry weight do not colocalize with QTLs for seed traits in cultivated grapevine (*Vitis vinifera* L.). *BMC Plant Biol* **13**: 217
- Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS** (1991) “Touchdown” PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res* **19**: 4008
- De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG** (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. *PLoS One* **8**: e62137
- Drašnarová A, Krak K, Vít P, Doudová J, Douda J, Hadincová V, Zákavský P, Mandák B** (2014) Cross-amplification and multiplexing of SSR markers for *Alnus glutinosa* and *A. incana*. *Tree Genet Genomes* **10**: 865–873
- Duchêne E, Butterlin G, Claudel P, Dumas V, Jaegli N, Merdinoglu D** (2009) A grapevine (*Vitis vinifera* L.) deoxy-d-xylulose synthase gene colocates with a major quantitative trait loci for terpenol content. *Theor Appl Genet* **118**: 541–552
- Duchêne E, Butterlin G, Dumas V, Merdinoglu D** (2012) Towards the adaptation of grapevine varieties to climate change: QTLs and candidate genes for developmental stages. *Theor Appl Genet* **124**: 623–635
- Eibach R, Töpfer R** (2015) Traditional grapevine breeding techniques. In AG Reynolds, ed, *Grapevine breeding programs for the wine industry: traditional and molecular techniques*. Elsevier Science, pp 3–22
- Eibach R, Zyprian E, Welter L, Töpfer R** (2007) The use of molecular markers for pyramiding resistance genes in grapevine breeding. *Vitis* **46**: 120–125

- Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE** (2000) Multiplex PCR: Optimization and application in diagnostic virology. *Clin Microbiol Rev* **13**: 559–570
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE** (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**: e19379
- Emanuelli F, Battilana J, Costantini L, Le Cunff L, Boursiquot J-M, This P, Grando MS** (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 JM, Zyprian E, Moreira FM, et al** (2013) Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biol* **13**: 39
- Emanuelli F, Sordo M, Lorenzi S, Battilana J, Grando MS** (2014) Development of user-friendly functional molecular markers for *VvDXS* gene conferring muscat flavor in grapevine. *Mol Breed* **33**: 235–241
- Ersoz ES, Yu J, Buckler ES** (2007) Applications of linkage disequilibrium and association mapping in crop plants. In RK Varshney, R Tuberosa, eds, *Genomics-assisted crop improvement*. Springer Netherlands, Dordrecht, pp 97–119
- 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
- Falconer DS, Mackay TFC** (1996) *Introduction to quantitative genetics*. Longman
- Fanizza G, Lamaj F, Costantini L, Chaabane R, Grando MS** (2005) QTL analysis for fruit yield components in table grapes (*Vitis vinifera*). *Theor Appl Genet* **111**: 658–664
- Fechter I, Hausmann L, Daum M, Sørensen TR, Viehöver P, Weisshaar B, Töpfer R** (2012) Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. *Mol Genet Genomics* **287**: 247–259
- Fernandez L, Le Cunff L, Tello J, Lacombe T, Boursiquot JM, Fournier-Level A, Bravo G, Lalet S, Torregrosa L, This P, et al** (2014) Haplotype diversity of *VvTFL1A* gene and association with cluster traits in grapevine (*V. vinifera*). *BMC Plant Biol* **14**: 209
- Fischer BM, Salakhutdinov I, Akkurt M, Eibach R, Edwards KJ, Töpfer R, Zyprian EM** (2004) Quantitative trait locus analysis of fungal disease resistance factors on a molecular map of grapevine. *Theor Appl Genet* **108**: 501–515
- Flint-Garcia S a, Thornsberry JM, Buckler ES** (2003) Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54**: 357–374

- Fodor A, Segura V, Denis M, Neuenschwander S, Fournier-Level A, Chatelet P, Homa FAA, Lacombe T, This P, Le Cunff L** (2014) Genome-wide prediction methods in highly diverse and heterozygous species: proof-of-concept through simulation in grapevine. *PLoS One* **9**: e110436
- Fournier-Level A, Le Cunff L, Gomez C, Doligez A, Ageorges A, Roux C, Bertrand Y, Souquet J-M, Cheynier V, This P** (2009) Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. ssp. *sativa*) berry: a quantitative trait locus to quantitative trait nucleotide integrated study. *Genetics* **183**: 1127–1139
- Galet P** (1998) Précis d ampé ographie pratique, 7th ed. Pierre Galet, Montpellier
- Gärke C, Ytournal F, Bed’hom B, Gut I, Lathrop M, Weigend S, Simianer H** (2012) Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Anim Genet* **43**: 419–428
- Gaut BS, Long AD** (2003) The lowdown on linkage disequilibrium. *Plant Cell* **15**: 1502–1506
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES** (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**: e90346
- Goldstein DB, Schlötterer C** (1999) Microsatellites: Evolution and Applications. Oxford University Press
- Grassi F, Labra M, Imazio S, Rubio RO, Failla O, Scienza a., Sala F** (2006) Phylogeographical structure and conservation genetics of wild grapevine. *Conserv Genet* **7**: 837–845
- Grassi F, Labra M, Imazio S, Spada A, Sgorbati S, Scienza A, Sala F** (2003) Evidence of a secondary grapevine domestication centre detected by SSR analysis. *Theor Appl Genet* **107**: 1315–1320
- Grattapaglia D, Bertolucci FL, Sederoff RR** (1995) Genetic mapping of QTLs controlling vegetative propagation in *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross strategy and RAPD markers. *Theor Appl Genet* **90**: 933–947
- Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD** (2011) Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* **108**: 11983–11988
- Grimplet J, Van Hemert J, Carbonell-Bejerano P, Díaz-Riquelme J, Dickerson J, Fennell A, Pezzotti M, Martínez-Zapater JM** (2012) Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences. *BMC Res Notes* **5**: 213

- Grzeskowiak L, Costantini L, Lorenzi S, Grando MS** (2013) Candidate loci for phenology and fruitfulness contributing to the phenotypic variability observed in grapevine. *Theor Appl Genet* **126**: 2763–2776
- Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, et al** (2011) Current trends in microsatellite genotyping. *Mol Ecol Resour* **11**: 591–611
- Gulcher J** (2012) Microsatellite markers for linkage and association studies. *Cold Spring Harb Protoc* **2012**: 425–432
- Gupta PK, Rustgi S, Kulwal PL** (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol* **57**: 461–485
- Hall JM, LeDuc CA, Watson AR, Roter AH** (1996) An approach to high-throughput genotyping. *Genome Res* **6**: 781–790
- Hamblin MT, Warburton ML, Buckler ES** (2007) Empirical comparison of Simple Sequence Repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS One* **2**: e1367
- Hardie WJ** (2000) Grapevine biology and adaptation to viticulture. *Aust J Grape Wine Res* **6**: 74–81
- Hardy OJ, Vekemans X** (2002) SPAGeDI: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* **2**: 618–620
- Harper AL, Trick M, Higgins J, Fraser F, Clissold L, Wells R, Hattori C, Werner P, Bancroft I** (2012) Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nat Biotechnol* **30**: 798–802
- Herzog E, Töpfer R, Hausmann L, Eibach R, Frisch M** (2013) Selection strategies for marker-assisted background selection with chromosome-wise SSR multiplexes in pseudo-backcross programs for grapevine breeding. *Vitis* **52**: 193–196
- Hill CR, Butler JM, Vallone PM** (2009) A 26plex autosomal STR assay to aid human identity testing. *J Forensic Sci* **54**: 1008–1015
- Hill W, Robertson A** (1968) Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**: 226–231
- Huang H, Lu J, Ren Z, Hunter W, Dowd SE, Dang P** (2011) Mining and validating grape (*Vitis* L.) ESTs to develop EST-SSR markers for genotyping and mapping. *Mol Breed* **28**: 241–254
- Huang X, Han B** (2014) Natural variations and genome-wide association studies in crop plants. *Annu Rev Plant Biol* **65**: 531–551

- Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, et al** (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* **42**: 961–967
- Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C, et al** (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet* **44**: 32–39
- Hui J, Oka A, James A, Palmer LJ, Musk AW, Beilby J, Inoko H** (2008) A genome-wide association scan for asthma in a general Australian population. *Hum Genet* **123**: 297–306
- Ibáñez J, Carreño J, Yuste J, Martínez-Zapater JM** (2015) Grapevine breeding and clonal selection programmes in Spain. In AG Reynolds, ed, *Grapevine breeding programs for the wine industry: Traditional and molecular techniques*. Elsevier Science, pp 183–209
- Ibáñez J, Vargas AM, Palancar M, Borrego J, de Andres MT** (2009) Genetic relationships among table-grape varieties. *Am J Enol Vitic* **60**: 35–42
- Van Inghelandt D, Reif JC, Dhillon BS, Flament P, Melchinger AE** (2011) Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. *Theor Appl Genet* **123**: 11–20
- International Rice Genome Sequencing Project** (2005) The map-based sequence of the rice genome. *Nature* **436**: 793–800
- Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al** (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467
- Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A** (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* **15**: 740
- Jewell MC, Frere CH, Prentis PJ, Lambrides CJ, Godwin ID** (2010) Characterization and multiplexing of EST-SSR primers in *Cynodon* (Poaceae) species1. *Am J Bot* **97**: 99–101
- Jia G, Huang X, Zhi H, Zhao Y, Zhao Q, Li W, Chai Y, Yang L, Liu K, Lu H, et al** (2013) A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat Genet* **45**: 957–961
- Jones ES, Sullivan H, Bhatramakki D, Smith JSC** (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L.). *Theor Appl Genet* **115**: 361–371
- Jones G V., White MA, Cooper OR, Storchmann K** (2005) Climate change and global wine quality. *Clim Change* **73**: 319–343

- Jorde LB** (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Res* **10**: 1435–1444
- Jorgenson E, Witte JS** (2006) A gene-centric approach to genome-wide association studies. *Nat Rev Genet* **7**: 885–891
- Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK** (2010) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* **177**: 309–334
- Kalinowski ST, Taper ML, Marshall TC** (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol Ecol* **16**: 1099–1106
- Karaagac E, Vargas AM, Andrés MT, Carreño I, Ibáñez J, Carreño J, Martínez-Zapater JM, Cabezas JA** (2012) Marker assisted selection for seedlessness in table grape breeding. *Tree Genet Genomes* **8**: 1003–1015
- Kawashima M, Tamiya G, Oka A, Hohjoh H, Juji T, Ebisawa T, Honda Y, Inoko H, Tokunaga K** (2006) Genomewide association analysis of human narcolepsy and a new resistance gene. *Am J Hum Genet* **79**: 252–263
- Keller M** (2015) *The Science of grapevines: Anatomy and physiology*, 2nd ed. Academic Press
- Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh J a L, Barris W, Schnabel RD, Taylor JF, Raadsma HW** (2008) Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* **9**: 187
- Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M** (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet* **39**: 1151–1155
- Kobayashi S, Goto-Yamamoto N, Hirochika H** (2004) Retrotransposon-induced mutations in grape skin color. *Science* **304**: 982
- Kobayashi S, Ishimaru M, Hiraoka K, Honda C** (2002) *Myb*-related genes of the Kyoho grape (*Vitis labruscana*) regulate anthocyanin biosynthesis. *Planta* **215**: 924–933
- Korte A, Farlow A** (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**: 29
- Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M** (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* **44**: 1066–1071
- Krivanek AF, Riaz S, Walker MA** (2006) Identification and molecular mapping of *PdR1*, a primary resistance gene to Pierce's disease in *Vitis*. *Theor Appl Genet* **112**: 1125–1131

- Kruglyak L** (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**: 139–144
- Kujur A, Bajaj D, Upadhyaya HD, Das S, Ranjan R, Shree T, Saxena MS, Badoni S, Kumar V, Tripathi S, et al** (2015) A genome-wide SNP scan accelerates trait-regulatory genomic loci identification in chickpea. *Sci Rep* **5**: 11166
- Lacombe T, Boursiquot JJ-M, Laucou V, Di Vecchi-Staraz M, Péros J-P, This P** (2013) Large-scale parentage analysis in an extended set of grapevine cultivars (*Vitis vinifera* L.). *Theor Appl Genet* **126**: 401–414
- Lander ES, Schork NJ** (1994) Genetic dissection of complex traits. *Science* **265**: 2037–2048
- Laucou V, Lacombe T, Dechesne F, Siret R, Bruno J-P, Dessup M, Dessup T, Ortigosa P, Parra P, Roux C, et al** (2011) High throughput analysis of grape genetic diversity as a tool for germplasm collection management. *Theor Appl Genet* **122**: 1233–1245
- Lebon G, Wojnarowicz G, Holzapfel B, Fontaine F, Vaillant-Gaveau N, Clément C** (2008) Sugars and flowering in the grapevine (*Vitis vinifera* L.). *J Exp Bot* **59**: 2565–2578
- Lefort F, Kyvelos CJ, Zervou M, Edwards KJ, Roubelakis-Angelakis KA** (2002) Characterization of new microsatellite loci from *Vitis vinifera* and their conservation in some *Vitis* species and hybrids. *Mol Ecol Notes* **2**: 20–21
- Le Paslier M-C, Choisne N, Scalabrin S, Bacilieri R, Berard A, Bounon R, Boursiquot J-M, Bras M, Brunel D, Chauveau A, et al** (2013) The GRAPERSEQ 18K *Vitis* genotyping chip. IX International Symposium on Grapevine Physiology & Biotechnology, La Serina
- Levadoux L** (1956) Les populations sauvages et cultivées de *Vitis vinifera* L. *Ann l'Amélioration des Plantes* **6**: 59–118
- Lewontin RC** (1964) The Interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**: 49–67
- Li H, Durbin R** (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760
- Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N, et al** (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat Genet* **45**: 43–50
- Li Y, Haseneyer G, Schön C-C, Ankerst D, Korzun V, Wilde P, Bauer E** (2011) High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (*Secale cereale* L.) genes involved in frost response. *BMC Plant Biol* **11**: 6

- Lijavetzky D, Cabezas JA, Ibáñez A, Rodríguez V, Martínez-Zapater JM** (2007) High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* **8**: 424
- Lijavetzky D, Ruiz-García L, Cabezas J a., De Andrés MT, Bravo G, Ibáñez A, Carreño J, Cabello F, Ibáñez J, Martínez-Zapater JM** (2006) Molecular genetics of berry colour variation in table grape. *Mol Genet Genomics* **276**: 427–435
- Lippman Z, Tanksley SD** (2001) Dissecting the genetic pathway to extreme fruit size in tomato using a cross between the small-fruited wild species *Lycopersicon pimpinellifolium* and *L. esculentum* var. Giant Heirloom. *Genetics* **158**: 413–422
- Liu H, Bayer M, Druka A, Russell JR, Hackett CA, Poland J, Ramsay L, Hedley PE, Waugh R** (2014) An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. *BMC Genomics* **15**: 104
- Liu K, Muse S V** (2005) PowerMarker : an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**: 2128–2129
- Liu L, Wang L, Yao J, Zheng Y, Zhao C** (2010) Association mapping of six agronomic traits on chromosome 4A of wheat (*Triticum aestivum* L.). *Mol. Plant Breed.* doi: 10.5376/mpb.2010.01.0005
- Liu L, Wu Y** (2012) Development of a genome-wide multiple duplex-SSR protocol and its applications for the identification of selfed progeny in switchgrass. *BMC Genomics* **13**: 522
- Liu N, Chen L, Wang S, Oh C, Zhao H** (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genet* **6** Suppl 1: S26
- Lodhi MA, Reisch BI** (1995) Nuclear DNA content of *Vitis* species, cultivars, and other genera of the *Vitaceae*. *Theor Appl Genet* **90**: 11–16
- Loiselle B, Sork V, Nason J, Graham C** (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (*Rubiaceae*). *Am J Bot* **82**: 1420 – 1425
- Lowe KM, Walker M a** (2006) Genetic linkage map of the interspecific grape rootstock cross Ramsey (*Vitis champinii*) x Riparia Gloire (*Vitis riparia*). *Theor Appl Genet* **112**: 1582–1592
- Mandel JR, Nambeesan S, Bowers JE, Marek LF, Ebert D, Rieseberg LH, Knapp SJ, Burke JM** (2013) Association mapping and the genomic consequences of selection in sunflower. *PLoS Genet* **9**: e1003378
- El Mannai Y, Shehzad T, Okuno K** (2011) Variation in flowering time in sorghum core collection and mapping of QTLs controlling flowering time by association analysis. *Genet Resour Crop Evol* **58**: 983–989

- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al** (2009) Finding the missing heritability of complex diseases. *Nature* **461**: 747–753
- Marguerit E, Boury C, Manicki A, Donnart M, Butterlin G, Némorin A, Wiedemann-Merdinoglu S, Merdinoglu D, Ollat N, Decroocq S** (2009) Genetic dissection of sex determinism, inflorescence morphology and downy mildew resistance in grapevine. *Theor Appl Genet* **118**: 1261–1278
- Martins WS, Lucas DCS, Neves KF de S, Bertoli DJ** (2009) WebSat - A web software for microsatellite marker development. *Bioinformation* **3**: 282–283
- Masi P, Zeuli P, Donini P** (2003) Development and analysis of multiplex microsatellite markers sets in common bean (*Phaseolus vulgaris* L.). *Mol Breed* **303**–313
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP a, Hirschhorn JN** (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**: 356–369
- McCarthy MI, Hirschhorn JN** (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet* **17**: R156–65
- McGovern PE** (2003) Ancient wine: The search for the origins of viniculture. Princeton University Press
- McInerney CE, Allcock AL, Johnson MP, Bailie DA, Prodöhl PA** (2011) Comparative genomic analysis reveals species-dependent complexities that explain difficulties with microsatellite marker development in molluscs. *Heredity* **106**: 78–87
- Mejía N, Gebauer M, Munoz L, Hewstone N, Munoz C, Hinrichsen P** (2007) Identification of qtls for seedlessness, berry size, and ripening date in a seedless x seedless table grape progeny. *Am J Enol Vitic* **58**: 499–507
- Mejía N, Soto B, Guerrero M, Casanueva X, Houel C, de los Ángeles Miccono M, Ramos R, Le Cunff L, Boursiquot J-M, Hinrichsen P, et al** (2011) Molecular, genetic and transcriptional evidence for a role of *VvAGL11* in stenospermocarpic seedlessness in grapevine. *BMC Plant Biol* **11**: 57
- 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
- Mesak F, Tatarenkov A, Earley RL, Avise JC** (2014) Hundreds of SNPs vs. dozens of SSRs: which dataset better characterizes natural clonal lineages in a self-fertilizing fish? *Front Ecol Evol*. doi: 10.3389/fevo.2014.00074

- Metzker ML** (2010) Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31–46
- Michikawa Y, Suga T, Ishikawa A, Hayashi H, Oka A, Inoko H, Iwakawa M, Imai T** (2010) Genome wide screen identifies microsatellite markers associated with acute adverse effects following radiotherapy in cancer patients. *BMC Med Genet* **11**: 123
- Missiaggia A, Grattapaglia D** (2006) Plant microsatellite genotyping with 4-color fluorescent detection using multiple-tailed primers. *Genet Mol Res* **5**: 72–78
- Monforte JA, Iban E, Abad S, Arús P** (2005) Inheritance mode of fruit traits in melon: Heterosis for fruit shape and its correlation with genetic distance. *Euphytica* **144**: 31–38
- Morrell PL, Buckler ES, Ross-Ibarra J** (2011) Crop genomics: advances and applications. *Nat Rev Genet* **13**: 85–96
- Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, et al** (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci U S A* **110**: 453–458
- Mueller JC** (2004) Linkage disequilibrium for different scales and applications. *Brief Bioinform* **5**: 355–364
- Myles S** (2013) Improving fruit and wine: what does genomics have to offer? *Trends Genet* **29**: 190–196
- Myles S, Boyko ARA, Owens CCL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D, et al** (2011) Genetic structure and domestication history of the grape. *Proc Natl Acad Sci U S A* **108**: 3530–3535
- Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, Buckler E, Ware D** (2010) Rapid genomic characterization of the genus *Vitis*. *PLoS One* **5**: e8219
- Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES** (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21**: 2194–2202
- Nakabayashi K, Komaki G, Tajima A, Ando T, Ishikawa M, Nomoto J, Hata K, Oka A, Inoko H, Sasazuki T, et al** (2009) Identification of novel candidate loci for anorexia nervosa at 1q41 and 11q22 in Japanese by a genome-wide association analysis with microsatellite markers. *J Hum Genet* **54**: 531–537
- Nakaya A, Isobe SN** (2012) Will genomic selection be a practical method for plant breeding? *Ann Bot* **110**: 1303–1316
- Negrul A** (1946) Origin and classification of cultured grape. *In* A Baranov, YF Kai, MA Lazarevski, T V Palibin, NN Prosmoserdov, eds, *The Ampelography of the USSR*. Pischepromizdat, Moscow, pp 159–216

- Nicolas SD, Péros J-P, Lacombe T, Launay A, Le Paslier M-C, Bérard A, Mangin B, Valière S, Martins F, Le Cunff L, et al** (2016) Genetic diversity, linkage disequilibrium and power of a large grapevine (*Vitis vinifera* L) diversity panel newly designed for association studies. *BMC Plant Biol* **16**: 74
- Nordborg M, Tavaré S** (2002) Linkage disequilibrium: what history has to tell us. *Trends Genet* **18**: 83–90
- Nordborg M, Weigel D** (2008) Next-generation genetics in plants. *Nature* **456**: 720–723
- Ohashi J, Tokunaga K** (2003) Power of genome-wide linkage disequilibrium testing by using microsatellite markers. *J Hum Genet* **48**: 487–491
- OIV (International Organization of Vine and Wine)** (2011) OIV vine and wine outlook 2010–2011. OIV, Paris
- OIV (International Organization of Vine and Wine)** (2009) The OIV descriptor list for grape varieties and *Vitis* species (Second Edition). OIV, Paris
- Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, Kilian B, Graner A** (2012) Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC Plant Biol* **12**: 16
- Patil DT, Bronner MP, Portier BP, Fraser CR, Plesec TP, Liu X** (2012) A five-marker panel in a multiplex PCR accurately detects microsatellite instability-high colorectal tumors without control DNA. *Diagnostic Mol Pathol* **21**: 127–133
- Payseur BA, Place M, Weber JL** (2008) Linkage disequilibrium between STRPs and SNPs across the human genome. *Am J Hum Genet* **82**: 1039–1050
- Pelsy F** (2010) Molecular and cellular mechanisms of diversity within grapevine varieties. *Heredity* **104**: 331–340
- Peng JH, Bai Y, Haley SD, Lapitan NL V** (2009) Microsatellite-based molecular diversity of bread wheat germplasm and association mapping of wheat resistance to the Russian wheat aphid. *Genetica* **135**: 95–122
- Platt A, Vilhjálmsson BJ, Nordborg M** (2010) Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**: 1045–1052
- Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, et al** (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome J* **5**: 103
- Poland JA, Rife TW** (2012) Genotyping-by-Sequencing for plant breeding and genetics. *Plant Genome J* **5**: 92
- Pompanon F, Bonin A, Bellemain E, Taberlet P** (2005) Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* **6**: 847–859

- Poni S, Casalini L, Bernizzoni F, Civardi S, Intrieri C** (2006) Effects of early defoliation on shoot photosynthesis, yield components, and grape composition. *Am J Enol Vitic* **57**: 397–407
- Postolache D, Leonarduzzi C, Piotti A, Spanu I, Roig A, Fady B, Roschanski A, Liepelt S, Vendramin GG** (2013) Transcriptome versus genomic microsatellite markers: highly informative multiplexes for genotyping *Abies alba* Mill. and congeneric species. *Plant Mol Biol Report* **32**: 750–760
- Powell W, Machray GC, Provan J** (1996) Polymorphism revealed by simple sequence repeats. *Trends Plant Sci* **1**: 215–222
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D** (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909
- Pritchard JK, Przeworski M** (2001) Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**: 1–14
- Pritchard JK, Stephens M, Donnelly P** (2000a) Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P** (2000b) Association mapping in structured populations. *Am J Hum Genet* **67**: 170–181
- Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H** (2010) The pattern of linkage disequilibrium in German Holstein cattle. *Anim Genet* **41**: 346–56
- Raabová J, Hans G, Risterucci A-M, Jacquemart A-L, Raspé O** (2010) Development and multiplexing of microsatellite markers in the polyploid perennial herb, *Menyanthes trifoliata* (Menyanthaceae). *Am J Bot* **97**: e31–3
- Rabbi I, Hamblin M, Gedil M, Kulakow P, Ferguson M, Ikpan AS, Ly D, Jannink J-L** (2014) Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. *Crop Sci* **54**: 1384
- Rafalski A** (2002a) Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* **5**: 94–100
- Rafalski A, Morgante M** (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet* **20**: 103–111
- Rafalski JA** (2002b) Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci* **162**: 329–333
- Rafalski JA, Tingey S V** (1993) Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. *Trends Genet* **9**: 275–280

- Reddy UK, Abburi L, Abburi VL, Saminathan T, Cantrell R, Vajja VG, Reddy R, Tomason YR, Levi A, Wehner TC, et al** (2014) A genome-wide scan of selective sweeps and association mapping of fruit traits using microsatellite markers in watermelon. *J Hered* **106**: 166–176
- Regner F, Stadlbauer A, Eisenheld C, Kaserer H** (2000) Genetic relationships among pinots and related cultivars. *Am J Enol Vitic* **51**: 7–14
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, et al** (2001) Linkage disequilibrium in the human genome. *Nature* **411**: 199–204
- Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES** (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A* **98**: 11479–11484
- Riaz S, Krivanek AF, Xu K, Walker MA** (2006) 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* **113**: 1317–1329
- Riaz S, Tenscher AC, Rubin J, Graziani R, Pao SS, Walker MA** (2008) Fine-scale genetic mapping of two Pierce's disease resistance loci and a major segregation distortion region on chromosome 14 of grape. *Theor Appl Genet* **117**: 671–681
- Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE** (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc Natl Acad Sci U S A* **109**: 8872–8877
- Rolle L, Siret R, Segade SR, Maury C, Gerbi V, Jourjon F** (2011) Instrumental texture analysis parameters as markers of table-grape and winegrape quality: a review. *Am J Enol Vitic* **63**: 11–28
- Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, et al** (2013) Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol* **14**: R55
- Rosenberg N a, Li LM, Ward R, Pritchard JK** (2003) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73**: 1402–1422
- Rozen S, Skaletsky H** (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**: 365–386
- Rutkoski JE, Poland J, Jannink J-L, Sorrells ME** (2013) Imputation of unordered markers and the impact on genomic selection accuracy. *G3* **3**: 427–439

- Saka S, Hirawa N, Oka A, Yatsu K, Hirukawa T, Yamamoto R, Matsusaka T, Imai E, Narita I, Endoh M, et al** (2015) Genome-wide association study of IgA nephropathy using 23 465 microsatellite markers in a Japanese population. *J Hum Genet*. doi: 10.1038/jhg.2015.88
- Sanchez LA, Dokoozlian NK** (2005) Bud microclimate and fruitfulness in *Vitis vinifera* L. *Am J Enol Vitic* **56**: 319–329
- Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M** (2014) Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol* **165**: 1120–1132
- Sayama T, Hwang T-Y, Komatsu K, Takada Y, Takahashi M, Kato S, Sasama H, Higashi A, Nakamoto Y, Funatsuki H, et al** (2011) Development and application of a whole-genome simple sequence repeat panel for high-throughput genotyping in soybean. *DNA Res* **18**: 107–115
- Schilling MP, Wolf PG, Duffy AM, Rai HS, Rowe CA, Richardson BA, Mock KE** (2014) Genotyping-by-sequencing for *Populus* population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS One* **9**: e95292
- Schlötterer C** (2004) The evolution of molecular markers--just a matter of fashion? *Nat Rev Genet* **5**: 63–69
- Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ** (2000) Analysis of SSRs derived from grape ESTs. *Theor Appl Genet* **100**: 723–726
- Sefc KM, Pejić I, Maletić E, Thomas MR, Lefort F** (2009) Microsatellite markers for grapevine: tools for cultivar identification & pedigree reconstruction. In KA Roubelakis-Angelakis, ed, *Grapevine molecular physiology & biotechnology*. Springer Netherlands, Dordrecht, pp 565–596
- Sefc KM, Regner F, Turetschek E, 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
- Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M** (2012) An efficient multi-locus mixed model approach for genome-wide association studies in structured populations. *Nat Genet* **44**: 825–830
- Sepahi A** (1980) Estimating cluster compactness in Yaghouti grapes. *Vitis* **19**: 81–90
- Sham P, Bader JS, Craig I, O'Donovan M, Owen M** (2002) DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet* **3**: 862–871
- Shavrukov YN, Dry IB, Thomas MR** (2004) Inflorescence and bunch architecture development in *Vitis vinifera* L. *Aust J Grape Wine Res* **10**: 116–124

- Shibata H, Yamamoto K, Sun Z, Oka A, Inoko H, Arinami T, Inada T, Ujike H, Itokawa M, Tochigi M, et al** (2013) Genome-wide association study of schizophrenia using microsatellite markers in the Japanese population. *Psychiatr Genet* **23**: 117–123
- Simko I, Eujayl I, van Hintum TJL** (2012) Empirical evaluation of DArT, SNP, and SSR marker-systems for genotyping, clustering, and assigning sugar beet hybrid varieties into populations. *Plant Sci* **184**: 54–62
- Singh N, Choudhury DR, Singh AK, Kumar S, Srinivasan K, Tyagi RK, Singh NK, Singh R** (2013) Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLoS One* **8**: e84136
- Slatkin M** (1994) Linkage disequilibrium in growing and stable populations. *Genetics* **137**: 331–336
- Slatkin M** (2008) Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**: 477–485
- Smith JR, Carpten JD, Brownstein MJ, Ghosh S, Magnuson VL, Gilbert DA, Trent JM, Collins FS** (1995) Approach to genotyping errors caused by nontemplated nucleotide addition by Taq DNA polymerase. *Genome Res* **5**: 312–317
- Sommer KJ, Islam MT, Clingeleffer PR** (2000) Light and temperature effects on shoot fruitfulness in *Vitis vinifera* L. cv. Sultana: Influence of trellis type and grafting. *Aust J Grape Wine Res* **6**: 99–108
- Stich B, Möhring J, Piepho H-P, Heckenberger M, Buckler ES, Melchinger AE** (2008) Comparison of mixed-model approaches for association mapping. *Genetics* **178**: 1745–1754
- Stram D** (2014) Design, analysis, and interpretation of genome-wide association scans (Series: Statistics for biology and health), 1st ed. Springer-Verlag, New York
- Syvänen a C** (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* **2**: 930–942
- Tamiya G, Shinya M, Imanishi T, Ikuta T, Makino S, Okamoto K, Furugaki K, Matsumoto T, Mano S, Ando S, et al** (2005) Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. *Hum Mol Genet* **14**: 2305–2321
- Tello J, Ibáñez J** (2014) Evaluation of indexes for the quantitative and objective estimation of grapevine bunch compactness. *Vitis* **53**: 9–16
- Tello J, Torres-Pérez R, Grimplet J, Ibáñez J** (2015) Association analysis of grapevine bunch traits using a comprehensive approach. *Theor Appl Genet* **129**: 227–242
- Tero N, Neumeier H, Gudavalli R, Schlötterer C** (2006) *Silene tatarica* microsatellites are frequently located in repetitive DNA. *J Evol Biol* **19**: 1612–1619

- The Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815
- This P, Lacombe T, Cadle-Davidson M, Owens CL** (2007) Wine grape (*Vitis vinifera* L.) color associates with allelic variation in the domestication gene *VvmybA1*. *Theor Appl Genet* **114**: 723–730
- This P, Lacombe T, Thomas MR** (2006) Historical origins and genetic diversity of wine grapes. *Trends Genet* **22**: 511–519
- This P, Zapater JM, Péros J-P, Lacombe T** (2011) Natural variation in *Vitis*. In AF Adam-Blondon, JM Martínez-Zapater, C Kole, eds, *Genetics, genomics, and breeding of grapes*. CRC Press, pp 30–67
- Thomas MR, Matsumoto S, Cain P, Scott NS** (1993) Repetitive DNA of grapevine: classes present and sequences suitable for cultivar identification. *Theor Appl Genet* **86**: 173–180
- Thomas MR, Scott NS** (1993) Microsatellite repeats in grapevine reveal DNA polymorphisms when analysed as sequence-tagged sites (STSs). *Theor Appl Genet* **86**: 985–990
- Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES** (2001) *Dwarf8* polymorphisms associate with variation in flowering time. *Nat Genet* **28**: 286–289
- Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES** (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* **43**: 159–162
- Tomasi D, Jones G V., Giust M, Lovat L, Gaiotti F** (2011) Grapevine phenology and climate change: relationships and trends in the veneto region of italy for 1964-2009. *Am J Enol Vitic* **62**: 329–339
- Töpfer R, Ludger H, Eibach R** (2011) Molecular breeding. In AF Adam-Blondon, JM Martínez-Zapater, C Kole, eds, *Genetics, genomics, and breeding of grapes*. CRC Press, pp 160–185
- Torregrosa L, Fernandez L, Bouquet A, Boursiquot J-M, Pelsy F, JM M-Z** (2011) Origins and consequences of somatic variation in grapevine. In AF Adam-Blondon, JM Martínez-Zapater, C Kole, eds, *Genetics, genomics, and breeding of grapes*. CRC Press, pp 68–92
- Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al** (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604
- Vargas AM** (2009) Estudio de diversidad en variedades de uva de mesa (*Vitis vinifera* l.) y su aplicación al análisis de asociación para caracteres de calidad. Universidad de Alcalá, Alcalá de Henares

- Vargas AM, Le Cunff L, This P, Ibáñez J, de Andrés MT (2013a)** *VvGAI1* polymorphisms associate with variation for berry traits in grapevine. *Euphytica* **191**: 85–98
- Vargas AM, Fajardo C, Borrego J, De Andrés MT, Ibáñez J (2013b)** Polymorphisms in *VvPel* associate with variation in berry texture and bunch size in the grapevine. *Aust J Grape Wine Res* **19**: 193–207
- Varilo T, Paunio T, Parker A, Perola M, Meyer J, Terwilliger JD, Peltonen L (2003)** The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Hum Mol Genet* **12**: 51–59
- Velasco R, Zharkikh A, Troggio M, Cartwright D a, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, et al (2007)** A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* **2**: e1326
- Vitulo N, Forcato C, Carpinelli E, Telatin A, Campagna D, D'Angelo M, Zimbello R, Corso M, Vannozzi A, Bonghi C, et al (2014)** A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biol* **14**: 99
- Wagner a P, Creel S, Kalinowski ST (2006)** Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity* **97**: 336–345
- Walker AR, Lee E, Bogs J, McDavid D a J, Thomas MR, Robinson SP (2007)** White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J* **49**: 772–785
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, et al (1998)** Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082
- Wang Y-H, Upadhyaya HD, Burrell a M, Sahraeian SME, Klein RR, Klein PE (2013)** Genetic structure and linkage disequilibrium in a diverse, representative collection of the C4 model plant, *Sorghum bicolor*. *G3* **3**: 783–793
- Ward JA, Bhangoo J, Fernández-Fernández F, Moore P, Swanson JD, Viola R, Velasco R, Bassil N, Weber CA, Sargent DJ (2013)** Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. *BMC Genomics* **14**: 2
- Weiss KM, Clark AG (2002)** Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* **18**: 19–24
- Welter L, Grando Ms, Zyprian E (2011)** Basics of grapevine genetic analysis. In AF Adam-Blondon, JM Martínez-Zapater, C Kole, eds, *Genetics, genomics, and breeding of grapes*. CRC Press, pp 137–159

- Welter LJ, Göktürk-Baydar N, Akkurt M, Maul E, Eibach R, Töpfer R, Zyprian EM (2007) Genetic mapping and localization of quantitative trait loci affecting fungal disease resistance and leaf morphology in grapevine (*Vitis vinifera* L). *Mol Breed* **20**: 359–374
- Winkler AJ, Cook JA, Kliewer WM, Lider LA (1974) General viticulture. University of California Press, CA
- Würschum T, Langer SM, Longin CFH, Korzun V, Akhunov E, Ebmeyer E, Schachschneider R, Schacht J, Kazman E, Reif JC (2013) Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. *Theor Appl Genet* **126**: 1477–1486
- Wycislo AP, Clark JR, Karcher DE (2008) Fruit shape analysis of *Vitis* using digital photography. *HortScience* **43**: 677–680
- Xiong M, Jin L (1999) Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods. *Am J Hum Genet* **64**: 629–640
- Xu K, Riaz S, Roncoroni NC, Jin Y, Hu R, Zhou R, Walker M a (2008) Genetic and QTL analysis of resistance to *Xiphinema index* in a grapevine cross. *Theor Appl Genet* **116**: 305–311
- Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* **30**: 105–111
- Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One* **4**: e8451
- Yang N, Lu Y, Yang X, Huang J, Zhou Y, Ali F, Wen W, Liu J, Li J, Yan J (2014) Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet* **10**: e1004573
- Yang X, Xu Y, Shah T, Li H, Han Z, Li J, Yan J (2011) Comparison of SSRs and SNPs in assessment of genetic relatedness in maize. *Genetica* **139**: 1045–1054
- Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**: 134
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–308

- Yu J, Zhang Z, Zhu C, Tabanao DA, Pressoir G, Tuinstra MR, Kresovich S, Todhunter RJ, Buckler ES** (2009) Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. *Plant Genome J* **2**: 63
- Zarouri B, Vargas AM, Gaforio L, Aller M, de Andrés MT, Cabezas JA** (2015) Whole-genome genotyping of grape using a panel of microsatellite multiplex PCRs. *Tree Genet Genomes* **11**: 17
- Zhang D-X** (2004) Lepidopteran microsatellite DNA: redundant but promising. *Trends Ecol Evol* **19**: 507–509
- Zhang J, Hausmann L, Eibach R, Welter LJ, Töpfer R, Zyprian EM** (2009) A framework map from grapevine V3125 (*Vitis vinifera* “Schiava grossa” x “Riesling”) x rootstock cultivar “Börner” (*Vitis riparia* x *Vitis cinerea*) to localize genetic determinants of phylloxera root resistance. *Theor Appl Genet* **3125**: 1039–1051
- Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore M a, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al** (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**: 355–360
- Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, et al** (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**: e4
- Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, et al** (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun* **2**: 467
- Zhao Y, Wang H, Chen W, Li Y** (2014) Genetic structure, linkage disequilibrium and association mapping of Verticillium wilt resistance in elite cotton (*Gossypium hirsutum* L.) germplasm population. *PLoS One* **9**: e86308
- Zhu C, Gore M, Buckler ES, Yu J** (2008) Status and Prospects of Association Mapping in Plants. *Plant Genome J* **1**: 5
- Zondervan KT, Cardon LR** (2004) The complex interplay among factors that influence allelic association. *Nat Rev Genet* **5**: 89–100

ANNEXES

ANNEXES

Table A.1. List of the 274 grapevine accessions included in the mapping collection and their main features (see the key at the bottom of the table)

Mapping collection							Phenotyping			GBS		Population structure		
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
ABONDANT	ABONDANT	BGVCAM2331	W	FRA	x		x	x	x	1473895	0.99	0.39	0.09	0.52
AGOSTENGA (PRIÉ BLANC)	AGOSTENGA	BGVCAM0923	T	ITA	x		x	x	x	235519	0.67	0.45	0.20	0.35
AGUDELO (CHENIN BLANC)	AGUDILLO	BGVCAM1664	W	ESP	x		x	x	x	26402	0.56	0.92	0.04	0.05
ALBA MAGNA	ALBA MAGNA	BGVCAM1216	T	ITA	x		x	x	x	59865	N	0.16	0.05	0.79
ALBARIÑO	ALBARIÑO	BGVCAM1931	W	ESP	x		x	x	x	1315842	0.98	0.93	0.02	0.05
ALBILLO DE POZO	ALBILLO	BGVCAM0926	W	ESP	x		x	x	x	337568	0.70	0.06	0.91	0.03
ALBILLO MAYOR	TURRUNTÉS	BGVCAM1177	W	ESP	x		x	x	x	88995	N	0.04	0.92	0.04
ALBILLO REAL	NIEVES TEMPRANO	BGVCAM0907	W	ESP	x		x	x	x	212615	0.92	0.16	0.42	0.42
ALEDO (CHERTA)	ALEDO	BGVCAM1191	T	ESP	x		x	x	x	663108	0.93	0.05	0.89	0.06
ALEDO REAL	ALEDO REAL	BGVCAM1402	T	ESP	x		x	x	x	217272	0.79	0.07	0.88	0.06
ALIGOTE	ALIGOTÉ	BGVCAM2603	W	FRA	x		x	x	x	1623013	0.99	0.95	0.03	0.02
ALJIBE	ALJIBE	BGVCAM2391	W	ESP	x		x	x	x	993123	0.97	0.05	0.17	0.78
ALPHONSE LAVALLEE	LEOPOLD III	BGVCAM1338	T	FRA	x		x	x	x	241138	0.81	0.02	0.02	0.95
ANGELO PIROVANO	PERGOLESE	BGVCAM0855	T	ITA	x		x	x	x	96457	0.65	0.10	0.02	0.88
ARAMON	ARAMON	BGVCAM1106	W	FRA	x		x	x	x	1550340	0.99	0.24	0.70	0.06
ARVINE PETITE	ARVINE PETITE	BGVCAM2600	W	CHE	x		x	x	x	1942415	0.99	0.75	0.10	0.15
ASSIRTICO	ASSIRTICO	BGVCAM0796	W	GRC	x		x	x	x	843985	0.95	0.13	0.29	0.58
ATALAYA	ATALAYA	BGVCAM2689	T/W	ESP		x			x	200617	0.50	0.41	0.29	0.30
ATTILIO RAGGIONERI	ATTILIO RAGGIONERI	BGVCAM1474	T	ITA	x		x	x	x	284715	0.65	0.19	0.06	0.75
AUBUN (CORVO)	AMBROSINA	BGVCAM2808	W	FRA		x			x	540814	0.90	0.49	0.40	0.11
AZAL TINTO	AZAL TINTO	BGVCAM2790	W	PRT	x		x	x	x	218691	0.85	0.95	0.03	0.02
BABIC	BABIC	BGVCAM2627	W	HRV	x		x	x	x	126586	0.84	0.06	0.08	0.86
BARBABLEU	BARBABLEU	BGVCAM0941	T	ITA	x		x	x	x	781247	0.93	0.04	0.33	0.64
BARESANA(KORITHI ASPRO)	THURCHESACA BLANCA	BGVCAM2824	T	IND		x			x	18786	N	0.03	0.79	0.17
BARLINKA FAUX (ALMERIA)	BARLINKA	BGVCAM1427	T		x		x	x	x	280137	0.93	0.03	0.12	0.85
BASCINA BIJELA	BASCINA BIJELA	BGVCAM2780	W	HRV	x		x	x	x	124996	0.84	0.09	0.20	0.71
BASTARDO BLANCO	PEDRO XIMÉNEZ	BGVCAM2223	W	ESP	x		x	x	x	699002	0.93	0.92	0.04	0.03
BEAUTY SEEDLESS	BEAUTY SEEDLESS	BGVCAM0797	T	USA	x		x	x	x	1901777	0.99	0.06	0.82	0.12
BEBA	BEBA DE JAÉN	BGVCAM2341	T/W	ESP	x		x	x	x	1614517	0.99	0.04	0.90	0.06
BEITAMOUNI	OSCARI ROSE	BGVCAM1272	T	LBN	x		x	x	x	1304911	0.98	0.09	0.62	0.30

Mapping collection							Phenotyping			GBS		Population structure		
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
BELLONE	BELLONE	BGVCAM2242	W	ITA	x		x	x	x	535007	0.87	0.10	0.20	0.70
BERBECEL	KOKORKO	BGVCAM1814	W	BGR	x		x	x	x	64043	0.72	0.26	0.13	0.61
BERMEJUELA	BERMEJUELA	BGVCAM2686	W	ESP	x		x	x	x	689603	0.93	0.52	0.43	0.05
BIBIOLA	TRENTAM BLACK	BGVCAM1510	T	ITA	x		x	x	x	823800	0.96	0.48	0.15	0.37
BLACK ALICANTE	BLACK ALICANTE	BGVCAM1477	T		x		x	x	x	298708	0.90	0.28	0.14	0.58
BLACK MONUCCA	BLACK MONUCCA	BGVCAM1722	T	USA	x		x	x	x	558105	0.88	0.02	0.96	0.02
BLAUFRAENKISCH	BLAUFRAENKISCH	BGVCAM1991	W	AUT	x		x	x	x	4241978	0.99	0.52	0.41	0.07
BOBAL	BOBAL	BGVCAM2051	W	ESP	x		x	x	x	22870	0.54	0.08	0.84	0.09
BOGNI 08	BOGNI 08	BGVCAM1377	T	ITA	x		x	x	x	62908	N	0.13	0.05	0.82
BONARDA (PIEMONTESE)	BONARDA	BGVCAM2642	W	ITA	x		x	x	x	771971	0.95	0.34	0.03	0.63
BRANCELLAO	SERRADILLO	BGVCAM1671	W	ESP	x		x	x	x	585356	0.89	0.92	0.03	0.05
BRUN ARGENTE	BRUN ARGENTE	BGVCAM2400	W	FRA		x			x	63610	N	0.62	0.27	0.11
BRUNI 125	BRUNI 125	BGVCAM1480	T	ITA	x		x	x	x	12076	0.65	0.61	0.02	0.37
BRUNI 415	BRUNI 415	BGVCAM1723	T	ITA	x		x	x	x	65778	0.72	0.19	0.16	0.64
BRUÑAL (ALBARÍN TINTO, BABOSO TINTO)	ALFROCHEIRO	BGVCAM1192	W	PRT		x			x	582546	0.91	0.85	0.08	0.07
BUDAI ZOELD	BUDAI ZOLD	BGVCAM2244	W	HUN	x		x	x	x	44784	0.75	0.42	0.10	0.48
CABERNET FRANC	VERDEJILLA TINTO	BGVCAM1637	W	ESP	x		x	x	x	1178337	0.98	0.87	0.08	0.05
CABERNET SAUVIGNON	CABERNET SAUVIGNON	BGVCAM2608	W	FRA	x		x	x	x	246656	0.90	0.95	0.03	0.02
CAIÑO BLANCO	CAIÑO BLANCO	BGVCAM2687	W	ESP		x			x	2880515	0.99	0.97	0.01	0.02
CAIÑO TINTO	CAIÑO DEL PAIS	BGVCAM1865	W	ESP	x		x	x	x	1305055	0.98	0.92	0.03	0.04
CALMERIA	FLORA	BGVCAM2838	T	USA		x			x	293338	0.66	0.03	0.94	0.03
CANAIOLO	CANAIOLO	BGVCAM2405	W	ITA	x		x	x	x	569029	0.89	0.21	0.29	0.50
CANNER SEEDLESS	CANNER SEEDLESS	BGVCAM1378	T	USA	x		x	x	x	3093755	0.99	0.03	0.95	0.02
CARMENERE	CARMENERE	BGVCAM2770	W	FRA		x			x	1244750	0.98	0.94	0.03	0.03
CASTAÑAL	CASTAÑAL	BGVCAM2581	W	ESP	x		x	x	x	48496	N	0.85	0.05	0.10
CASTELLANA NEGRA (TINTO CÃO)	TINTO CAO	BGVCAM2787	W	PRT	x		x	x	x	854379	0.97	0.43	0.32	0.26
CERCIAL	CERCIAL	BGVCAM1208	W	PRT	x		x	x	x	206833	0.91	0.51	0.46	0.03
CHAOUCH BLANC	CAVUS	BGVCAM1760	T	TUR	x		x	x	x	40949	0.71	0.03	0.47	0.50
CHAOUCH ROZOVI	CHAOUCH ROSE	BGVCAM1727	T	TUR	x		x	x	x	198837	0.91	0.03	0.12	0.86
CHARDONNAY	CHARDONNAY	BGVCAM1121	W	FRA	x		x	x	x	2007840	0.99	0.93	0.05	0.02
CHASSELAS H.S.	CHASSELAS H.S.	BGVCAM0804	W		x		x	x	x		0.14	0.74	0.12	
CHASSELAS VIOLET	CHASSELAS VIOLET	BGVCAM1798	T	CHE	x		x	x	x	1599453	0.99	0.64	0.02	0.34
CHONDROSTAFIDA	LOCAL BLACK	BGVCAM1213	T	GRC	x		x	x	x	31136	0.63	0.04	0.17	0.78

Mapping collection						Phenotyping			GBS		Population structure			
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
CHRISTMAS ROSE	CHRISTMAS ROSE	BGVCAM2753	T	USA		x			x	2258135	0.99	0.05	0.75	0.20
CLAIRETTE DE GASCOGNE	BLANC DAME	BGVCAM2243	W	FRA		x				4595828	0.99			
CLAIRETTE MAZEL	MAZEL	BGVCAM1011	W	FRA	x		x	x	x	185228	0.84	0.22	0.03	0.75
CLOTILDE PROSPERI	CLOTILDE PROSPERI	BGVCAM1355	T	ITA	x		x	x	x	26776	N	0.03	0.06	0.91
COLOMBARD	COLOMBARD	BGVCAM2250	W	FRA		x			x	1393374	0.99	0.88	0.09	0.03
CORINTO (KORINTHIAKI)	BLACK CURRANT	BGVCAM1212	T	GRC	x		x	x	x	1447715	0.98	0.02	0.65	0.33
CORNIOLA DI MILAZZO	CORNIOLA DI MILAZZO	BGVCAM1220	T	ITA	x		x	x	x	20683	N	0.03	0.54	0.43
CORVINA VERONESE	CORVINA	BGVCAM2776	W	ITA	x		x	x	x	254029	0.89	0.58	0.27	0.15
CROATINA	CROATINA	BGVCAM2779	W	ITA	x		x	x	x	254083	0.94	0.40	0.15	0.45
CUATENDRÁ	CUATENDRÁ	BGVCAM1129	W	ESP	x		x	x	x	3215948	0.99	0.75	0.20	0.06
DABOUKI	DABOUKI	BGVCAM1384	T	ARM	x		x	x	x	78173	0.72	0.02	0.90	0.09
DANLAS	DANLAS	BGVCAM2349	T	FRA	x		x	x	x	582310	0.90	0.27	0.38	0.35
DATTIER NOIR	DATTIER NOIR	BGVCAM1491	T		x		x	x	x	1017283	0.98	0.02	0.77	0.21
DE CILINDRO	DE CILINDRO	BGVCAM1610	T	ESP	x		x	x	x	1821261	0.99	0.08	0.88	0.05
DE CUERNO	DE CUERNO	BGVCAM2078	T	ESP	x		x	x	x	1596065	0.99	0.02	0.33	0.66
DELHRO	DELHRO	BGVCAM2835	T	FRA		x			x	117600	N	0.03	0.08	0.90
DELIGHT	DELIGHT	BGVCAM1253	T	USA	x		x	x	x	1011616	0.98	0.02	0.89	0.09
DIMYAT	SMEDEREVKA	BGVCAM1230	T/W	BGR	x		x	x	x	145153	0.88	0.53	0.09	0.38
DODRELYABI	GROS COLMAN	BGVCAM0984	T	GEO	x		x	x	x	287899	0.89	0.04	0.04	0.92
DOLCETTO	DOLCETTO	BGVCAM2414	W	ITA	x		x	x	x	278449	0.85	0.31	0.35	0.34
DOMINGA	DON MARIANO	BGVCAM2181	T/W	ESP		x			x	269894	0.63	0.05	0.17	0.79
DON MARIANO	DON MARIANO	BGVCAM2127	T	ESP	x		x	x	x	1921168	0.99	0.02	0.95	0.03
DOÑA MARÍA (DONA MARIA)	DOÑA MARÍA	BGVCAM2836	T	ESP		x			x	711798	0.94	0.02	0.03	0.96
DRNEKUSA	DRNEKUSA	BGVCAM2781	W	YUG		x			x	20918	N	0.05	0.37	0.59
DURIF	PRECOCE HERMING	BGVCAM1050	W	FRA	x		x	x	x	1194572	0.98	0.76	0.03	0.21
EL FARRYALI	EL FARRYALI	BGVCAM1494	T	MAR	x		x	x	x	123703	0.64	0.10	0.84	0.06
ELBLING WEISS	ELBLING	BGVCAM2353	W	DEU	x		x	x	x	5232	N	0.92	0.04	0.03
EMERALD SEEDLESS	EMERALD SEEDLESS	BGVCAM1472	T	USA	x		x	x	x	26415	0.58	0.03	0.49	0.48
FAKADI	FAKADI	BGVCAM2888	W	IND		x			x	2309127	0.99	0.04	0.03	0.93
FERNAN PIRIZ (RABO DE OVELHA)	FERNAN PIRIZ	BGVCAM0970	W	ESP	x		x	x	x	1159258	0.98	0.44	0.51	0.05
FERRAL	FERRAL	BGVCAM0814	T/W	ESP	x		x	x	x	1942239	0.99	0.02	0.96	0.02
FETEASCA REGALA	FETEASCA REGALA	BGVCAM2599	W	ROU		x			x	968925	0.97	0.32	0.07	0.61
FLAME SEEDLESS	FLAME SEEDLESS	BGVCAM2680	T	USA		x			x	1511194	0.98	0.03	0.90	0.07
FURMINT	FURMINT	BGVCAM2356	W	HUN	x		x	x	x	92498	0.86	0.27	0.04	0.70

Mapping collection							Phenotyping			GBS		Population structure		
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
FUSCA	FUSCA	BGVCAM1221	T	ITA	x		x	x	x	1796263	0.99	0.23	0.05	0.73
GAFARRO	GAFARRO	BGVCAM2803	W	ESP		x			x	505330	0.88	0.15	0.67	0.18
GALLETTA ROSA	620 PIROVANO	BGVCAM2588	T	ITA	x		x	x	x	945545	0.97	0.11	0.34	0.54
GALLURAZENNI DE DAMASCO	CONCA D'ORO	BGVCAM1434	T	ITA	x		x	x	x	2199978	0.99	0.03	0.47	0.50
GAMAY	GAMAY TEINTURIER DE FREAU	BGVCAM2428	W	FRA	x		x	x	x	673874	0.94	0.95	0.04	0.02
GARNACHA	GARNACHA	BGVCAM0975	W	ESP	x		x	x	x	480640	0.82	0.35	0.62	0.03
GARNACHA TINTORERA (ALICANTE BOUSCHET)	ALICANTE	BGVCAM2172	W	ESP	x		x	x	x	261603	0.89	0.30	0.56	0.15
GARRIDO MACHO	PRIMITIVO	BGVCAM2620	T/W	ESP		x			x	2862153	0.99	0.03	0.79	0.18
GARRO (MANDON)	GALMETE	BGVCAM1542	W	ESP		x			x	1434227	0.99	0.48	0.49	0.04
GENERAL DE LA MARMORA	GENERAL DE LA MARMORA	BGVCAM1388	T	FRA	x		x	x	x	1147812	0.98	0.17	0.04	0.79
GEWÜRZTRAMINER	ALBARÍÑO	BGVCAM1103	W			x			x	129933	N	0.97	0.01	0.02
GIRO (NEGRO, RUBIO)	GIRO NEGRO	BGVCAM2765	W	ESP		x			x	504868	0.78	0.07	0.64	0.29
GRACIANO	GRACIANO	BGVCAM2133	W	ESP	x		x	x	x	312878	0.91	0.71	0.24	0.06
GRASA DE COTNARI	KOVER SZOLO	BGVCAM2255	T/W	MDA	x		x	x	x	641101	0.93	0.34	0.06	0.60
GRAZIELLA	GRAZIELLA I	BGVCAM1325	T	ITA	x		x	x	x	2197979	0.99	0.07	0.22	0.71
GROLLEAU (GRIS, NOIR)	GROLLEAU	BGVCAM2771	W	FRA		x			x	830624	0.96	0.74	0.06	0.21
GUAL (TORRONTÉS, BOAL CACHUDO, MALVASÍA FINA, TARANTEY DE NERGA)	MOSCATEL	BGVCAM1669	W	ESP		x			x	4131063	0.99	0.36	0.60	0.04
HARSLEVELU	HARSLEVELU	BGVCAM2782	W	HUN	x		x	x	x	14424	0.75	0.08	0.04	0.88
HEBÉN	HEBÉN	BGVCAM2030	W	ESP	x		x	x	x	606529	0.91	0.03	0.96	0.02
HEUNISCH WEISS	HEUNISCH WEISS	BGVCAM2767	W		x		x	x	x	1366368	0.98	0.87	0.10	0.03
HIBOU BLANC	GIBI	BGVCAM2637	W	FRA	x		x	x	x	23161	0.74	0.82	0.13	0.05
HONDARRABI ZURI (COURBU BLANC)	ONDARRABI ZURI	BGVCAM2683	W	ESP		x			x	1627455	0.98	0.87	0.06	0.07
HORA	CORNICHON VIOLET	BGVCAM1448	T	BGR	x		x	x	x	202003	0.92	0.20	0.67	0.13
IMPERIAL ROJA	IMPERIAL ROJA	BGVCAM1615	T	ESP	x		x	x	x	2047705	0.98	0.02	0.94	0.03
ITALIA	REGINA	BGVCAM1347	T	ITA	x		x	x	x	1321420	0.99	0.07	0.05	0.88
JUAN GARCIA (MOURATON)	MENCIA	BGVCAM1861	W	ESP		x			x	466871	0.84	0.52	0.45	0.03
KADARKA	KADARKA	BGVCAM1996	W	HUN	x		x	x	x	525354	0.88	0.12	0.03	0.85
KERNER (WES25-30)	KERNER	BGVCAM2595	W	DEU		x			x	41744	N	0.50	0.03	0.48
KHALILI BELYI	KALILI	BGVCAM2362	T	AZE	x		x	x	x	1375557	0.98	0.05	0.56	0.39
KOKOUR BELY	KOKUR	BGVCAM2644	W	UKR	x		x	x	x	114641	0.80	0.04	0.68	0.29

Mapping collection						Phenotyping			GBS		Population structure			
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
KOSHU	KOSHU	BGVCAM2635	W	JPN	x		x	x	x	326057	0.80	0.03	0.87	0.10
KRATOSIJA	DIMITIS	BGVCAM1292	T/W	GRC	x		x	x	x	1630272	0.99	0.30	0.24	0.46
LADY DOWNE'S SEEDLING	LADY DOWNE'S SEEDLING	BGVCAM1497	T	GBR	x		x	x	x	3500955	0.99	0.21	0.03	0.76
LASINA	LASSINA	BGVCAM1238	T	HRV	x		x	x	x	828092	0.96	0.70	0.11	0.19
LATINA	LATINA	BGVCAM1326	T	ITA	x		x	x	x	924604	0.96	0.03	0.07	0.90
LATTUARIO NERO	DORONI MACERON	BGVCAM2351	T	ISR	x		x	x	x	432439	0.79	0.03	0.57	0.40
LISTÁN PRIETO (MISSION)	LISTÁN NEGRO	BGVCAM1590	W	ESP	x		x	x	x	328007	0.71	0.31	0.15	0.54
LJUTAC	LJUTAC	BGVCAM2778	W	YUG		x			x	244143	0.59	0.04	0.27	0.69
LOUREIRA	LOUREIRO BLANCO	BGVCAM1138	W	ESP	x		x	x	x	99382	0.83	0.79	0.18	0.04
LUGLIENGA BIANCA	LUGLIENGA BIANCA	BGVCAM0997	T	ITA	x		x	x	x	22876	N	0.38	0.27	0.35
LUISA BLANCA	FORCALLAT	BGVCAM0973	W	ESP	x		x	x	x	1753203	0.99	0.02	0.94	0.04
LUNGYEN	LUNGYEN	BGVCAM2791	T/W	CHN	x		x	x	x	35376	0.49	0.12	0.75	0.13
MALBEC (COT)	COT VERT	BGVCAM0961	W	FRA	x		x	x	x	3946068	0.99	0.85	0.06	0.09
MALMSEY	MALMSEY	BGVCAM2259	W	USA	x		x	x	x	20786	N	0.12	0.84	0.04
MALVASÍA AROMÁTICA (MALVASÍA DE SITGES)	MALVASÍA	BGVCAM1957	T/W	ESP	x		x	x	x	794475	0.96	0.38	0.56	0.06
MALVASIA BIANCA LUNGA	MALVASIA CHIANTI	BGVCAM2260	W	ITA	x		x	x	x	212550	0.87	0.08	0.38	0.54
MALVASIA DI CANDIA	MALVASIA CANDIA AROMÁTICA	BGVCAM2614	W	ITA		x			x	9116	N	0.29	0.13	0.58
MALVASIA VOLCÁNICA	MALVASÍA BLANCA	BGVCAM0829	W	ESP	x		x	x	x	2471428	0.99	0.44	0.52	0.05
MANSENG VERT	MANSENG VERT	BGVCAM2261	W	FRA	x		x	x	x	1062425	0.97	0.96	0.02	0.02
MARIA ROSA	MARIA ROSA	BGVCAM1327	T	ITA	x		x	x	x	319004	0.93	0.04	0.02	0.94
MAROCAIN NOIR	GROS MAROC	BGVCAM1271	T	FRA	x		x	x	x	460216	0.85	0.50	0.05	0.46
MARSANNE	MARSANNE	BGVCAM1147	W	FRA	x		x	x	x	85745	0.73	0.82	0.15	0.04
MATHIASZ JANOSNE	MUSCAT JULIUS	BGVCAM2831	T/W	HUN		x			x	503289	0.78	0.13	0.02	0.85
MATILDE	MATILDE	BGVCAM2751	T	ITA		x			x	182575	N	0.02	0.02	0.96
MATRASSA	MATRASSA	BGVCAM2631	W	AZE	x		x	x	x	57122	N	0.28	0.54	0.18
MELON	MELAR	BGVCAM1352	W	FRA	x		x	x	x	1428576	0.99	0.96	0.03	0.02
MERLOT	MERLOT	BGVCAM2609	W	FRA	x		x	x	x	137446	0.81	0.74	0.10	0.16
MERSEGUERA (EXQUITSAGOS,VERDOSILLA, SUMOLL BLANCO)	MERSEGUERA	BGVCAM1850	W	ESP		x			x	43184	N	0.04	0.94	0.02
MESLIER SAINT FRANÇOIS	MESLIER SAINT FRANÇOIS	BGVCAM0835	W	FRA	x		x	x	x	383166	0.74	0.91	0.05	0.04
MIGUEL DE ARCO	MIGUEL DE ARCO	BGVCAM2685	W	ESP		x			x	1030290	0.98	0.15	0.81	0.04
MOLINERA	CASTELLANO MORADO	BGVCAM2246	T	ESP	x		x	x	x	1541983	0.99	0.03	0.94	0.03

Mapping collection						Phenotyping			GBS		Population structure			
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
MOLL	PENSAL BLANCO	BGVCAM2682	W	ESP	x		x	x	x	37534	N	0.03	0.93	0.04
MONBADON	MONBADON	BGVCAM2636	W	USA		x			x	3538	N	0.63	0.13	0.24
MONTEPULCIANO	MONTEPULCIANO	BGVCAM2617	W	ITA	x		x	x	x	1077431	0.95	0.09	0.32	0.60
MORENILLO	MORENILLO	BGVCAM1023	W	ESP		x			x	3125464	0.99	0.58	0.34	0.09
MOSCATEL DE HAMBURGO	MOLINERA	BGVCAM2798	T	GBR		x			x	590083	0.88	0.03	0.02	0.96
MOSCATO DI TERRACINA	MOSCATO DI TERRACINA	BGVCAM1357	T	ITA	x		x	x	x	237570	0.81	0.08	0.12	0.80
MOSCATUEL	MOSCATUEL	BGVCAM2702	T	ARG		x			x	876802	0.97	0.03	0.03	0.95
MÜLLER-THURGAU	MÜLLER THURGAU	BGVCAM1365	W	DEU	x		x	x	x	243547	0.91	0.74	0.10	0.16
MUNKATSY JOZSEF	TRIOMPHE DE JEAN MATHIASZ	BGVCAM2817	T	HUN		x			x	341520	0.71	0.07	0.03	0.90
MUSCAT OF ALEXANDRIA	MOSCATEL DE ALEJANDRÍA	BGVCAM1997	T/W	GRC	x		x	x	x	639188	0.99	0.01	0.02	0.97
MUSCAT SAINT LAURENT	MUSCAT SAINT LAURENT	BGVCAM1393	T	FRA	x		x	x	x	40327	0.83	0.44	0.03	0.54
MUSKAT USBEKISTANSKII	MOSCATEL RUSO	BGVCAM2830	T	UZB		x			x	258629	0.62	0.04	0.26	0.70
NAPARO	NÁPARO	BGVCAM1717	T	ESP		x			x	566263	0.91	0.02	0.96	0.02
NEBBIOLO	NEBBIOLO	BGVCAM2618	W	ITA		x			x	898521	0.97	0.42	0.04	0.54
NEDELCESEV MAGVATLANA	ITALIA X SULTANINA V-6	BGVCAM1331	T	BGR	x		x	x	x	1066246	0.98	0.10	0.44	0.46
NEGRA DORADA	NEGRA DORADA	BGVCAM2270	T	ESP	x		x	x	x	694997	0.93	0.05	0.66	0.28
NEGRA TARDÍA	NEGRA TARDÍA	BGVCAM1410	T	ESP	x		x	x	x	611896	0.92	0.05	0.89	0.06
NEGRO AMARO	NEGRO AMARO	BGVCAM2619	W	ITA	x		x	x	x	201786	0.75	0.28	0.26	0.46
NEGRÓN DE ALDÁN (PETIT BOUSCHET)	NEGRÓN DE ALDÁN	BGVCAM1587	W	ESP	x		x	x	x	292539	0.90	0.84	0.10	0.07
NEHELESCOL	NEHELESCOL	BGVCAM1394	T	ISR	x		x	x	x	1607666	0.99	0.04	0.74	0.22
NIELLUCCIO	SANGIOVESE	BGVCAM2622	W	ITA	x		x	x	x	65939	0.85	0.50	0.21	0.29
OJO DE BUEY	OJO DE BUEY	BGVCAM1978	T	ESP		x			x	617854	0.91	0.03	0.93	0.04
ÖKÜZ GÖZÜ	ÖKÜZ GÖZÜ	BGVCAM2639	W	TUR		x			x	10339	N	0.22	0.72	0.06
OLIVETTE NOIRE	OLIVETTA NERA	BGVCAM1358	T	FRA	x		x	x	x	3087543	0.99	0.03	0.93	0.04
OPHTALMO	OPHTALMO	BGVCAM1214	T	CYP	x		x	x	x	539726	0.87	0.02	0.37	0.61
PALOMA	PALOMA	BGVCAM2458	W	ESP	x		x	x	x	1335600	0.98	0.04	0.63	0.33
PALOTE	PALOTE	BGVCAM1036	W	ESP	x		x	x	x	1263941	0.98	0.05	0.88	0.07
PARDUCA	PARDUCA	BGVCAM1186	W	ESP		x			x	20671	N	0.82	0.07	0.11
PARREL	PARREL	BGVCAM2802	W	ESP		x			x	711152	0.94	0.61	0.34	0.06
PASCAL	PASCAL	BGVCAM2459	W	FRA	x		x	x	x	1693636	0.99	0.41	0.10	0.49
PECSI ALDAS	SUPERZIBIBBO	BGVCAM1801	T/W	HUN	x		x	x	x	510171	0.86	0.58	0.04	0.39
PEDRAL (DOZAL)	PEDROL	BGVCAM1156	W	ESP		x			x	685829	0.93	0.91	0.05	0.04
PEDRO XIMENES SEEDLESS	CORINTHE BLANC	BGVCAM1447	T		x		x	x	x	1557725	0.99	0.27	0.60	0.13

Mapping collection							Phenotyping			GBS		Population structure		
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
PERLA NERA	PERLA NERA	BGVCAM0856	T	ITA	x		x	x	x	22141	0.60	0.07	0.04	0.89
PERLETTE	SULINA	BGVCAM2821	T	USA		x			x	348514	0.72	0.02	0.87	0.11
PERRUNO EXTREMEÑO	PERRUNO	BGVCAM2806	W	ESP		x			x	329953	0.69	0.03	0.95	0.02
PETIT VERDOT	PETIT VERDOT	BGVCAM2462	W	FRA	x		x	x	x	925933	0.96	0.95	0.02	0.03
PHRAOULA	HELVANY ROUGE	BGVCAM1496	T	LBN	x		x	x	x	519552	0.88	0.02	0.88	0.10
PINCE'S BLACK	MISTRESS HALL	BGVCAM1498	T	GBR	x		x	x	x	304877	0.91	0.03	0.03	0.94
PINOT	PINOT BLANC	BGVCAM2275	W	FRA	x		x	x	x	44850	0.61	0.96	0.02	0.02
PIROVANO 166A	PIROVANO 166A	BGVCAM2589	T	ITA	x		x	x	x	1221163	0.98	0.03	0.43	0.54
PIROVANO 28	PIROVANO 28	BGVCAM0857	T	ITA	x		x	x	x	831675	0.96	0.05	0.32	0.64
PIROVANO 671	PIROVANO 671	BGVCAM1504	T	ITA	x		x	x	x	49625	0.78	0.04	0.02	0.94
PIROVANO 771	PIROVANO 771	BGVCAM1328	T	ITA	x		x	x	x	3013188	0.99	0.05	0.03	0.92
PIZZUTELLO	PIZZUTELLO NERO	BGVCAM1225	T	ITA	x		x	x	x	1126817	0.98	0.03	0.39	0.58
PIZZUTELLO MOSCATO BIONDO	PIZZUTELLO MOSCATO BIONDO	BGVCAM1329	T	ITA	x		x	x	x	3399174	0.99	0.06	0.38	0.56
PLANTA FINA DE PEDRALBA	PLANTA FINA DE PEDRALBA	BGVCAM1516	W	ESP		x			x	222530	0.54	0.03	0.94	0.03
PLANTA MULA	PLANTA MULA	BGVCAM1158	T	ESP	x		x	x	x	1135337	0.98	0.14	0.59	0.27
PLANTA NOVA	PLANTA NOVA	BGVCAM1047	T	ESP	x		x	x	x	796997	0.93	0.05	0.45	0.51
PLAVAC MALI	PLAVAC MALI	BGVCAM2629	W	YUG		x			x	1007756	0.98	0.08	0.20	0.72
PLAVINA CRNA	PLAVINA	BGVCAM1243	W	YUG		x			x	80302	N	0.26	0.12	0.62
PRECOCE DE MALINGRE	PRECOCE DE MALINGRE	BGVCAM1746	T	FRA	x		x	x	x	470436	0.83	0.90	0.05	0.05
PRIMIERA	PRIMIERA	BGVCAM0859	T	ITA	x		x	x	x	312851	0.81	0.08	0.06	0.86
QUEBRANTA	QUEBRANTA	BGVCAM2641	W	PER		x			x	21977	N	0.07	0.74	0.19
QUEEN	QUEEN	BGVCAM1397	T	USA	x		x	x	x	667415	0.94	0.24	0.58	0.18
QUIEBRATINAJAS	COLGADERO	BGVCAM1888	T	ESP	x		x	x	x	969390	0.98	0.04	0.90	0.05
QUIGAT	FOSTER'S WHITE SEEDLESS	BGVCAM2355	W	ESP	x		x	x	x	443076	0.81	0.43	0.55	0.02
RABIGATO	RABIGATO	BGVCAM2468	W	PRT	x		x	x	x	858173	0.97	0.68	0.10	0.22
RAGOL	RAGOL	BGVCAM1426	T/W	ESP	x		x	x	x	298626	0.93	0.04	0.92	0.04
RASAKI (AFUZ ALI, DATTIER DE BEYROUT, REGINA, ROSETTI)	ROSETI	BGVCAM2163	T	LBN		x			x	95269	0.64	0.03	0.09	0.88
RED OHANES	RED OHANES	BGVCAM2470	T	PRT	x		x	x	x	1527347	0.99	0.09	0.27	0.64
REDONDAL	REDONDAL	BGVCAM0913	W	ESP	x		x	x	x	3233588	0.99	0.92	0.04	0.03
RIESLANER	RIESLANER	BGVCAM1367	W	DEU	x		x	x	x	169463	0.85	0.94	0.04	0.03
RIESLING	RIESLING RHÉNAN	BGVCAM1999	W	DEU	x		x	x	x	56728	0.54	0.85	0.12	0.03
RKATSITELI	RKATSITELI	BGVCAM2630	W	RUS		x			x	2935	N	0.06	0.63	0.31
RODI	RODI	BGVCAM1265	T	ITA	x		x	x	x	3017078	0.99	0.03	0.47	0.50

Mapping collection						Phenotyping			GBS		Population structure			
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
RODITIS	RHODITIS	BGVCAM2885	W	GRC		x			x	75182	N	0.07	0.43	0.51
ROI DES PRECOCES	ROI DES PRECOCES	BGVCAM1506	T	FRA	x		x	x	x	791201	0.94	0.88	0.04	0.08
ROJAL CONQUENSE	ROJAL	BGVCAM1520	W	ESP	x		x	x	x	907336	0.97	0.34	0.52	0.15
ROSAKI NOIR DE SEMIS	ROSAKI NOIR DES SEMIS	BGVCAM1507	T		x		x	x	x	352187	0.73	0.04	0.44	0.52
RUSAKE	RUSAKE	BGVCAM2038	T			x			x	225456	0.61	0.03	0.84	0.14
RUTILIA	RUTILIA	BGVCAM2705	T	ARG		x			x	1573698	0.99	0.23	0.74	0.03
SABRO	SABRO	BGVCAM2234	W	ESP	x		x	x	x	564509	0.90	0.20	0.76	0.04
SALVADOR	COLGADERA	BGVCAM2309	W	ESP		x			x	1478024	0.99	0.02	0.87	0.11
SANTA MAGDALENA	TEMPRANILLA BLANCA	BGVCAM2281	T/W	ESP	x		x	x	x	480636	0.85	0.43	0.44	0.13
SAPERAVI	SAPERAVI	BGVCAM2632	W	GEO	x		x	x	x	1148317	0.97	0.14	0.71	0.15
SAUVIGNON BLANC	GEWÜRZTRAMINER	BGVCAM2596	W	FRA	x		x	x	x	547829	0.89	0.95	0.02	0.02
SAVVATIANO	SAVATIANO	BGVCAM2625	W	GRC	x		x	x	x	27138	0.72	0.02	0.32	0.67
SCHIAVA GROSSA	FRANKENTHAL	BGVCAM1356	T	ITA	x		x	x	x	521726	0.88	0.04	0.03	0.93
SCHIRAS DR. HOUDBINE	SCHIRAZ	BGVCAM1066	T		x		x	x	x	78074	0.86	0.03	0.46	0.52
SELECCION BRUNI 1	SELECCION BRUNI 1	BGVCAM0866	T	ITA	x		x	x	x	364799	0.75	0.69	0.17	0.15
SEMILLON	SEMILLON	BGVCAM1163	W	FRA	x		x	x	x	730537	0.94	0.70	0.13	0.17
SILVANER	SILVANER	BGVCAM1261	W	AUT	x		x	x	x	473616	0.85	0.71	0.02	0.27
SÍRIA (DOÑA BLANCA, BLANCA DE MONTERREI, CIGÜENTE, DONA BRANCA, MALVASÍA CASTELLANA)	DONA BLANCA	BGVCAM1976	T/W	PRT		x			x	172507	N	0.47	0.51	0.02
SOUSÓN	SOUSÓN	BGVCAM1534	W	ESP	x		x	x	x	907454	0.96	0.93	0.04	0.03
SULTANINA	SULTANINA	BGVCAM1077	T	GRC	x		x	x	x	692342	0.93	0.03	0.95	0.02
SYRAH	SYRAH	BGVCAM1078	W	FRA	x		x	x	x	23835	0.74	0.84	0.06	0.10
TALISMANO	TALISMANO	BGVCAM1330	T	ESP	x		x	x	x	316540	0.91	0.03	0.95	0.02
TEBRIZI	SCHIRADZOULI BIANCO	BGVCAM1065	T	AZE	x		x	x	x	1399495	0.98	0.05	0.82	0.13
TEMPRANILLO	TEMPRANILLO	BGVCAM1172	W	ESP	x		x	x	x	12295	0.52	0.12	0.83	0.05
TETA DE VACA (AHMEUR BOU AHMEUR)	AHMEUR BOU AHMEUR	BGVCAM1375	T	DZA	x		x	x	x	249917	0.87	0.05	0.81	0.15
THALLOCSY LAJOS MUSKOTALY	THALLOCSY LAJOS	BGVCAM1361	T	HUN	x		x	x	x	170385	0.73	0.04	0.03	0.94
TINTO DE NAVALCARNERO	TINTO NAVALCARNERO	BGVCAM1555	W	ESP	x		x	x	x	433764	0.84	0.39	0.54	0.07
TOCAI FRIULANO	TOCAI FRIULANO	BGVCAM2623	W	ITA	x		x	x	x	181893	0.71	0.86	0.08	0.07
TOMPA MIHALY	ALFONSO LAVALLÉE X FERRAL	BGVCAM2335	T	ESP	x		x	x	x	420833	0.82	0.02	0.02	0.97

Mapping collection						Phenotyping			GBS		Population structure			
Variety name	Accession name	Accession ID	Usage	Country of origin	Core col.	Supp. set	2010	2012	2013	Number of reads	call rate	q1	q2	q3
TOURIGA NACIONAL	TOURIGO	BGVCAM1209	W	PRT		x			x	71168	N	0.94	0.03	0.04
TREIXADURA	TREIXADURA	BGVCAM1176	W	ESP		x			x	838905	0.97	0.56	0.14	0.30
TRIESTE	TRIESTE	BGVCAM1755	T	ITA	x		x	x	x	193980	0.47	0.03	0.04	0.93
TROBAT	TROBAT	BGVCAM1089	W	ESP	x		x	x	x	1004242	0.98	0.54	0.30	0.16
TSCHETIRBAK	TSCHETIRBAK	BGVCAM2283	W	UZB	x		x	x	x	685409	0.94	0.02	0.88	0.09
TURKI	TURKI	BGVCAM1511	T	TUN	x		x	x	x	3257366	0.99	0.09	0.08	0.83
UVA DE OLAZ (JERÓNIMO)	JERÓNIMO DE TUDELA	BGVCAM0825	T	ESP	x		x	x	x	858734	0.96	0.02	0.95	0.02
VELTLINER FRUEHROT	CAILLABA	BGVCAM0884	W	ITA	x		x	x	x	622622	0.54	0.43	0.07	0.50
VELTLINER GRUEN	VELTLINER GRUEN	BGVCAM2597	W	ITA	x		x	x	x	435932	0.83	0.68	0.09	0.23
VERDEJO	VERDEJO	BGVCAM1768	W	ESP		x			x	201055	0.51	0.85	0.13	0.03
VERDEJO DE SALAMANCA	VERDEJO	BGVCAM1916	W	ESP		x			x	154494	N	0.12	0.79	0.09
VÉRICO	VÉRICO	BGVCAM1215	T	CYP	x		x	x	x	628960	0.92	0.03	0.79	0.19
VERMANSENG VERT	VERMANSENG VERT	BGVCAM2284	W		x		x	x	x	30544	0.75	0.71	0.06	0.23
VERNACCIA DI SAN GIMIGNANO	VERNACCIA	BGVCAM2774	W	ITA	x		x	x	x	4548891	0.99	0.22	0.39	0.39
VINYATER	VINATE	BGVCAM2327	W	ESP	x		x	x	x	615984	0.92	0.03	0.95	0.02
VIOLA	DALMASSO XVIII-03	BGVCAM1440	T	ITA	x		x	x	x	3559748	0.99	0.12	0.13	0.75
VIVONA 378	VIVONA 378	BGVCAM1437	T	ITA	x		x	x	x	117711	0.71	0.04	0.05	0.91
XYNOMAVRO (XINOGALTSO, POPOLKA)	XYNOMAVRO	BGVCAM2626	W	GRC		x			x	3060289	0.99	0.76	0.15	0.10
ZEINI ABIAD	ZEINI ABIAD	BGVCAM1513	T	LBN	x		x	x	x	113309	0.86	0.05	0.88	0.08
ZILAVKA	ZILAVKA	BGVCAM1252	W	YUG		x			x	428143	0.82	0.05	0.05	0.90
ZWEIGELTREBE BLAU	ZWEIGELTREBE BLAU	BGVCAM2601	W	DEU		x			x	51536	N	0.89	0.07	0.04

Variety name is the curated name of the accession in the El Encin germplasm bank based on genotypic, phenotypic and bibliographic data; synonyms of some varieties are given between parentheses. **Accession name** is the introduction name (as found in the passport data). **Usage**: T: table, W: making, T/W: mixt use; **Country of origin**: abbreviations of country names are given according to ISO 3166-1 alpha-3 code (<https://www.iso.org/obp/ui/#search>); **Core col.**: core collection; **Supp. Set**: supplementary set; **Number of reads** from a single Illumina flow cell channel; **Call rate** after filtering (from DS5 in Fig 3.); **Population structure**: Probabilities of assignment to sub-populations at K = 3.

Table A.2. List of trio pedigrees included in the mapping collection

Parent 1	Parent 2	Progeny
AGUDELO (CHENIN BLANC)	HEUNISCH WEISS	MESLIER SAINT FRANÇOIS
ALBILLO MAYOR	HEBÉN	VINYATER
BRUNI 415	MUSCAT OF ALEXANDRIA	VIVONA 378
CABERNET FRANC	SAUVIGNON BLANC	CABERNET SAUVIGNON
DABOUKI	CHASSELAS VIOLET	DANLAS
GARNACHA	NEGRÓN DE ALDÁN (PETIT BOUSCHET)	GARNACHA TINTORERA (ALICANTE BOUSCHET)
HEBÉN	MALVASÍA AROMÁTICA (MALVASÍA DE SITGES)	MALMSEY
HEUNISCH WEISS	PINOT	ALIGOTE
		CHARDONNAY
		GAMAY
		MELON
MAROCAIN NOIR	MUSCAT OF ALEXANDRIA	LADY DOWNE'S SEEDLING
		PINCE'S BLACK
RIESLING	SILVANER	RIESLANER

Cultivars are identified via "variety name"

Table A.3. Mendelian inheritance checking of SNP alleles in trio pedigrees

Parent 1	Parent 2	Progeny	SNPs without missing values	Number of genotypic values ^a	Number of Mendelian errors	Error rate (%) ^b
AGUDELO	HEUNISCH WEISS	MESLIER SAINT FRANÇOIS	16,084	48,252	170	0.35
BRUNI 415	MUSCAT OF ALEXANDRIA	VIVONA 378	20,343	61,029	167	0.27
CABERNET FRANC	SAUVIGNON BLANC	CABERNET SAUVIGNON	28,821	86,463	315	0.36
DABOUKI	CHASSELAS VIOLET	DANLAS	23,834	71,502	244	0.34
GARNACHA	NEGRON DE ALDAN	GARNACHA TINTORERA	24,829	74,487	250	0.34
HEUNISCH WEISS	PINOT	ALIGOTE	21,203	63,609	222	0.35
HEUNISCH WEISS	PINOT	CHARDONNAY	21,207	63,621	211	0.33
MAROCAIN NOIR	MUSCAT OF ALEXANDRIA	LADY DOWNE'S SEEDLING	30,343	91,029	439	0.48
MAROCAIN NOIR	MUSCAT OF ALEXANDRIA	PINCE'S BLACK	28,299	84,897	265	0.31
RIESLING	SILVANER	RIESLANER	15,922	47,766	183	0.38
Mean						0.35

^a: Number of SNPs without missing values in the pedigree multiplied by the number of pedigree members (three).

^b: Assuming that each Mendelian error detected in a pedigree is the consequence of only one genotype calling error (since the occurrence of two genotyping errors in the same pedigree should be negligible), the SNP calling error rate was estimated as the number of detected Mendelian errors over the number of individual genotypic values.

Table A.4. Chromosome-wise summary statistics of the filtered SNP dataset DS5

Chr.	Chr. size (Mb)	Number of SNPs	Average SNP call rate (%)	Position of the first SNP	Position of the last SNP	Average inter-SNPs distance (Kb)	Average MAF ^a	Average H _{obs} ^b	Average H _{exp} ^c
1	24.23	1396	88.31	0.01	24.20	17.36	0.22	0.32	0.30
2	18.89	855	88.27	0.10	18.86	22.10	0.21	0.32	0.29
3	20.70	1159	88.10	0.04	20.67	17.86	0.21	0.32	0.30
4	24.71	1314	88.02	0.06	24.69	18.81	0.21	0.32	0.30
5	25.65	1764	87.98	0.02	25.58	14.54	0.20	0.31	0.29
6	22.65	1318	88.29	0.10	22.55	17.18	0.22	0.33	0.31
7	27.36	1545	88.49	0.09	27.35	17.71	0.22	0.32	0.30
8	22.55	1415	88.33	0.05	22.46	15.94	0.22	0.34	0.31
9	23.01	1161	88.01	0.06	23.00	19.82	0.22	0.34	0.30
10	23.50	1084	88.00	0.04	23.48	21.68	0.23	0.33	0.32
11	20.12	1099	87.99	0.03	20.08	18.31	0.21	0.32	0.30
12	24.27	1390	88.13	0.00	24.24	17.46	0.21	0.31	0.29
13	29.08	1613	88.05	0.12	29.06	18.03	0.21	0.32	0.30
14	30.27	1781	88.03	0.03	30.26	17.00	0.21	0.32	0.30
15	20.30	1112	87.48	0.10	20.27	18.26	0.21	0.32	0.30
16	23.57	1075	88.19	0.07	23.50	21.93	0.22	0.33	0.30
17	18.69	970	88.39	0.00	18.09	19.27	0.22	0.32	0.30
18	34.57	2136	88.34	0.02	34.55	16.18	0.20	0.31	0.29
19	24.70	1341	88.12	0.02	24.63	18.42	0.21	0.32	0.30

Chromosome sizes are given according to the 12x.2 version of the reference genome sequence; MAF: minor allele frequency, H_{obs}: observed heterozygosity, H_{exp}: average expected heterozygosity. ^a, ^b and ^c were calculated using the software HaploView 4.2.

Figure A.1. Relationship between SNP and gene density distributions in the grapevine genome

The SNP dataset presented here was called from only one sequencing lane by alignment to the 12x.0 genome sequence, to which corresponds also the gene annotation data

(<http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis>); gene and SNP densities are presented as their number in chromosomal bins of 500 kb.

Table A.5. Main features of the developed SSR multiplex PCRs (see the key at the bottom of the table)

Identifier	Cycling program	Markers /PCR	Marker name	Control loci	Chrom.	Position (bp)	Reference	Fluorescent dye	Primers concent. (μM)
Mx01	Tp-A	15	VVIP60		1	9,196,429	12	PET	0.22
			VVIB01		2	2,161,587	12	NED	0.07
			VRZAG83		4	20,627,004	16	PET	0.07
			VRZAG79		5	5,692,460	16	PET	0.17
			VVMD7		7	1,560,528	3	NED	0.10
			VRZAG62		7	2,148,817	16	VIC	0.10
			VMC1B11		8	14,039,245	18	FAM	0.28
			VVIQ52		9	21,558,043	12	PET	0.12
			VVMD25		11	2,971,693	4	FAM	0.18
			VVS2		11	3,909,894	17	FAM	0.22
			VVIH54		13	3,434,661	12	VIC	0.17
			VRZAG112		14	22,042,800	16	VIC	0.10
			VVMD24		14	24,251,864	4	PET	0.18
			VVIN73		17	5,636,037	12	VIC	0.13
			VVIP31		19	6,697,524	12	NED	0.12
Mx02	Tp-B	11	VRZAG29		1	5,286,106	16	VIC	0.06
			VVMD28		3	12,005,376	4	NED	0.20
			VVMD32		4	18,402,264	4	VIC	0.23
			VVMD27		5	4,472,022	4	PET	0.10
			VVMD21		6	14,883,818	4	PET	0.10
			VRZAG67		10	5,006,368	16	PET	0.08
			VVIV37		10	14,590,969	12	VIC	0.15
			VMC4F3-1		12	16,417,442	7	NED	0.08
			VVIV67		15	10,897,392	12	PET	0.32
			VVMD5		16	23,305,077	3	FAM	0.35
Mx03	Tp-A	8	VVIN16		18	26,361,305	12	FAM	0.07
			VMC4F8		1	632,258	13	NED	0.07
			VMC5G7		2	8,334,940	18	VIC	0.10
			VVMD27	x	5	4,472,022	4	PET	0.10
			VVIM63		13	4,909,035	12	PET	0.40
			VMC3E5		18	321,045	18	VIC	0.15
			VMC2B1.1		18	3,752,073	13	FAM	0.07
			UDV-005		18	11,025,140	8	FAM	0.10
Mx04	Tp-A	10	VMC7f2-2		18	30,313,811	14	PET	0.10
			VVIQ57		1	6,081,664	12	VIC	0.07
			VVIN61		1	10,569,689	12	NED	0.20
			VChr1b		1	11,850,801	5	NED	0.07
			VChr5c		5	4,080,604	5	VIC	0.07
			VVIT68		5	7,099,947	12	VIC	0.10
			UDV-053		5	8,541,098	8	PET	0.10
			VMC9B5		5	20,174,295	9	PET	0.10
			VVIN40		5	24,429,944	12	FAM	0.10
			UDV-117		18	13,168,914	8	NED	0.10
Mx05	Tp-B	8	VVIN16	x	18	26,361,305	12	FAM	0.10
			VMCNG2g7		1	6,993,233	13	PET	0.15
			UDV-109		2	132,291	8	NED	0.15
			VVC6		5	1,470,620	6	FAM	0.15
			VRZAG79	x	5	5,692,460	16	PET	0.10
			VVIN33		5	11,721,184	12	FAM	0.15
			VMC4C6		5	25,493,869	7	VIC	0.07
			VMC1E8		8	17,261,639	18	VIC	0.15
Mx06	Tp-A	8	VVIP05		14	2,861,324	12	NED	0.10
			VVIB01	x	2	2,161,587	12	NED	0.10
			VMCNG1D3		2	18,860,462	13	NED	0.07

Identifier	Cycling program	Markers /PCR	Marker name	Control loci	Chrom.	Position (bp)	Reference	Fluorescent dye	Primers concent. (μM)
			VChr5b		5	9,474,726	5	VIC	0.05
			VMC3C7		5	12,716,770	18	FAM	0.30
			UDV-077		5	21,538,151	8	PET	0.30
			VVIN62		13	1,387,838	12	VIC	0.15
			VVIV08		17	12,161,420	12	PET	0.30
			VVIQ44		18	5,309,353	12	FAM	0.15
Mx07	Tp-A	9	VVIP12		1	4,502,036	12	FAM	0.08
			VMC9D3		1	24,039,899	9	VIC	0.06
			VChr2c		2	7,998,935	5	PET	0.12
			VMC7g3		2	18,382,886	14	VIC	0.08
			VMCNG2E1		4	5,232,859	13	NED	0.06
			VVMD21	x	6	14,883,818	4	PET	0.12
			VVS16		15	9,570,305	14	NED	0.45
			VMC2A3		18	948,244	18	FAM	0.08
			VVIN74		19	213,243	12	VIC	0.08
Mx08	Tp-B	8	VChr1a		1	2,819,297	5	PET	0.30
			VMC2C10-1		2	6,975,274	18	NED	0.10
			VV6718		2	16,446,991	13	VIC	0.07
			VChr9a		9	6,426,756	5	PET	0.15
			VVS2	x	11	3,909,894	17	FAM	0.12
			VVIP02		11	11,931,480	12	FAM	0.10
			VMC4D9-2		15	16,581,058	13	NED	0.20
			VMC8G3		15	17,666,601	1	VIC	0.07
Mx09	Tp-A	7	VRZAG29	x	1	5,286,106	16	VIC	0.04
			VMC8F10		3	3,324,131	18	PET	0.10
			VMC2F10		6	7,615,634	18	PET	0.10
			VChr8a		8	15,200,499	5	NED	0.10
			VMC9H4_2		13	18,412,652	9	VIC	0.15
			UDV-046		19	7,755,770	8	FAM	0.10
			VChr19a		19	977,575	5	FAM	0.10
Mx10	Tp-A	7	VMC2B3		1	19,375,299	18	FAM	0.07
			Vchr4a		4	946,454	5	NED	0.07
			VVIM43		6	20,141,244	12	VIC	0.10
			VMC1F10		8	292,910	18	VIC	0.07
			VChr13c		13	4,525,996	5	NED	0.07
			VChr13b		13	24776322	5	PET	0.10
			VVMD5	x	16	23,305,077	3	FAM	0.25
Mx11	Tp-B	9	VVIN54		3	11,093,607	12	NED	0.15
			VRZAG21		4	13,794,845	16	NED	0.07
			VVIQ67		7	22,899,285	12	FAM	0.12
			VVIV37	x	10	14,590,969	12	VIC	0.10
			GT183H07		13	83,697	13	PET	0.15
			VMCNG1E1		14	1,413,666	18	PET	0.07
			VVIS70		14	27,477,542	12	FAM	0.10
			VVIQ61		15	11,761,810	12	FAM	0.15
			VMC7B1		19	23,628,578	14	VIC	0.10
Mx12	Tp-A	9	VMC4D4		4	3,753,977	7	PET	0.07
			VV5799		6	11,606,406	13	NED	0.10
			VVMD21	x	6	14,883,818	4	PET	0.15
			VVIP33		15	13,060,296	12	PET	0.22
			VChr16a		16	3,187,867	5	NED	0.10
			VMC4B7-2		16	23,329,085	18	FAM	0.20
			VMC2H3		17	3,684,833	18	VIC	0.07
			VChr18c		18	33,971,369	5	FAM	0.05
			VMC5E9		19	4,183,125	18	VIC	0.10
Mx13	Tp-B	8	VVIN75		4	14,969,759	12	VIC	0.20
			VChr7a		7	23,070,829	5	FAM	0.05
			VVIV04		7	23,860,671	12	FAM	0.20

Identifier	Cycling program	Markers /PCR	Marker name	Control loci	Chrom.	Position (bp)	Reference	Fluorescent dye	Primers concent. (μM)
			VChr8b		8	4,343,076	5	NED	0.30
			VMC3G7		8	8,553,958	18	PET	0.15
			VMC1B11	x	8	14,039,245	18	FAM	0.25
			VMC2H10		8	20,816,914	18	VIC	0.06
			VMC2H4		12	5,447,330	18	PET	0.17
Mx14	Tp-B	6	VMC3B10		2	5,000,200	18	NED	0.13
			UDV-093		3	1,335,056	8	VIC	0.25
			VRZAG83	x	4	20,627,004	16	PET	0.08
			VMC5G1-1		6	17,486,959	18	FAM	0.25
			VVIS58		7	5,335,402	12	NED	0.10
			VVIM42A		15	15,943,609	12	VIC	0.15
Mx15	Tp-B	8	UDV-022		3	7,546,783	8	VIC	0.10
			VChr3a		3	9,542,014	5	PET	0.10
			VMC8B11		4	17,682,875	13	VIC	0.30
			VMC16D4		5	17,606,754	18	FAM	0.10
			VVMD7	x	7	1,560,528	3	NED	0.10
			VVIN70		14	29,853,616	12	VIC	0.12
			VMC8B5		18	3,362,208	1	NED	0.10
			VChr18b		18	14,028,762	5	PET	0.08
Mx16	Tp-A	7	VVIQ66		9	10,808,981	12	NED	0.12
			UDV-040		9	20,091,003	8	NED	0.10
			VVIH01		10	1,473,028	12	FAM	0.12
			VVIV67	x	15	10,897,392	12	PET	0.12
			VVIM93		18	6,962,288	12	VIC	0.12
			VMC8F4.2		18	10,120,948	1	PET	0.08
			VChr18a		18	8,512,593	5	FAM	0.10
Mx17	Tp-B	7	VMC9F2		1	20,980,935	9	PET	0.15
			VChr5a		5	14,448,749	5	FAM	0.12
			VVIN31		6	18,937,746	12	VIC	0.10
			VVIB66		8	18,270,246	12	FAM	0.10
			VMC1E12		14	7,136,672	18	NED	0.10
			UDV-062		19	21,598,416	8	PET	0.10
			VVMD32	x	4	18,402,264	4	VIC	0.10
Mx18	Tp-A	10	VMC1G7		3	8,547,609	18	NED	0.12
			VVIB22		7	3,521,457	12	NED	0.12
			VMC2E11		9	22,676,139	18	VIC	0.10
			VMC2D9		9	18,966,688	18	PET	0.10
			VVMD25	x	11	2,971,693	4	FAM	0.10
			UDV-100		11	7,911,371	8	VIC	0.12
			VMC3D12		13	8,193,616	18	PET	0.10
			UDV-020		13	19,113,495	8	FAM	0.10
			VVIP22		14	18,364,396	12	VIC	0.10
			VVIM03		19	7,078,556	12	NED	0.12
Mx19	Tp-A	10	VVIP60	x	1	9,196,429	12	PET	0.10
			VVIP77		4	19,684,609	12	NED	0.10
			VChr7c		7	545,849	5	PET	0.10
			VMC5H2		8	15,407,796	13	VIC	0.10
			VMC3E12		11	774,942	18	FAM	0.10
			VVIB19		11	5,865,805	12	NED	0.10
			VVIM11		12	11,103,248	12	VIC	0.10
			VVIN52		16	619,678	12	VIC	0.10
			B003		19	5,927,868	13	PET	0.10
			VVIV33		19	14,690,189	12	FAM	0.10
Mx20	Tp-A	8	VVIO61		1	22,221,400	12	FAM	0.10
			VVIP72		6	2,228,161	12	NED	0.10
			VVIN56		7	25,737,398	12	NED	0.10
			VMC7h2		8	10,149,243	14	FAM	0.10
			VMC1C10		9	563,536	18	VIC	0.10

Identifier	Cycling program	Markers /PCR	Marker name	Control loci	Chrom.	Position (bp)	Reference	Fluorescent dye	Primers concent. (μM)
			VChr16b		16	17,432,683	5	PET	0.10
			VVIN73	x	17	5,636,037	12	VIC	0.10
			VVIV22		17	7,660,564	12	PET	0.10
Mx21	Tp-A	8	VMC1A5		3	5,954,608	18	VIC	0.30
			VVMD28	x	3	12,005,376	4	NED	0.12
			VMC1A12		7	22,589,274	18	FAM	0.15
			VMCNG1E4-2		9	5,448,330	13	NED	0.10
			VMC2A10		10	5,982,405	18	VIC	0.08
			UDV-124		13	11,412,148	8	PET	0.12
			VMC8E6		13	27,180,657	1	FAM	0.10
			VMC2H12		14	10,303,957	18	PET	0.10
Mx22	Tp-A	5	VVIB72		5	2,848,403	12	NED	0.10
			VRZAG62	x	7	2,148,817	16	VIC	0.10
			UDV-059		10	9,016,614	8	FAM	0.10
			VMC3C11-1		17	2,296,673	18	PET	0.10
			VRZAG15		17	6,588,595	16	PET	0.10
Mx23	Tp-B	7	VVIB01	x	2	2,161,587	12	NED	0.06
			VMC8C2		2	17,626,961	1	FAM	0.10
			UDV-063		10	15,454,368	8	PET	0.18
			VMC2A12		11	1,542,611	18	NED	0.45
			VMC6G1		11	8,983,138	18	VIC	0.08
			VMC8G9		12	21,945,210	18	PET	0.30
			VVIB54		19	20,755,629	12	NED	0.05
Mx24	Tp-B	8	VVIO02		1	9,813,247	12	FAM	0.05
			VMC6g10		4	23,949,251	13	NED	0.20
			VVIQ52	x	9	21,558,043	12	PET	0.07
			VVIM01		13	6,609,298	12	PET	0.07
			UDV-095		14	26,102,129	8	VIC	0.20
			VVIN94		14	28,772,134	12	NED	0.10
			VMC5G8		15	14,895,031	18	FAM	0.12
			VChr17a		17	9,315,344	5	VIC	0.05
Mx25	Tp-A	7	VVMD27	x	5	4,472,022	4	PET	0.10
			UDV-090		6	1,012,094	8	PET	0.10
			UDV-085		6	4,906,382	8	NED	0.15
			UDV-038		13	9,494,599	8	VIC	0.10
			VMC4G6		6	16,304,885	7	FAM	0.10
			VMC8D11		7	18964323	18	VIC	0.10
			VMC2B11		14	20,429,749	18	NED	0.10
Mx26	Tp-B	6	VMC7h3		4	4,719,044	14	FAM	0.10
			VVIN78		10	10,476,651	12	NED	0.10
			VVIP10		13	26,977,865	12	PET	0.15
			UDV-050		14	4,943,488	8	PET	0.10
			VVMD24	x	14	24,251,864	4	PET	0.10
			UDV-032		16	2,288,155	8	VIC	0.08
Mx27	Tp-A	6	VVIF52		1	23,292,680	12	PET	0.10
			VMCNG2H2.2		8	7,184,551	13	PET	0.20
			UDV-026		8	21,473,580	8	NED	0.12
			VVMD25	x	11	2,971,693	4	FAM	0.10
			UDV-058		12	19,001,459	8	VIC	0.08
			UDV-116		15	7,789,873	8	FAM	0.20
Mx28	Tp-A	7	VMC2H9		6	5,819,480	18	NED	0.08
			VMC7A4		7	7,936,887	14	PET	0.15
			UDV-075		8	2,400,696	8	VIC	0.08
			VChr12a		12	6,174,660	5	FAM	0.12
			VMC4F3-1	x	12	16,417,442	7	NED	0.15
			VVC34		14	13,207,677	6	PET	0.15
			A105		18	28,306,457	13	VIC	0.20
Mx29	Tp-B	6	VMC6d12		9	3,792,236	2	FAM	0.05

Identifier	Cycling program	Markers /PCR	Marker name	Control loci	Chrom.	Position (bp)	Reference	Fluorescent dye	Primers concent. (µM)
			VMC5C1		9	9,396,293	13	NED	0.08
			VChr11b		11	17,380,059	5	PET	0.08
			VMC8G6		12	2,354,142	13	VIC	0.05
			TT251F02		12	418,480	13	PET	0.45
			VVIV67	x	15	10,897,392	12	PET	0.15
Mx30	Tp-A	8	FAM140		2	9,166,714	10	VIC	0.10
			FAM110		6	3,652,483	10	FAM	0.08
			FAM96		6	21,584,205	10	NED	0.08
			VVIQ52	x	9	21,558,043	12	PET	0.10
			FAM63		10	23,407,091	10	VIC	0.10
			VMC1G3-2		12	9,589,001	18	FAM	0.15
			VChr15a		15	19,977,749	5	NED	0.10
			UDV-104		16	10,785,478	8	PET	0.30
Mx31	Tp-A	6	FAM141		1	17,732,854	10	VIC	0.05
			FAM91		5	782,861	10	FAM	0.10
			FAM12		5	2,783,225	10	FAM	0.10
			FAM45		10	11,414,249	10	NED	0.10
			VRZAG112	x	14	22,042,800	16	VIC	0.10
			FAM18		16	19,490,645	10	PET	0.06
Mx32	Tp-D	8	FAM68		2	13,188,507	10	PET	0.10
			VVIQ25		4	2,036,638	12	PET	0.20
			FAM122		5	15,723,310	10	FAM	0.10
			FAM116		7	20,388,473	10	VIC	0.05
			FAM70		16	21,189,258	10	NED	0.10
			VChr17b		17	5,017,367	5	VIC	0.05
			VVIP31	x	19	6,697,524	12	NED	0.10
			VChr19b		19	8,693,674	5	FAM	0.06
Mx33	Tp-A	9	FAM24		2	13,508,269	10	NED	0.15
			VVIU20		2	16,650,964	12	NED	0.15
			VRZAG89		5	22,919,299	16	FAM	0.06
			FAM42		9	7,551,193	10	VIC	0.10
			VVIQ52	x	9	21,558,043	12	PET	0.10
			VVMD8		11	19,974,886	3	VIC	0.45
			FAM135		14	8,755,200	10	PET	0.10
			p3_VvAGL11		18	30,313,541	11	NED	0.10
			UDV-023		19	2,346,454	8	PET	0.12
Mx34	Tp-D	4	UDV-110		4	11,543,793	8	FAM	0.50
			VVIC50		6	8,450,955	12	NED	0.10
			VVIV37	x	10	14,590,969	12	VIC	0.05
			FAM92		16	7,976,858	10	PET	0.25
Mx35	Tp-B	5	VChr1a-2		1	2,819,297	15	PET	0.26
			FAM140-2		2	9,166,787	15	VIC	0.07
			VVMD25	x	11	2,971,693	4	FAM	0.10
			TT251F02-2		12	418,506	15	FAM	0.12
			VChr15a-2		15	19,977,651	15	NED	0.16
Mx36	Tp-B	5	VVIV67	x	15	10,897,392	12	PET	0.20
			VMC8G3-2		15	17,666,601	15	VIC	0.12
			UDV-009		16	9,642,339	8	VIC	0.08
			VChr18a-2		18	8,512,550	15	FAM	0.12
			UDV-062-2		19	21,594,816	15	PET	0.10
Mx37	Tp-B	5	VRZAG29	x	1	5,286,106	16	VIC	0.08
			VMC2D9-2		9	18,966,688	15	PET	0.15
			VChr17b-2		17	5,017,280	15	VIC	0.08
			VChr19a-2		19	977,575	15	FAM	0.12
			VVIM03-2		19	7,078,556	15	NED	0.20
Mx38	Tp-A	4	VMC16D4-2		5	17,606,676	15	FAM	0.10
			VVIP72-2		6	2,280,007	15	NED	0.10
			VRZAG62	x	7	2,148,817	16	VIC	0.07

Identifier	Cycling program	Markers /PCR	Marker name	Control loci	Chrom.	Position (bp)	Reference	Fluorescent dye	Primers concent. (µM)
Mx39	Tp-A	5	VVIS58-2		7	5,335,417	15	NED	0.10
			VVS3-2		2	4,014,388	15	PET	0.14
			VVIN75-2		4	14,969,759	15	VIC	0.12
			VChr8a-2		8	15,200,499	15	NED	0.12
			VVIH54	x	13	3,434,661	12	VIC	0.12
			FAM70-2		16	21,189,056	15	FAM	0.10
Mx40	Tp-C	5	VVIQ25-2		4	2,036,638	15	PET	0.15
			VVIB22-2		7	3,521,457	15	NED	0.20
			VVIV37	x	10	14,590,969	12	VIC	0.08
			VMC4H2		12	4,073,541	13	FAM	0.35
			VVIS70-2		14	27,477,542	15	FAM	0.07
Mx41	Tp-A	9	VVIM25		1	12,617,249	12	NED	0.12
			RTAChr3-17a		3	18,564,018	19	PET	0.08
			RTAChr3-19b		3	20,400,237	19	VIC	0.20
			RTAChr4-9a		4	9,511,967	19	PET	0.20
			RTAChr13-20a		13	20,382,252	19	VIC	0.10
			RTAChr15-4a		15	4019057	19	FAM	0.30
			VVIV67	x	15	10,897,392	12	PET	0.20
			VVIP08		18	13,796,392	12	NED	0.08
			RTAChr19-11b		19	11,230,059	19	FAM	0.12
Mx42	Tp-C	8	RTAChr1-14a		1	13,253,163	19	NED	0.45
			RTAChr3-15a		3	15,937,188	19	FAM	0.50
			FAM65		9	1,202,649	10	FAM	0.06
			VVIQ52	x	9	21,558,043	12	PET	0.08
			FAM62-2		13	16,059,252	15	FAM	0.06
			RTAChr15-3b		15	2650923	19	PET	0.20
			RTAChr16-6b		16	6,021,381	19	VIC	0.06
			RTAChr17-15a		17	16,316,158	19	VIC	0.12
Mx43	Tp-A	8	RTAChr2-11a		2	10,707,832	19	FAM	0.15
			UDV-021-2		3	5,287,429	15	NED	0.12
			VRZAG67	x	10	5,006,368	16	PET	0.08
			RTAChr13-13b		13	13,446,986	19	VIC	0.15
			RTAChr15-1b		15	1167751	19	VIC	0.08
			RTAChr15-7a		15	8,439,107	19	FAM	0.20
			VMC1E11-2		16	15,219,141	15	PET	0.10
			RTAChr19-16b		19	18,589,198	19	VIC	0.08
Mx44	Tp-A	6	VVIV36		7	8,779,579	12	NED	0.08
			RTAChr9-13a		9	13,102,063	19	NED	0.08
			VRZAG67	x	10	5,006,368	16	PET	0.10
			RTAChr13-16a		13	16,119,425	19	VIC	0.12
			RTAChr17-15b		17	16,474,837	19	FAM	0.20
			UDV-134-2		18	18,100,286	15	PET	0.15
Mx45	Tp-C	5	VVMD27	x	5	4,472,022	4	PET	0.08
			RTAChr7-9		7	9,947,798	19	NED	0.12
			FAM104		9	18,036,034	10	FAM	0.60
			RTAChr11-11b		11	11,365,896	19	VIC	0.30
			RTAChr13-23b		13	27,762,807	19	FAM	0.08
Discarded			UDV-021		3	5,287,429	8		
Discarded			UDV-128		7	6,208,791	8		
Discarded			FAM62		13	16,059,563	10		
Discarded			VMC1E11		16	15,219,204	13		
Discarded			UDV-134		18	18,100,286	8		
Discarded			VVS3		2	4,014,310	17		

New SSRs developed in this study are identified with the number 19 in the “Reference” column. SSRs for which one or both primers were redesigned (indicated by the number 15 in the “Reference” column) were named adding “-2” at the end of their original names; their names are highlighted in green, and names of the original markers are highlighted in red. Physical positions are given according to the 12x.2 version of the grapevine genome (Canaguier

et al., 2014; <https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences>). Primer sequences of all the markers can be found in “Online resource 1” in Zarouri et al., (2015).

List of references: **1** (Adam-Blondon et al., 2004), **2** (Arroyo-García and Martínez-Zapater, 2004), **3** (Bowers et al., 1996), **4** (Bowers et al., 1999), **5** (Cipriani et al., 2008), **6** (Decroocq et al., 2003), **7** (Di Gaspero et al., 2000), **8** (Di Gaspero et al., 2005), **9** (Greek Vitis Database: <http://gvd.biology.uoc.gr/gvd/>), **10** (Huang et al., 2011), **11** (Mejía et al., 2011), **12** (Merdinoglu et al., 2005), **13** (NCBI: <http://www.ncbi.nlm.nih.gov/probe/>), **14** (Pellerone et al., 2001), **15** “Redesigned in this study”, **16** (Sefc et al., 1999), **17** (Thomas and Scott, 1993), **18** “Vitis Microsatellite Consortium at (NCBI <http://www.ncbi.nlm.nih.gov/probe/>)”, **19** (Zarouri et al., 2015).

Table A.6. Thermocycling programs optimized to run the designed multiplex PCRs

Identifier	Cycles	Temperature	Time
Tp-A (Ibanez et al., 2009, with modifications)	1	94 °C	15 min.
	8	94 °C	1 min.
		60 °C	90 s.
		72 °C	2 min.
	10	94 °C	1 min.
		60°C - (0,3°C/cycle)	90 s.
Tp-B (Ibanez et al., 2009, with modifications)	20	72 °C	2 min.
		94 °C	1 min.
		57°C- (0,2°C/ cycle)	90 s.
	1	72 °C	2 min.
		94 °C	1 min.
		72 °C	90 min.
Tp-C	1	94 °C	15 min.
	5	94 °C	1 min.
		58 °C	90 s.
		72 °C	2 min.
	5	94 °C	1 min.
		58°C - (0,4°C/cycle)	90 s.
Tp-D	25	72 °C	2 min.
		94 °C	1 min.
		56°C-(0,2°C/ cycle)	90 s.
	1	72 °C	2 min.
		94 °C	1 min.
		72 °C	90 min.
Tp-D	1	94 °C	15 min.
	10	94 °C	1 min.
		52 °C	90 s.
		72 °C	2 min.
	30	94 °C	1 min.
		52°C - (0,2°C/cycle)	90 s.
Tp-D	1	72 °C	2 min.
		94 °C	1 min.
		72 °C	90 min.
	1	94 °C	1 min.
		50 °C	90 s.
		72 °C	2 min.
Tp-D	1	72 °C	2 min.
		94 °C	1 min.
		72 °C	90 min.
	1	94 °C	1 min.
		50 °C	90 s.
		72 °C	2 min.
Tp-D	1	72 °C	2 min.
		94 °C	1 min.
		72 °C	90 min.
	1	94 °C	1 min.
		50 °C	90 s.
		72 °C	2 min.
Tp-D	1	72 °C	2 min.
		94 °C	1 min.
		72 °C	90 min.
	1	94 °C	1 min.
		50 °C	90 s.
		72 °C	2 min.

Table A.7. Quality of amplification profiles, data checking and diversity parameters of the analyzed SSR markers and their utilization in the study (see the key at the bottom of the table)

Marker			Quality of amplification profiles							Checking		Genetic diversity								Use			
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PI _C	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis
VMC4F8	1	632,258	2	no	no	no	1	1	1	1		207	12	112-141	0.787	0.86	0.844	0.034	0.047	C		x	x
VChr1a	1	2,819,297	1	no	no	no	1	1	1	2	1	207	13	172-244	0.763	0.742	0.71	0.097	-0.011				
VChr1a-2	1	2,819,297	1	no	no	no	1	1	1	1		207	11	145-219	0.725	0.735	0.701	0.103	0.012	C		x	x
VVIP12	1	4,502,036	1	no	no	no	1	1	1	1		207	2	284-288	0.357	0.404	0.322	0.437	0.06	C+S		x	x
VRZAG29	1	5,286,106	2	no	no	no	1	1	1	1		207	5	105-119	0.338	0.359	0.331	0.439	0.035	C+S		x	x
VVIQ57	1	6,081,664	2	no	no	no	1	1	1	1		207	5	167-174	0.464	0.449	0.412	0.34	-0.016	C		x	x
VMCNG2g7	1	6,993,233	2	no	no	no	1	1	1	1		207	13	102-150	0.826	0.802	0.782	0.058	-0.017	C		x	x
VVIP60	1	9,196,429	2	no	no	no	2	1	2	1		207	15	306-341	0.71	0.763	0.73	0.089	0.037	C+S	x	x	x
VVIO02	1	9,813,247	1	no	no	no	1	1	1	1		207	4	130-141	0.222	0.212	0.194	0.64	-0.021	C		x	x
VVIN61	1	10,569,689	2	no	no	yes	1	1	2	2	1	206	11	343-370	0.636	0.801	0.775	0.065	0.108	C		x	x
VChr1b	1	11,850,801	1	no	no	yes	1	1	2	1		206	5	95-119	0.544	0.561	0.504	0.249	0.016	C		x	x
VVIM25	1	12,617,249	2	no	no	no	1	1	1	1		207	17	112-156	0.744	0.762	0.728	0.09	0.011	C+S		x	x
RTAChr1-14a	1	13,253,163	3	no	no	no	1	1	2	2	1	193	25	152-218	0.554	0.928	0.92	0.01	0.252	C+S		x	x
FAM141	1	17,732,854	2	no	yes	no	1	1	2	1		207	4	170-182	0.618	0.618	0.547	0.217	0.001	C+S		x	x
VMC2B3	1	19,375,299	3	no	no	no	1	1	2	1		207	10	162-188	0.821	0.813	0.787	0.06	-0.007	C		x	x
VMC9F2	1	20,980,935	2	no	yes	no	1	1	2	1		207	15	209-311	0.812	0.841	0.823	0.042	0.015	C		x	x
VVIO61	1	22,221,400	2	no	no	no	1	1	2	1		207	7	217-238	0.715	0.678	0.628	0.153	-0.033	C		x	x
VVIF52	1	23,292,680	2	no	no	no	1	1	1	1		207	14	249-306	0.763	0.771	0.747	0.076	-0.001	C		x	x
VMC9D3	1	24,039,899	2	no	no	no	1	1	1	1		207	15	178-210	0.715	0.728	0.696	0.105	0.014	C+S		x	x
UDV-109	2	132,291	2	no	yes	no	1	1	2	1		206	11	128-164	0.51	0.557	0.5	0.253	0.039	C		x	x
VVIB01	2	2,161,587	2	no	no	no	1	1	1	1		207	5	288-307	0.667	0.65	0.587	0.185	-0.016	C+S	x	x	x
VVS3-2	2	4,014,388	1	no	no	no	1	1	1	1		207	4	121-128	0.657	0.512	0.397	0.353	-0.127	C		x	x
VMC3B10	2	5,000,200	3	no	no	no	1	1	2	1		207	22	96-160	0.928	0.743	0.71	0.098	-0.131	C+S		x	x
VMC2C10-1	2	6,975,274	2	no	no	no	1	1	1	1		207	11	135-178	0.729	0.739	0.703	0.103	0.005	C+S		x	x
VChr2c	2	7,998,935	1	no	yes	no	1	1	2	1		207	2	150-153	0.29	0.473	0.361	0.39	0.239	C+S		x	x
VMC5G7	2	8,334,940	2	no	no	no	1	1	1	2	2	207	14	179-217	0.758	0.754	0.72	0.094	-0.006	C		x	x
FAM140	2	9,166,714	1	no	yes	no	1	1	2	2	1	206	5	248-264	0.563	0.652	0.582	0.19	0.066				
FAM140-2	2	9,166,787	1	no	no	no	1	1	1	1		207	4	174-177	0.56	0.542	0.439	0.312	-0.018	C		x	x

Marker		Quality of amplification profiles								Checking		Genetic diversity								Use		
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis
RTAChr2-11a	2	10,707,832	2	no	no	no	1	1	1	2	1	203	6	271-300	0.478	0.559	0.461	0.292	0.08	C+S	x	x
FAM68	2	13,188,507	1	yes	no	no	1	1	2	1		207	2	305-308	0.01	0.01	0.01	0.981	0	C		x
FAM24	2	13,508,269	3	no	no	no	1	1	1	1		207	6	262-272	0.609	0.596	0.521	0.238	-0.017	C	x	x
VV6718	2	16,446,991	2	no	no	no	1	1	1	1		207	7	179-190	0.309	0.301	0.288	0.502	-0.027	C+S	x	x
VVIU20	2	16,650,964	3	no	no	no	1	1	2	1		207	5	363-385	0.372	0.373	0.348	0.418	-0.008	C	x	x
VMC8C2	2	17,626,961	2	no	no	no	1	1	1	1		207	7	130-151	0.585	0.539	0.494	0.257	-0.052	C	x	x
VMC7g3	2	18,382,886	2	no	no	no	1	1	1	1		207	17	115-161	0.585	0.587	0.571	0.186	-0.005	C+S	x	x
VMCNG1D3	2	18,860,462	3	no	no	no	1	2	3													
UDV-093	3	1,335,056	2	no	no	no	1	1	1	1		207	9	136-171	0.657	0.675	0.643	0.137	0.017	C+S	x	x
VMC8F10	3	3,324,131	2	no	no	no	1	1	1	1		207	14	199-237	0.787	0.808	0.784	0.06	0.012	C	x	x
UDV-021-2	3	5,287,429	2	no	no	no	1	1	1	1		207	11	250-273	0.662	0.744	0.7	0.109	0.056	C+S	x	x
VMC1A5	3	5,954,608	2	no	no	no	1	1	1	1		207	12	178-205	0.773	0.791	0.766	0.068	0.01	C+S	x	x
UDV-022	3	7,546,783	2	no	yes	no	1	1	2	1		207	13	192-252	0.729	0.77	0.736	0.086	0.022	C	x	x
VMC1G7	3	8,547,609	2	no	no	no	1	1	1	2	1	206	17	221-264	0.786	0.885	0.872	0.025	0.059	C	x	x
VChr3a	3	9,542,014	2	no	no	no	1	1	1	1		207	13	175-229	0.831	0.861	0.844	0.035	0.017	C	x	x
VVIN54	3	11,093,607	2	yes	no	no	1	1	2	1		207	14	82-127	0.729	0.843	0.823	0.043	0.072	C	x	x
VVMD28	3	12,005,376	2	no	no	no	1	1	1	1		207	18	218-276	0.87	0.879	0.865	0.027	0.006	C+S	x	x
RTAChr3-15a	3	15,937,188	2	no	no	no	1	1	1	1		207	12	261-290	0.686	0.798	0.773	0.065	0.079	C+S	x	x
RTAChr3-17a	3	18,564,018	2	no	no	no	1	1	1	2	1	206	7	167-185	0.33	0.457	0.427	0.324	0.156	C+S		x
RTAChr3-19b	3	20,400,237	3	no	no	no	1	1	2	1		207	31	279-363	0.899	0.931	0.925	0.009	0.017	C+S	x	x
Vchr4a	4	946,454	1	no	no	no	1	1	1	1		207	7	181-204	0.618	0.647	0.58	0.191	0.024	C	x	x
VVIQ25	4	2,036,638	2	no	no	no	1	1	1	2	1	156	2	185-187	0.141	0.401	0.32	0.441	0.478			
VVIQ25-2	4	2,036,638	1	yes	yes	no	1	1	2	1		201	7	249-295	0.632	0.723	0.673	0.125	0.066	C	x	x
VMC4D4	4	3,753,977	1	no	no	no	1	1	1	1		207	7	152-176	0.594	0.558	0.515	0.238	-0.047	C+S	x	x
VMC7h3	4	4,719,044	2	no	no	no	1	1	1	1		207	21	119-175	0.749	0.783	0.754	0.075	0.022	C	x	x
VMCNG2E1	4	5,232,859	2	no	no	no	1	1	1	1		207	11	100-133	0.729	0.711	0.673	0.12	-0.019	C+S	x	x
RTAChr4-9a	4	9,511,967	3	no	no	no	2	1	2	1		207	9	253-268	0.778	0.815	0.788	0.06	0.024	C+S	x	x
UDV-110	4	11,543,793	2	no	yes	no	1	1	2	1		207	13	151-200	0.57	0.709	0.662	0.131	0.116	C	x	x
VRZAG21	4	13,794,845	2	no	no	no	1	1	1	1		207	9	190-213	0.816	0.796	0.765	0.071	-0.012	C	x	x
VVIN75	4	14,969,759	2	no	no	no	1	1	1	2	1	207	6	178-195	0.367	0.519	0.434	0.316	0.17			
VVIN75-2	4	14,969,759	2	no	no	no	1	1	1	1		207	7	271-288	0.705	0.662	0.6	0.176	-0.035	C	x	x
VMC8B11	4	17,682,875	1	no	yes	no	1	1	2	1		207	6	158-186	0.633	0.652	0.582	0.191	0.016	C	x	x

Marker			Quality of amplification profiles							Checking		Genetic diversity							Use				
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PIC	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis
VVMD32	4	18,402,264	2	no	no	no	1	1	1	1		207	13	238-288	0.86	0.84	0.819	0.045	-0.014	C+S	x	x	x
VVIP77	4	19,684,609	2	no	no	no	1	1	1	1		207	13	170-211	0.816	0.838	0.816	0.047	0.011	C		x	x
VRZAG83	4	20,627,004	1	no	no	no	1	1	1	1		207	7	159-201	0.802	0.73	0.68	0.122	-0.052	C+S		x	x
VMC6g10	4	23,949,251	2	no	no	no	1	1	1	1		207	12	121-195	0.614	0.618	0.582	0.182	-0.001	C		x	x
FAM91	5	782,861	1	no	no	no	1	1	1	1		207	2	156-159	0.324	0.389	0.313	0.45	0.09	C+S		x	x
VVC6	5	1,470,620	2	no	no	no	1	1	1	1		207	6	109-123	0.43	0.405	0.381	0.379	-0.051	C		x	x
FAM12	5	2,783,225	3	no	no	no	1	1	1	1		207	8	352-373	0.763	0.753	0.715	0.098	-0.007	C+S		x	x
VVIB72	5	2,848,403	1	no	no	no	1	1	1	1		207	2	181-183	0.053	0.07	0.067	0.868	0.132	C		x	x
VChr5c	5	4,080,604	1	no	no	yes	1	1	2	1		205	8	84-124	0.776	0.769	0.734	0.088	-0.009	C		x	x
VVMD27	5	4,472,022	2	no	no	no	1	1	1	1		207	11	175-194	0.802	0.836	0.811	0.05	0.019	C+S	x	x	x
VRZAG79	5	5,692,460	2	no	no	no	1	1	1	1		207	12	237-267	0.855	0.86	0.844	0.034	0.002	C+S		x	x
VVIT68	5	7,099,947	1	no	no	no	1	1	1	1		207	8	250-285	0.715	0.715	0.669	0.126	-0.004	C		x	x
UDV-053	5	8,541,098	2	no	no	no	1	1	1	2	1	202	11	116-156	0.594	0.791	0.764	0.069	0.141	C		x	x
VChr5b	5	9,474,726	1	no	yes	no	1	1	2	1		207	9	184-210	0.783	0.722	0.673	0.125	-0.042	C+S		x	x
VVIN33	5	11,721,184	2	no	no	no	1	1	1	1		207	11	266-300	0.807	0.808	0.783	0.061	-0.004	C		x	x
VMC3C7	5	12,716,770	2	no	yes	no	1	1	2	1		206	11	104-135	0.82	0.85	0.83	0.041	0.016	C		x	x
VChr5a	5	14,448,749	1	no	no	no	1	1	1	1		207	12	179-257	0.821	0.805	0.783	0.059	-0.012	C		x	x
FAM122	5	15,723,310	2	no	no	no	1	1	1	1		206	6	348-360	0.515	0.661	0.62	0.156	0.125	C		x	x
VMC16D4-2	5	17,606,676	2	no	no	no	1	1	1	1		207	20	209-278	0.855	0.839	0.819	0.044	-0.012	C		x	x
VMC16D4	5	17,606,754	2	no	no	no	1	1	1	2	1	207	18	143-218	0.464	0.828	0.806	0.051	0.279				
VMC9B5	5	20,174,295	2	no	no	no	1	1	1	1		207	9	236-262	0.783	0.781	0.746	0.082	-0.004	C		x	x
UDV-077	5	21,538,151	2	no	no	no	1	1	1	1		207	14	149-197	0.681	0.705	0.671	0.12	0.018	C+S		x	x
VRZAG89	5	22,919,299	1	no	no	no	1	1	1	1		207	2	160-171	0.188	0.179	0.163	0.691	-0.027	C		x	x
VVIN40	5	24,429,944	2	no	no	no	1	1	1	1		205	12	218-250	0.673	0.692	0.665	0.121	0.004	C		x	x
VMC4C6	5	25,493,869	1	no	yes	no	1	1	2	1		207	7	153-177	0.696	0.676	0.628	0.153	-0.013	C		x	x
UDV-090	6	1,012,094	2	no	yes	no	1	1	2	1		207	8	105-174	0.773	0.754	0.721	0.092	-0.017	C		x	x
VVIP72	6	2,228,161	2	no	no	no	1	1	1	2	1	201	8	73-92	0.368	0.524	0.471	0.279	0.16				
VVIP72-2	6	2,280,007	2	no	no	no	1	1	1	1		207	12	230-256	0.72	0.681	0.646	0.136	-0.034	C		x	x
FAM110	6	3,652,483	1	no	no	no	1	1	1	1		207	3	269-284	0.546	0.516	0.432	0.318	-0.033	C		x	x
UDV-085	6	4,906,382	2	no	no	no	1	1	1	1		207	14	114-151	0.831	0.875	0.859	0.03	0.025	C		x	x
VMC2H9	6	5,819,480	2	no	no	no	1	1	1	1		207	11	115-156	0.671	0.719	0.672	0.125	0.037	C+S		x	x
VMC2F10	6	7,615,634	2	no	no	no	1	1	1	1		207	11	92-132	0.768	0.753	0.716	0.097	-0.016	C		x	x

Marker			Quality of amplification profiles								Checking		Genetic diversity								Use			
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PIC	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis	
VVIC50	6	8,450,955	2	no	yes	yes	1	1	2	2	1	207	7	75-89	0.44	0.518	0.482	0.268	0.081	C		x	x	
VV5799	6	11,606,406	2	no	no	no	1	1	1	1		207	11	209-239	0.836	0.826	0.806	0.049	-0.008	C+S		x	x	
VVMD21	6	14,883,818	2	no	no	no	1	1	1	1		207	8	243-265	0.691	0.696	0.658	0.13	-0.001	C+S	x	x	x	
VMC4G6	6	16,304,885	2	no	yes	no	1	1	2	1		207	11	119-139	0.657	0.794	0.765	0.07	0.096	C		x	x	
VMC5G1-1	6	17,486,959	2	no	no	no	1	1	1	1		207	14	120-282	0.643	0.664	0.638	0.139	0.017	C+S		x	x	
VVIN31	6	18,937,746	2	no	no	no	1	1	1	1		207	13	141-200	0.787	0.813	0.791	0.056	0.01	C		x	x	
VVIM43	6	20,141,244	2	no	no	no	1	1	1	1		207	12	77-100	0.87	0.876	0.86	0.029	0.002	C		x	x	
FAM96	6	21,584,205	1	no	no	no	1	1	1	1		207	2	283-308	0.005	0.005	0.005	0.99	0	C			x	
VChr7c	7	545,849	1	no	yes	no	1	1	2	1		207	4	105-114	0.884	0.662	0.587	0.188	-0.145	C		x	x	
VVMD7	7	1,560,528	2	no	no	no	1	1	1	1		207	12	231-261	0.812	0.819	0.795	0.055	0.003	C+S	x	x	x	
VRZAG62	7	2,148,817	2	no	no	no	1	1	1	1		207	9	186-204	0.802	0.81	0.785	0.06	0.002	C+S		x	x	
VVIB22	7	3,521,457	2	no	no	no	1	1	1	2	1	207	13	130-196	0.449	0.726	0.706	0.094	0.243					
VVIB22-2	7	3,521,457	2	no	no	no	1	1	1	2	1	207	11	163-224	0.502	0.741	0.719	0.088	0.186	C		x	x	
VVIS58	7	5,335,402	1	no	no	no	1	1	1	2	1	207	10	281-310	0.469	0.671	0.613	0.166	0.189					
VVIS58-2	7	5,335,417	2	no	no	no	1	1	1	1		207	10	266-294	0.671	0.652	0.605	0.168	-0.023	C		x	x	
VMC7A4	7	7,936,887	2	no	yes	no	1	1	2	1		207	21	112-173	0.884	0.871	0.856	0.03	-0.009	C+S		x	x	
VVIV36	7	8,779,579	1	no	no	no	1	1	1	1		207	5	218-228	0.662	0.648	0.577	0.195	-0.013	C+S		x	x	
RTAChr7-9	7	9,947,798	3	no	no	no	1	1	2	1		207	23	254-304	0.821	0.883	0.871	0.024	0.037	C+S		x	x	
VMC8D11	7	18964323	2	no	no	no	1	1	1	1		207	11	117-143	0.874	0.849	0.829	0.041	-0.016	C		x	x	
FAM116	7	20,388,473	2	no	no	no	1	1	1	1		207	4	314-322	0.488	0.573	0.496	0.26	0.089	C		x	x	
VMC1A12	7	22,589,274	2	yes	no	no	1	1	1	1		207	12	120-167	0.821	0.794	0.766	0.07	-0.02	C+S		x	x	
VVIQ67	7	22,899,285	2	yes	no	no	1	1	2	1		207	12	252-272	0.56	0.76	0.725	0.091	0.145	C		x	x	
VChr7a	7	23,070,829	1	no	yes	no	1	1	2	1		207	3	126-140	0.454	0.524	0.409	0.341	0.073	C+S		x	x	
VVIV04	7	23,860,671	2	yes	no	no	2	1	2	1		207	16	253-276	0.783	0.829	0.81	0.047	0.027	C+S		x	x	
VVIN56	7	25,737,398	1	no	no	no	1	1	1	1		207	6	161-176	0.522	0.712	0.661	0.133	0.155	C		x	x	
VMC1F10	8	292,910	2	no	yes	no	1	1	2	1		207	10	190-208	0.643	0.816	0.792	0.056	0.125	C		x	x	
UDV-075	8	2,400,696	2	no	yes	no	1	1	2	1		207	7	145-168	0.652	0.689	0.629	0.156	0.026	C+S		x	x	
VChr8b	8	4,343,076	2	no	no	no	1	1	2	1		207	20	100-455	0.787	0.892	0.88	0.022	0.061	C+S		x	x	
VMCNG2H2.2	8	7,184,551	2	no	yes	no	1	1	2	1		207	17	115-219	0.638	0.861	0.844	0.035	0.152	C		x	x	
VMC3G7	8	8,553,958	3	no	no	no	1	2	3															
VMC7h2	8	10,149,243	1	no	no	no	1	1	1	1		207	5	119-132	0.671	0.678	0.62	0.161	-0.001	C		x	x	
VMC1B11	8	14,039,245	2	no	no	no	1	1	1	1		207	12	167-196	0.874	0.825	0.803	0.052	-0.031	C+S	x	x	x	

Marker			Quality of amplification profiles							Checking		Genetic diversity								Use			
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PIC	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis
VChr8a	8	15,200,499	2	no	no	no	1	1	1	2	1	207	14	173-225	0.618	0.832	0.81	0.049	0.146				
VChr8a-2	8	15,200,499	2	no	no	no	1	1	1	2	1	202	14	208-260	0.629	0.824	0.8	0.053	0.133	C		x	x
VMC5H2	8	15,407,796	2	no	no	no	1	1	1	1		207	8	187-210	0.502	0.507	0.471	0.279	0.007	C		x	x
VMC1E8	8	17,261,639	2	no	no	no	1	1	1	1		207	10	206-231	0.773	0.805	0.779	0.063	0.021	C		x	x
VVIB66	8	18,270,246	2	no	yes	no	1	1	2	1		207	9	85-105	0.841	0.825	0.803	0.052	-0.012	C		x	x
VMC2H10	8	20,816,914	2	no	yes	no	1	1	2	1		207	11	105-133	0.841	0.838	0.816	0.047	-0.008	C+S		x	x
UDV-026	8	21,473,580	2	no	no	no	1	1	1	1		207	8	133-155	0.812	0.781	0.746	0.082	-0.022	C		x	x
VMC1C10	9	563,536	2	no	no	no	1	1	1	1		207	16	129-185	0.841	0.863	0.848	0.033	0.013	C		x	x
FAM65	9	1,202,649	1	no	no	no	1	1	1	1		207	2	126-129	0.208	0.224	0.198	0.628	0.036	C+S		x	x
VMC6d12	9	3,792,236	2	no	no	no	1	1	1	1		207	13	131-181	0.826	0.827	0.81	0.046	-0.007	C+S		x	x
VMCNG1E4-2	9	5,448,330	3	no	no	no	2	2	3														
VChr9a	9	6,426,756	1	no	no	no	1	1	1	1		207	8	91-124	0.797	0.761	0.727	0.09	-0.031	C+S		x	x
FAM42	9	7,551,193	3	no	yes	no	1	1	2	2	1	207	8	371-391	0.643	0.804	0.776	0.065	0.106	C		x	x
VMC5C1	9	9,396,293	2	no	yes	no	1	1	2	2	1	207	11	143-190	0.57	0.724	0.695	0.105	0.105	C+S		x	x
VVIQ66	9	10,808,981	1	no	no	no	1	1	1	1		207	4	179-185	0.29	0.277	0.258	0.542	-0.021	C		x	x
RTAChr9-13a	9	13,102,063	3	no	no	no	1	1	2	1		207	18	250-319	0.879	0.883	0.869	0.026	0.001	C+S		x	x
FAM104	9	18,036,034	1	no	no	no	1	1	1	1		207	1	112	0	0	0	1	ND				
VMC2D9	9	18,966,688	1	yes	yes	no	1	1	1	2	1	206	9	69-117	0.684	0.792	0.765	0.069	0.069				
VMC2D9-2	9	18,966,688	2	no	no	yes	1	1	2	1		207	10	139-195	0.821	0.862	0.844	0.035	0.024	C		x	x
UDV-040	9	20,091,003	2	no	no	no	1	1	1	1		207	12	135-174	0.773	0.761	0.732	0.085	-0.006	C		x	x
VVIQ52	9	21,558,043	2	no	no	no	1	1	1	1		207	5	82-93	0.686	0.667	0.604	0.173	-0.014	C+S	x	x	x
VMC2E11	9	22,676,139	2	yes	yes	no	1	1	2	1		207	15	64-162	0.807	0.86	0.844	0.034	0.033	C		x	x
VVIH01	10	1,473,028	2	no	no	no	2	1	2	1		207	14	237-262	0.807	0.809	0.782	0.063	-0.002	C		x	x
VRZAG67	10	5,006,368	2	no	no	no	1	1	1	1		207	13	124-162	0.836	0.834	0.812	0.048	-0.004	C+S		x	x
VMC2A10	10	5,982,405	2	no	yes	no	1	1	2	1		207	15	98-148	0.681	0.814	0.793	0.054	0.082	C+S		x	x
UDV-059	10	9,016,614	2	no	no	no	1	1	1	1		207	9	133-171	0.802	0.812	0.782	0.064	0.004	C		x	x
VVIN78	10	10,476,651	1	no	no	no	1	1	1	1		207	2	160-162	0.483	0.497	0.373	0.377	0.013	C		x	x
FAM45	10	11,414,249	3	no	no	no	2	1	2	1		207	11	365-377	0.787	0.782	0.758	0.071	-0.004	C+S		x	x
VVIV37	10	14,590,969	2	no	no	no	1	1	1	1		207	15	150-181	0.807	0.86	0.844	0.034	0.03	C+S	x	x	x
UDV-063	10	15,454,368	2	no	no	no	1	1	1	1		207	12	117-150	0.812	0.83	0.808	0.05	0.01	C		x	x
FAM63	10	23,407,091	1	no	no	no	1	1	1	1		207	4	234-337	0.121	0.247	0.219	0.595	0.34	C		x	x
VMC3E12	11	774,942	2	no	no	no	1	1	1	1		207	14	107-157	0.633	0.625	0.597	0.168	-0.015	C		x	x

Marker		Quality of amplification profiles									Checking		Genetic diversity							Use		
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis
VMC2A12	11	1,542,611	2	no	no	no	1	1	1	1		207	9	96-124	0.372	0.38	0.369	0.396	-0.003	C	x	x
VVMD25	11	2,971,693	2	no	no	no	1	1	1	1		207	12	237-269	0.826	0.792	0.757	0.077	-0.023	C+S	x	x
VVS2	11	3,909,894	2	no	no	no	1	1	1	1		207	17	123-168	0.874	0.866	0.85	0.033	-0.007	C+S	x	x
VVIB19	11	5,865,805	2	no	no	no	1	1	1	1		207	7	374-405	0.7	0.71	0.656	0.137	0.008	C	x	x
UDV-100	11	7,911,371	2	no	no	no	1	1	1	2	1	201	8	216-240	0.259	0.71	0.675	0.118	0.463	C		x
VMC6G1	11	8,983,138	2	no	no	no	1	1	1	1		207	11	171-199	0.845	0.826	0.801	0.054	-0.014	C	x	x
RTAChr11-11b	11	11,365,896	3	no	no	no	1	1	3	2	1	207	15	141-198	0.246	0.699	0.664	0.125	0.484	C+S		x
VVIP02	11	11,931,480	2	no	no	no	1	1	1	1		207	5	266-274	0.478	0.505	0.445	0.305	0.026	C+S	x	x
VChr11b	11	17,380,059	1	no	yes	no	1	1	2	1		207	5	154-166	0.618	0.714	0.662	0.132	0.072	C+S	x	x
VVMD8	11	19,974,886	3	no	no	yes	1	1	2	1		207	23	131-215	0.754	0.798	0.772	0.066	0.026	C	x	x
TT251F02	12	418,480	2	no	no	no	1	1	1	2	1	207	6	190-205	0.522	0.592	0.507	0.252	0.063			
TT251F02-2	12	418,506	2	no	no	no	1	1	1	1		207	6	163-178	0.623	0.628	0.552	0.214	0.001	C	x	x
VMC8G6	12	2,354,142	2	no	no	no	1	1	1	1		207	19	134-193	0.773	0.85	0.832	0.04	0.049	C+S	x	x
VMC4H2	12	4,073,541	2	no	no	yes	1	1	2	1		207	25	56-110	0.831	0.909	0.9	0.015	0.047	C	x	x
VMC2H4	12	5,447,330	2	no	yes	no	1	1	2	1		207	15	118-237	0.816	0.86	0.845	0.033	0.024	C+S	x	x
VChr12a	12	6,174,660	1	no	no	no	1	1	1	1		207	8	122-147	0.671	0.657	0.593	0.181	-0.014	C+S	x	x
VMC1G3-2	12	9,589,001	2	no	no	no	1	1	1	1		207	14	115-170	0.696	0.687	0.658	0.126	-0.009	C	x	x
VVIM11	12	11,103,248	2	no	no	no	1	1	1	1		207	12	260-298	0.831	0.821	0.794	0.058	-0.008	C	x	x
VMC4F3-1	12	16,417,442	2	no	no	no	2	1	2	1		207	17	165-208	0.855	0.872	0.859	0.028	0.008	C+S	x	x
UDV-058	12	19,001,459	1	no	no	no	1	1	1	1		207	3	155-160	0.556	0.548	0.455	0.297	-0.01	C	x	x
VMC8G9	12	21,945,210	2	no	no	no	1	1	1	1		207	15	162-221	0.86	0.858	0.841	0.035	-0.002	C	x	x
GT183H07	13	83,697	2	no	no	no	1	1	1	1		207	3	251-260	0.498	0.491	0.377	0.373	-0.008	C	x	x
VVIN62	13	1,387,838	2	no	no	no	1	1	1	1		207	7	353-367	0.531	0.53	0.499	0.252	-0.005	C+S	x	x
VVIH54	13	3,434,661	2	no	no	no	1	1	1	1		207	15	140-181	0.725	0.726	0.691	0.109	0.004	C+S	x	x
VChr13c	13	4,525,996	1	no	no	no	1	1	1	1		207	5	114-135	0.686	0.644	0.597	0.173	-0.035	C	x	x
VVIM63	13	4,909,035	2	no	no	no	1	1	1	1		207	3	301-310	0.527	0.551	0.481	0.271	0.026	C	x	x
VVIM01	13	6,609,298	1	no	no	no	1	1	1	1		207	4	182-188	0.295	0.285	0.266	0.531	-0.023	C	x	x
VMC3D12	13	8,193,616	2	no	no	no	1	1	1	1		207	23	197-251	0.841	0.871	0.856	0.029	0.017	C	x	x
UDV-038	13	9,494,599	3	no	no	no	1	2	3													
UDV-124	13	11,412,148	2	no	no	no	1	1	1	1		207	16	188-239	0.845	0.891	0.879	0.023	0.025	C+S	x	x
RTAChr13-13b	13	13,446,986	3	no	no	no	1	1	2	2	1	178	9	226-247	0.27	0.574	0.547	0.209	0.37	C+S		x
FAM62-2	13	16,059,252	1	no	no	no	1	1	1	1		207	5	321-335	0.271	0.24	0.214	0.604	-0.057	C+S	x	x

Marker			Quality of amplification profiles								Checking		Genetic diversity								Use			
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PIC	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis	
RTAChr13-16a	13	16,119,425	3	no	no	no	1	1	2	1		207	20	242-324	0.812	0.864	0.849	0.032	0.031	C+S	x	x		
VMC9H4_2	13	18,412,652	2	no	no	yes	1	1	2	1		207	19	255-316	0.884	0.904	0.894	0.018	0.011	C		x	x	
UDV-020	13	19,113,495	3	no	no	no	1	2	3															
RTAChr13-20a	13	20,382,252	2	no	no	no	1	1	1	1		207	11	224-263	0.72	0.718	0.68	0.117	0	C+S		x	x	
VChr13b	13	24776322	1	no	no	no	1	1	1	1		207	9	149-178	0.696	0.667	0.636	0.141	-0.026	C		x	x	
VVIP10	13	26,977,865	2	no	no	no	1	1	1	2	1	207	10	277-327	0.444	0.775	0.736	0.088	0.269	C		x	x	
VMC8E6	13	27,180,657	2	no	no	no	1	1	1	2	1	207	11	209-276	0.623	0.823	0.799	0.054	0.135	C+S		x	x	
RTAChr13-23b	13	27,762,807	2	no	no	no	1	1	1	1		207	11	273-309	0.812	0.829	0.809	0.048	0.004	C+S		x	x	
VMCNG1E1	14	1,413,666	2	no	no	no	1	1	1	1		207	16	91-129	0.807	0.842	0.822	0.043	0.023	C		x	x	
VVIP05	14	2,861,324	2	no	no	no	1	1	1	2	1	189	10	243-282	0.143	0.791	0.759	0.075	0.696	C			x	
UDV-050	14	4,943,488	2	no	no	no	1	1	1	1		207	7	152-184	0.478	0.568	0.524	0.23	0.081	C		x	x	
VMC1E12	14	7,136,672	2	no	no	no	1	1	1	1		207	11	233-254	0.816	0.809	0.782	0.062	-0.007	C		x	x	
FAM135	14	8,755,200	1	no	no	no	1	1	1	1		207	4	394-402	0.285	0.285	0.266	0.531	-0.006	C		x	x	
VMC2H12	14	10,303,957	2	no	yes	no	1	1	2	1		207	7	96-117	0.71	0.707	0.667	0.126	-0.007	C+S		x	x	
VVC34	14	13,207,677	2	no	no	no	1	1	1	1		207	11	185-210	0.734	0.711	0.664	0.13	-0.019	C+S		x	x	
VVIP22	14	18,364,396	2	no	no	no	1	1	1	1		207	13	327-374	0.802	0.848	0.828	0.042	0.028	C		x	x	
VMC2B11	14	20,429,749	1	no	no	no	1	1	1	1		207	6	168-184	0.739	0.783	0.752	0.077	0.033	C		x	x	
VRZAG112	14	22,042,800	2	no	no	no	1	1	1	1		207	11	228-260	0.768	0.834	0.812	0.048	0.043	C+S		x	x	
VVMD24	14	24,251,864	2	no	no	no	1	1	1	1		207	10	206-222	0.681	0.709	0.673	0.12	0.011	C+S	x	x	x	
UDV-095	14	26,102,129	2	no	no	no	1	1	1	1		207	15	121-145	0.773	0.739	0.717	0.089	-0.028	C		x	x	
VVIS70	14	27,477,542	2	no	no	no	1	1	1	2	1	207	13	185-217	0.43	0.702	0.676	0.114	0.242					
VVIS70-2	14	27,477,542	2	no	yes	no	1	1	2	1		207	18	185-231	0.812	0.829	0.811	0.046	0.01	C		x	x	
VVIN94	14	28,772,134	2	no	no	no	1	1	1	1		206	14	271-307	0.777	0.77	0.737	0.085	-0.009	C		x	x	
VVIN70	14	29,853,616	2	no	no	no	1	1	1	1		207	7	356-373	0.7	0.74	0.692	0.114	0.028	C		x	x	
RTAChr15-1b	15	1167751	2	no	no	no	1	1	1	1		207	14	270-304	0.763	0.821	0.795	0.057	0.039	C+S		x	x	
RTAChr15-3b	15	2650923	2	no	yes	no	1	1	2	2	1	197	11	278-321	0.548	0.677	0.65	0.13	0.122	C+S		x	x	
RTAChr15-4a	15	4019057	3	no	no	no	1	1	3	2	1	190	19	223-324	0.495	0.833	0.812	0.048	0.253	C+S		x	x	
UDV-116	15	7,789,873	2	no	no	no	1	1	1	1		207	12	119-154	0.86	0.836	0.814	0.048	-0.016	C		x	x	
RTAChr15-7a	15	8,439,107	3	no	no	no	1	1	2	2	1	153	9	195-230	0.229	0.771	0.741	0.081	0.546	C			x	
VVS16	15	9,570,305	1	no	no	no	1	1	1	1		207	7	264-293	0.536	0.551	0.526	0.227	0.001	C		x	x	
VVIV67	15	10,897,392	2	no	no	no	2	1	2	1		207	19	348-401	0.879	0.87	0.854	0.031	-0.007	C+S	x	x	x	
VVIO61	15	11,761,810	1	no	no	no	1	1	1	1		207	3	359-364	0.609	0.583	0.499	0.258	-0.024	C		x	x	

Marker			Quality of amplification profiles								Checking	Genetic diversity							Use				
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PIC	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis
VVIP33	15	13,060,296	2	no	no	no	1	1	1	1		207	10	392-412	0.85	0.819	0.794	0.057	-0.018	C+S	x	x	
VMC5G8	15	14,895,031	2	no	no	no	1	1	1	1		207	11	281-319	0.739	0.778	0.741	0.085	0.023	C	x	x	
VVIM42A	15	15,943,609	1	yes	no	no	1	1	1	1		207	11	258-308	0.763	0.771	0.742	0.08	0.001	C+S	x	x	
VMC4D9-2	15	16,581,058	2	no	no	no	1	1	1	1		207	13	222-268	0.773	0.841	0.823	0.042	0.042	C+S	x	x	
VMC8G3	15	17,666,601	3	yes	no	no	2	1	2	2	1	206	17	253-302	0.835	0.855	0.838	0.037	0.01				
VMC8G3-2	15	17,666,601	3	no	no	no	2	1	2	1		207	17	196-244	0.812	0.848	0.831	0.039	0.025	C	x	x	
VChr15a-2	15	19,977,651	1	no	yes	no	1	1	2	1		207	7	227-262	0.551	0.668	0.611	0.167	0.096	C	x	x	
VChr15a	15	19,977,749	1	no	no	no	1	1	1	2	1	207	9	128-201	0.271	0.689	0.642	0.143	0.432				
VVIN52	16	619,678	2	no	no	no	1	1	1	2	1	206	15	76-135	0.524	0.826	0.806	0.049	0.229	C	x	x	
UDV-032	16	2,288,155	2	no	yes	no	1	1	2	1		207	9	141-179	0.71	0.775	0.739	0.086	0.045	C	x	x	
VChr16a	16	3,187,867	1	no	no	no	1	1	1	1		207	7	100-168	0.459	0.474	0.446	0.305	0.026	C+S	x	x	
RTAChr16-6b	16	6,021,381	2	no	no	no	1	1	1	1		207	9	199-224	0.841	0.785	0.75	0.08	-0.037	C+S	x	x	
FAM92	16	7,976,858	3	no	no	no	1	1	2	1		207	15	382-456	0.797	0.745	0.718	0.091	-0.034	C	x	x	
UDV-009	16	9,642,339	2	no	no	no	1	1	1	1		207	12	112-175	0.744	0.742	0.716	0.092	0.002	C	x	x	
UDV-104	16	10,785,478	3	no	no	no	1	1	2	2	1	205	26	134-241	0.649	0.89	0.88	0.02	0.157	C		x	
VMC1E11-2	16	15,219,141	2	no	no	no	1	1	1	2	1	207	11	253-316	0.763	0.828	0.804	0.053	0.039	C+S	x	x	
VChr16b	16	17,432,683	1	yes	no	no	1	1	2	1		207	10	168-200	0.565	0.592	0.56	0.199	0.028	C	x	x	
FAM18	16	19,490,645	1	no	no	no	1	1	1	1		207	1	205	0	0	0	1	ND	C		x	
FAM70-2	16	21,189,056	2	no	no	no	1	1	1	1		207	15	221-352	0.874	0.864	0.848	0.033	-0.007	C	x	x	
FAM70	16	21,189,258	2	no	yes	no	1	1	2	2	1	207	18	279-409	0.362	0.895	0.884	0.021	0.425				
VVMD5	16	23,305,077	2	no	no	no	1	1	1	1		207	10	219-243	0.816	0.866	0.848	0.034	0.028	C+S	x	x	x
VMC4B7-2	16	23,329,085	2	no	no	no	1	2	3														
VMC3C11-1	17	2,296,673	2	no	no	no	1	1	1	1		207	12	97-119	0.758	0.757	0.717	0.098	-0.002	C	x	x	
VMC2H3	17	3,684,833	2	no	no	no	1	1	1	1		207	20	68-131	0.845	0.849	0.831	0.039	0.001	C+S	x	x	
VChr17b-2	17	5,017,280	1	no	no	no	1	1	1	1		207	2	242-249	0.246	0.272	0.234	0.568	0.048	C	x	x	
VChr17b	17	5,017,367	1	yes	no	no	1	1	2	2	1	204	2	163-164	0.064	0.062	0.06	0.882	-0.008				
VVIN73	17	5,636,037	2	no	no	no	1	1	1	1		207	7	254-267	0.29	0.318	0.3	0.483	0.072	C+S	x	x	x
VRZAG15	17	6,588,595	2	no	no	no	1	1	1	1		207	11	167-201	0.401	0.422	0.403	0.353	0.034	C	x	x	
VVIV22	17	7,660,564	1	no	no	no	1	1	1	1		199	2	99-101	0	0.01	0.01	0.98	0.239	C		x	
VChr17a	17	9,315,344	1	no	yes	no	1	1	2	2	1	206	5	159-185	0.301	0.439	0.35	0.404	0.193	C	x	x	
VVIV08	17	12,161,420	2	no	no	no	1	1	1	1		207	12	341-382	0.633	0.727	0.693	0.108	0.074	C+S	x	x	
RTAChr17-15a	17	16,316,158	2	no	no	no	1	1	1	1		207	9	271-289	0.662	0.67	0.638	0.14	-0.003	C+S	x	x	

Marker		Quality of amplification profiles							Checking		Genetic diversity										Use		
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PI _C	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis
RTAChr17-15b	17	16,474,837	3	no	no	yes	1	1	2	2	1	205	24	240-307	0.493	0.893	0.882	0.02	0.289	C+S		x	
VMC3E5	18	321,045	2	no	no	no	1	1	1	1		207	8	108-154	0.42	0.395	0.362	0.399	-0.035	C		x	x
VMC2A3	18	948,244	2	no	no	no	1	1	1	1		207	15	149-185	0.812	0.814	0.795	0.052	0	C+S		x	x
VMC8B5	18	3,362,208	2	no	no	no	1	1	1	1		207	9	145-162	0.778	0.8	0.77	0.069	0.014	C		x	x
VMC2B1.1	18	3,752,073	2	no	no	no	1	1	1	2	1	207	6	90-105	0.42	0.499	0.448	0.302	0.094	C			x
VVIQ44	18	5,309,353	2	no	yes	no	1	1	2	1		207	10	185-220	0.787	0.824	0.798	0.056	0.02	C+S		x	x
VVIM93	18	6,962,288	2	no	no	no	1	1	1	1		207	13	107-151	0.783	0.791	0.772	0.062	0.005	C		x	x
VChr18a-2	18	8,512,550	1	no	no	no	1	1	1	1		207	11	176-234	0.768	0.783	0.754	0.075	0.009	C		x	x
VChr18a	18	8,512,593	1	no	no	no	1	1	1	2	1	207	10	133-192	0.671	0.757	0.723	0.093	0.06				
VMC8F4.2	18	10,120,948	1	no	no	no	1	1	1	1		207	7	92-113	0.628	0.646	0.586	0.185	0.023	C		x	x
UDV-005	18	11,025,140	2	no	no	no	1	1	1	1		206	8	184-219	0.65	0.785	0.75	0.081	0.095	C		x	x
UDV-117	18	13,168,914	2	no	no	no	1	1	1	1		206	12	135-167	0.772	0.809	0.783	0.061	0.025	C		x	x
VVIP08	18	13,796,392	2	no	no	no	1	1	1	1		207	2	263-267	0.507	0.488	0.368	0.382	-0.021	C+S		x	x
VChr18b	18	14,028,762	1	no	no	no	1	1	1	1		207	6	145-158	0.676	0.684	0.625	0.158	0.006	C		x	x
UDV-134-2	18	18,100,286	2	no	yes	no	1	1	2	1		207	13	258-331	0.734	0.713	0.685	0.109	-0.019	C+S		x	x
VVIN16	18	26,361,305	2	no	no	no	1	1	1	1		207	5	149-159	0.715	0.695	0.644	0.144	-0.019	C+S	x	x	x
A105	18	28,306,457	1	no	no	no	1	1	1	1		207	3	266-272	0.353	0.38	0.31	0.454	0.037	C+S		x	x
p3_VvAGL11	18	30,313,541	2	no	no	yes	1	1	2	1		207	8	175-197	0.609	0.593	0.57	0.189	-0.03	C		x	x
VMC7f2-2	18	30,313,811	2	no	no	no	1	1	1	1		207	6	122-132	0.691	0.641	0.602	0.168	-0.047	C		x	x
VChr18c	18	33,971,369	2	yes	no	no	1	1	2	1		207	5	126-134	0.271	0.276	0.261	0.54	0.016	C+S		x	x
VVIN74	19	213,243	2	no	no	no	1	1	1	1		207	9	279-299	0.691	0.746	0.706	0.104	0.036	C+S		x	x
VChr19a	19	977,575	1	no	no	no	1	1	1	2	1	207	9	121-150	0.758	0.755	0.721	0.093	-0.001				
VChr19a-2	19	977,575	1	no	no	no	1	1	1	1		207	9	188-215	0.754	0.756	0.722	0.093	0.002	C		x	x
UDV-023	19	2,346,454	2	no	no	no	1	1	1	1		207	9	184-223	0.778	0.8	0.768	0.071	0.013	C		x	x
VMC5E9	19	4,183,125	2	yes	no	no	1	1	2	1		207	16	193-230	0.874	0.867	0.851	0.032	-0.005	C+S		x	x
B003	19	5,927,868	1	no	no	no	1	1	1	2	2	207	10	159-230	0.821	0.785	0.752	0.079	-0.029	C		x	x
VVIP31	19	6,697,524	2	no	no	no	1	1	1	1		207	13	172-196	0.923	0.893	0.881	0.022	-0.017	C+S	x	x	x
VVIM03	19	7,078,556	2	no	no	no	1	1	1	2	1	207	10	361-382	0.502	0.708	0.665	0.127	0.174				
VVIM03-2	19	7,078,556	2	no	yes	no	1	1	2	1		207	11	224-245	0.667	0.67	0.631	0.148	0.003	C		x	x
UDV-046	19	7,755,770	1	no	no	no	1	1	1	1		207	6	243-260	0.681	0.643	0.577	0.193	-0.025	C		x	x
VChr19b	19	8,693,674	1	no	no	no	1	1	1	2	1	207	5	157-171	0.517	0.719	0.669	0.128	0.166	C		x	x
RTAChr19-11b	19	11,230,059	3	no	no	no	1	1	2	1		206	14	164-202	0.85	0.877	0.863	0.028	0.016	C+S		x	x

Marker			Quality of amplification profiles							Checking		Genetic diversity								Use			
name	Chromosome	Position (bp)	Stuttering pattern	+A effect	Low heterozygote peak height ratio	Artifacts	Bins separation	Multilocus	Ease of Scoring	Allele inheritance	Causes of incompatibility	N	K	Allele size range (bp)	H _{obs}	H _{exp}	PIC	PI	F	Accessions genotyping	Q/K analysis	LD analysis	Association analysis
VVIV33	19	14,690,189	2	no	no	no	1	1	1	1		207	10	339-381	0.816	0.823	0.797	0.056	0.003	C	x	x	x
RTAChr19-16b	19	18,589,198	2	no	no	no	1	1	1	2	1	202	12	340-381	0.53	0.782	0.753	0.075	0.189	C+S	x	x	x
VVIB54	19	20,755,629	1	no	no	no	1	1	1	1		207	1	197	0	0	0	1	ND	C		x	x
UDV-062-2	19	21,594,816	2	no	no	no	1	1	1	1		207	11	219-261	0.71	0.717	0.667	0.129	0.002	C	x	x	x
UDV-062	19	21,598,416	2	no	no	no	1	1	1	2	1	206	6	114-146	0.515	0.688	0.635	0.15	0.145				
VMC7B1	19	23,628,578	3	yes	no	no	1	1	2	1		207	17	210-249	0.768	0.836	0.82	0.041	0.041	C	x	x	x

Stuttering pattern: 1: no or little stuttering, 2: regular stuttering, 3: irregular and/or excessive stuttering. **+A effect** refers to irregular adenylation and/or split peaks. **Low heterozygote peak height ratio** refers to the large differences observed between alleles peak height in heterozygous individuals. **Bins separation:** refers to the magnitude of size difference between successive bins or alleles, 1: clearly separated, 2: not clearly separated (differences of 1bp), 3: unstable and/or not clearly separated (differences of 1bp). **Multilocus:** amplification of multiple loci by the same primer pair, 1: monolocus, 2: multilocus. **Ease of scoring:** 1: easy, 2: difficult, 3: medium. **Allele inheritance:** check of allele inheritance in trio pedigrees; 1: compatible, 2: incompatible, which means that incompatible genotypes were identified in at least one pedigree. **Causes of incompatibility (most likely):** 1: null alleles 2: somaclonal mutations. Diversity parameters were calculated only from the genotypic data of the core collection (207 accessions). **N:** number of successfully genotyped individuals over the core collection, **K:** number of alleles, **H_{obs}:** observed heterozygosity, **H_{exp}:** expected heterozygosity, **PIC:** polymorphic information content, **F:** estimated frequency of null alleles, **Q/K:** population structure/relatedness, **LD:** linkage disequilibrium analysis, **GWAS:** genome wide association analysis.

Figure A.2. LD decay Plots across the grapevine genome as estimated using a set of 25,726 SNPs over 242 genotypes. Red dotted lines represent the logarithmic trend line.

Figure A.3. LD decay plots across the grapevine genome as estimated using a set of 250 SSRs over 207 genotypes. Red lines depict the logarithmic trend curves.

Mean number of seeds per Berry-2010

Mean number of seeds per Berry-2012

Mean seed dry weight per berry-2010

Figure A.4. Genome-wide association profiles obtained using four analysis models for seed content traits in 2010 and 2012

The identifier of the analysis model is shown at the upper-right side of each plot. The 19 grapevine chromosomes are identified in the legend at the bottom of each panel. The SNPs assigned to the last column of each plot (identified as “chr20”) correspond to the 433 SNPs that were located in unmapped contigs. The genome-wide significance threshold is $P < 1.4 \times 10^{-6}$ ($-\log_{10}(P) = 5.856$).

Figure A.5. Genome-wide association profiles obtained using the Q+K model for seven grape traits across three growing seasons

Only the traits for which no significant associations were detected using the Q+K model are presented here. Each trait is represented in a single line of three plots that correspond respectively from the left to the right to the growing seasons 2010, 2012 and 2013. Exceptionally, there were no data to present for veraison time in 2010. The 19 grapevine chromosomes are identified in the legends. The SNPs assigned to the last column of each plot (identified as “chr20”) correspond to the 433SNPs that were located in unmapped contigs. The genome-wide significance threshold is $P < 1.4 \times 10^{-6}$ ($-\log_{10}(P) = 5.856$).

Figure A.6. GLM model-based association profiles for 16 grape traits across three growing seasons

Each trait is represented in a single line of three plots that correspond respectively from the left to the right to the growing seasons 2010, 2012 and 2013. Exceptionally, for veraison and ripening times, there were no data to present for the year 2010. For chromosomes identifiers, see Fig.1 above. The genome-wide significance threshold is $P < 1.4 \times 10^{-6}$ ($-\log_{10}(P) = 5.856$).

Table A.8. Positions and *P* values of all the significant trait-SNP associations identified on chr1 using Q+K model

SNP position	<i>P</i> value					
	MNS-10	MNS-12	MNS-13	MSDW-10	MSDW-12	MSDW-13
5,789,433				9.64E-07		
5,789,434				9.64E-07		
6,289,513			4.61E-07			
12,796,528	7.17E-07	2.23E-07	9.52E-09	1.38E-06		
17,686,234		4.96E-07	7.31E-08			
17,932,489	6.9E-07	8.71E-08	4.06E-09	1.15E-06		
19,361,896		1.49E-07	1.45E-08			
19,614,285		7.59E-07				
19,854,230	1.22E-06	5.21E-09			2.75E-07	
20,294,168		8.9E-08				
20,329,648		1.97E-08			8.23E-07	
20,341,178		5.38E-08				
20,997,589			2.65E-10	3.12E-07	3.04E-07	3.26E-07
21,032,006			4.3E-08	5.88E-07	7.15E-07	
21,032,027			4.07E-08	5.89E-07	7.22E-07	
21,057,398			1.21E-06	8.49E-07	7.03E-07	
21,075,421			5.17E-08			
21,948,581			8.88E-09			
22,315,931			3.19E-08			
22,315,946			3.19E-08			
23,238,804			5.21E-09		9.55E-07	1.11E-06

MNS: mean number of seeds, MSDW: mean seed dry weight, the endings -10, -12 and -13 refers to the growing seasons 2010, 2012 and 2013 respectively.

Table A.9. Positions and *P* values of all the significant trait-SNP associations identified on chr2 using Q+K model

SNP position (bp)	<i>P</i> value		
	Berry color	Flower sex	Muscat flavor
2,711,805			3.03E-08
2,711,806			3.03E-08
2,811,281			2.94E-07
2,947,880			2.43E-07
2,974,643			4.43E-09
3,508,212			1.24E-07
3,508,233			2.08E-07
3,629,148			4.21E-07
4,109,392		1.10E-06	
4,567,920		1.98E-07	
4,629,101		1.32E-12	
4,699,221		1.37E-08	
4,726,583		1.26E-07	
4,726,746		1.31E-10	
4,761,375		1.91E-16	
4,793,688		1.25E-06	
4,848,223		7.89E-10	
7,597,669	1.52E-07		
9,022,191	9.70E-07		
13,264,449	1.01E-06		
13,534,426	7.92E-07		
13,916,989	1.36E-06		
13,917,054	4.77E-07		
14,090,984	2.81E-10		
14,108,150	1.06E-10		
14,295,837	8.70E-12		
14,300,736	1.07E-06		
14,391,735	1.72E-12		
14,428,045	6.88E-23		
14,428,146	1.36E-12		
14,934,673	7.27E-08		
14,936,915	3.54E-23		
14,936,921	2.59E-09		
14,970,720	3.35E-12		
15,014,545	9.03E-17		
15,014,566	1.93E-28		
15,014,571	9.03E-17		
15,014,599	2.41E-16		
15,291,548	4.92E-09		
15,380,330	7.42E-13		
15,380,343	8.11E-13		
15,525,457	8.34E-26		
15,748,503	3.02E-26		
15,954,285	5.40E-15		

SNP position (bp)	<i>P</i> value		
	Berry color	Flower sex	Muscat flavor
16,373,294	5.35E-08		
16,654,145	1.86E-10		
16,659,997	3.52E-19		
16,660,076	4.12E-19		
16,660,145	1.37E-07		
16,664,554	5.09E-08		
16,994,368	1.90E-15		
17,001,371	3.38E-13		
17,017,127	1.34E-07		
17,017,153	5.96E-07		
17,454,742	2.34E-08		
17,631,970	5.04E-09		
17,632,084	1.50E-08		
17,678,681	3.69E-09		
17,722,108	2.53E-07		
17,723,411	3.65E-15		
18,012,029	1.39E-07		
18,193,649	5.22E-07		
18,193,667	1.71E-07		
18,193,679	1.71E-07		
18,193,737	1.09E-07		
18,193,745	3.05E-08		
18,329,553	6.35E-07		
18,514,381	4.38E-08		
18,514,385	4.38E-08		
18,514,388	4.38E-08		
18,514,390	4.38E-08		
18,514,473	1.81E-10		
18,514,506	1.81E-10		
18,592,574	3.38E-07		

Table A.10. Positions and *P* values of all the significant trait-SNP associations identified on chr5 using Q+K model

SNP position (bp)	<i>P</i> value						
	Mst	MCW-13	MBW-10	MBW-12	MBW-13	MBL-10	FI-12
735,377	7.3245E-07						
1,543,067	1.4845E-07						
2,202,620	4.6899E-09						
2,523,895	4.4905E-09						
2,852,619	2.5414E-08						
3,068,001	7.0716E-10						
3,127,078	3.4766E-07						
3,167,686	2.6882E-07						
3,585,011	2.5124E-08						
4,431,496	1.1236E-06						
4,913,499					7.7724E-07		
4,913,506					7.7724E-07		
8,734,944				2.2578E-08	6.4677E-08		
14,046,799			4.7273E-07		1.676E-07		9.114E-07
14,327,183					2.1695E-07		4.9472E-07
16,010,197						1.1891E-06	
16,010,204						1.1794E-06	
20,051,857		1.2903E-06					
22,075,614				1.2357E-06			
22,075,642				1.2357E-06			

Mst: muscat flavor, MCW: mean cluster weight, MBW: mean berry weight, MBL: mean berry length, ShI: berry shape index, FI: flesh firmness index, the endings -10, -12 and -13 refers to the growing seasons 2010, 2012 and 2013 respectively.

Table A.11. Positions and *P* values of all the significant trait-SNPs associations identified on chr8 using Q+K model

SNP position (bp)	<i>P</i> value			
	CI-12	MBW	MBL	ShI
3,405,149			7.05E-07	
3,452,355	5.88E-07			
10,371,958				3.62E-07
20,789,025		2.97E-07		
21,184,639			1.30E-06	

CI-12: cluster compactness index according to Tillo et al (2014), MBW: mean berry weight, MBL: mean berry length, ShI: berry shape index.

Table A.12. Positions and *P* values of all the significant trait-SNP associations identified on chr17 using Q+K model

SNP position (bp)	<i>P</i> values					
	MNS-12	MNS-13	MSDW-10	MSDW-12	MBW-13	MBD-13
6,256,364		3.91E-07			7.283E-08	1.384E-08
6,256,381		3.91E-07		4.99E-08	7.283E-08	1.384E-08
6,256,402		3.91E-07		4.99E-08	7.283E-08	1.384E-08
6,256,405		3.91E-07		4.99E-08	7.283E-08	1.384E-08
6,256,518		1.1E-06		1.41E-07	1.987E-07	4.596E-08
6,366,721				9.57E-07	2.67E-07	2.997E-07
6,914,850				1.15E-06		
8,916,863		8.45E-07	1.19E-07			
9,201,088	5.66E-07	4.63E-07				
9,201,093	5.66E-07	4.63E-07				
9,294,776	1.14E-08	4.18E-07		1.10E-07		
9,739,079	3.25E-07	1.29E-06		2.55E-07		
9,739,112			1.10 E-06			
9,970,623	1.01E-06			7.28E-07		
10,741,556	5.41E-07			4.26E-07		

MNS: mean number of seeds, MSDW: mean seed dry weight, MBD: mean berry diameter, MBW: mean berry weight, the endings -10, -12 and -13 refers to the growing seasons 2010, 2012 and 2013 respectively.

Table A.13. Positions and *P* values of all the significant trait-SNP associations identified on chr18 using Q+K model

SNP position (bp)	<i>P</i> values							
	MNS-10	MNS-12	MNS-13	MSDW-10	MSDW-13	MBD-10	MBD-12	MBD-13
4,859,533							2.9E-07	
4,914,521							2.07E-07	1.19E-06
5,079,438						2.33E-07	2.3E-07	
5,079,501						3.6E-07	5.82E-07	
6,697,371						4.8E-07	3.86E-07	
10,065,456	2.85E-07		1.45E-09	3E-07	1.9E-07			
10,177,341	1.08E-06		6.73E-09		1.28E-06			
24,729,405			3.39E-07					
29,343,740			2.69E-07					
30,234,060			2.13E-07					
32,027,523	7.22E-08	1.93E-08	1.8E-11					
32,125,496	3.89E-09	5.41E-08	1E-11	3.64E-08	8.63E-08			
32,712,250	1.25E-07	1.36E-07	1.45E-11	5.4E-07				
32,712,280	1.47E-07	2.26E-08	9.55E-11					

MNS: mean number of seeds, MSDW: mean seed dry weight, MBD: mean berry diameter, the endings -10, -12 and -13 refers to the growing seasons 2010, 2012 and 2013 respectively.

Table A.14. List of new unstable genetic associations detected by SSR markers in comparison with SNP association results

Trait	Chr	SSRs					SNPs					
		marker	position	P value			2010		2012		2013	
				2010	2012	2013	position	P value	position	P value	position	P value
Bb	1	FAM141	17,732,854	0.00118		0.00376	17932477	0.00196				
	11	VVIB19	5,865,805	0.00892			6064867	0.00971	5471210	1.40E-04		
Fl	1	VVIF52	23,292,680			0.00131	23,821,326	0.00317	22,931,400	0.00744		
	4	VMC8B11	17,682,875			0.00162	17,150,245	0.00609	17,140,062	0.00955		
	15	VVIQ61	11,761,810		0.0053						11,212,733	0.00879
	17	VRZAG15	6,588,595	0.00597			6,936,420	0.00318	6,184,504	0.00222	6,936,420	6.42E-04
	17	VChr17a	9,315,344	0.00681			9,332,071	0.00961	9,295,637	6.18E-04	9,295,664	1.84E-04
	17	VVIV08	12,161,420			6.71E-04	12,687,319	0.00135			11,670,422	0.00974
	18	VVIQ44	5,309,353	0.00483			4,737,616	0.00244			5,820,550	0.00396
	18	UDV-134-2	18,100,286			0.00385	17,425,242	0.00124			18,139,390	0.00296
	18	VVIN16	26,361,305			3.44E-04			26,392,298	0.00245	26,392,453	0.00489
	3	RTAChr3-15a	15,937,188		0.00768						15,425,799	0.0081
	10	VVIH01	1,473,028			0.00717	2,066,641	0.00107	620,749	7.22E-04	1,285,480	0.00162
	14	VVC34	13,207,677			0.00322	13,362,814	0.00243			12,832,703	0.00496
	18	VMC8F4.2	10,120,948		0.00537		10,064,577	2.32E-04	10,064,577	8.59E-04	10,064,577	0.00135
Rp	1	VVIO02	9,813,247			0.00961			9,871,357	0.00214	9,471,988	5.10E-04
	1	VMCNG2g7	6,993,233			0.00832					7,563,590	0.00141
	1	VVIM25	12,617,249			3.08E-04					12,608,150	7.28E-04
	3	UDV-021-2	5,287,429			5.97E-05					5,872,174	2.97E-05
	4	VMC8B11	17,682,875			4.63E-07					17,960,693	4.68E-03
	4	VVIP77	19,684,609			6.89E-04					19,974,850	6.06E-05
	4	VRZAG83	20,627,004			0.0079					20,708,463	2.11E-06
	5	UDV-053	8,541,098			1.68E-08					8,866,891	4.82E-06
	5	FAM122	15,723,310			0.00201					15,840,776	4.42E-06
	5	VRZAG89	22,919,299			0.00795					22,286,687	3.27E-04
FSx	6	VMC2H9	5,819,480			0.00526					5,732,869	1.54E-05
	8	VMC1B11	14,039,245			0.00405					14,009,096	3.81E-08
	8	VMC1E8	17,261,639			1.62E-04					16,801,728	4.05E-04
	8	VVIB66	18,270,246			0.00111					18,437,860	0.00398
	8	UDV-026	21,473,580			0.00849					21,357,940	2.33E-05
	9	VMC6d12	3,792,236			9.41E-04					4,271,085	5.36E-06
	10	FAM45	11,414,249			0.00208						
	13	FAM62-2	16,059,252			2.29E-04						

Trait	Chr	SSRs					SNPs					
		marker	position	P value			2010		2012		2013	
				2010	2012	2013	position	P value	position	P value	position	P value
	14	VVC34	13,207,677			2.76E-04						
	14	VRZAG112	22,042,800			1.96E-04					21,836,162	6.64E-05
	15	RTAChr15-1b	1,167,751			1.05E-06					1,566,698	2.74E-04
	16	UDV-104	10,785,478			0.00812					10,369,129	9.94E-05
	16	VMC4B7-2	23,329,085			0.00287					22,356,804	0.00275
	17	VRZAG15	6,588,595			1.56E-05					7,637,587	1.42E-08
	18	VVIQ44	5,309,353			0.00997					4,586,117	5.89E-04
	18	VMC8F4.2	10,120,948			1.62E-04					9,684,380	6.01E-06
	18	p3_VvAGL11	30,313,541			0.00213					29,644,844	3.81E-04
	18	VMC7f2-2	30,313,811			0.00137					29,644,844	3.81E-04
	18	VChr18c	33,971,369			0.0073					34,335,091	6.13E-04
	19	VVIN74	213,243			0.00882						
	19	UDV-062-2	21,594,816			1.64E-07					20,806,512	8.27E-05
	19	VMC7B1	23,628,578			1.59E-04					22,968,759	7.71E-06
Frt	4	VVIQ25-2	2,036,638	0.00163			2,194,163	4.53E-04	1,720,476	0.00139		
	4	VMC8B11	17,682,875			0.00656	17,721,820	0.00211			17,748,286	0.00162
	4	VRZAG83	20,627,004		0.00168				21,025,981	0.00879	20,280,269	0.00422
	6	UDV-090	1,012,094		0.00346				709,089	0.00337	687,574	0.00618
	6	FAM110	3,652,483		0.00132				3,857,229	0.00934	3,838,482	0.0064
MCW	4	VMC8B11	17,682,875			0.0065	17,547,365	0.0081	17,140,100	0.0026	17,030,682	0.00442
	4	VVIP77	19,684,609		0.00535				19,622,380	0.00967		
	9	VVIQ66	10,808,981			0.00411					10,566,841	0.00787
	14	UDV-095	26,102,129			0.00699			25,355,123	0.00735	25,641,223	6.95E-04
	18	p3_VvAGL11	30,313,541			0.00787	30,861,327	0.00495	30,861,327	0.0044		
	19	VVIM03-2	7,078,556			0.00266	6,628,738	0.00221	6,831,801	0.00162	6,782,152	1.39E-04
MCL	1	VChr1b	11,850,801	0.004								
	5	VRZAG79	5,692,460			0.00502			5,461,779	1.02E-04	6,187,373	1.38E-04
	5	VMC4C6	25,493,869			0.0073			25,079,253	0.00138	25,336,619	0.0075
	6	VVIP72-2	2,280,007			0.00164			2,950,270	0.0043	1,668,240	0.00791
	8	VChr8a-2	15,200,499			0.00923	15,166,087	0.002	14,750,609	0.00383	14,750,609	0.0063
	12	VChr12a	6,174,660	0.00462			5,837,639	0.00637	5,837,639	8.63E-05	5,140,615	0.00751
	15	VMC8G3-2	17,666,601			0.00512			18,512,080	0.00974	18,248,447	5.66E-04
	18	VChr18b	14,028,762		0.00206		13,727,668	0.00913			13,727,668	0.00401
	19	VVIM03-2	7,078,556			2.01E-04			7,059,538	2.08E-04		
	19	UDV-046	7,755,770			1.74E-04	7,756,361	0.00115				

Trait	Chr	SSRs					SNPs					
		marker	position	P value			2010		2012		2013	
				2010	2012	2013	position	P value	position	P value	position	P value
MCD	5	VMC4C6	25,493,869	0.00569			25,183,882	0.00325			24,906,298	0.00489
	13	FAM62-2	16,059,252			0.00532			15,518,582	0.00227	15,175,477	0.00107
	13	VMC8E6	27,180,657			0.00669			27,455,358	0.00481	26,797,677	0.00101
	14	UDV-050	4,943,488			0.0043	5,737,558	4.60E-04	4,291,492	0.00704		
	18	VMC2B1.1	3,752,073	0.00221			3,792,894	0.00111	3,792,894	5.16E-04	3,644,906	3.27E-04
	18	p3_VvAGL11	30,313,541	0.00485					30,861,327	0.00664		
	18	VMC7f2-2	30,313,811	0.00221					30,861,327	0.00664		
MPL	12	TT251F02-2	418,506		0.0032		159,331	0.004				
	15	RTAChr15-7a	8,439,107			0.00103	8,340,490	0.00606			8,524,574	0.00801
	18	VMC7f2-2	30,313,811			0.00244	30,201,598	0.0023				
CI-OIV	4	VMC4D4	3,753,977	0.0062			4,501,530	0.00372				
	14	VMC2H12	10,303,957	0.0039					10,295,462	0.0079		
	18	A105	28,306,457			0.00747					28,413,360	0.00266
CI-12	1	VChr1b	11,850,801			0.00897	11,127,190	1.21E-04	11,236,250	0.00158	12,090,586	0.00722
	3	VVIN54	11,093,607			0.00166	11,054,835	4.69E-06	3	VVIN54	11,093,607	
	4	VMC4D4	3,753,977	0.00473			3,284,361	2.35E-04	3,692,622	0.00944		
MBW	19	VVIN74	213,243		4.44E-04		312,975	4.36E-04	136,792	3.62E-04	502,566	1.16E-04
MBL	1	VChr1a-2	2,819,297	0.00408			3,040,405	5.25E-05	3,040,462	7.50E-04	2,878,002	0.00341
	2	FAM140-2	9,166,787	0.00435					8,441,882	0.00283	8,857,385	1.85E-04
	3	RTAChr3-17a	18,564,018	0.00703			18,702,260	0.00723	18,877,053	0.00245	18,748,644	0.0097
	9	VVIQ52	21,558,043		0.00498		21,092,627	5.59E-04	21,407,835	0.00551	21,882,366	0.00954
	10	FAM45	11,414,249			2.75E-04	11,79155	7.49E-04	10,520,743	0.00767	11,379,767	0.00196
	14	FAM135	8,755,200		0.00853		7,861,496	6.84E-04	8,532,838	0.00322	7,842,269	0.00784
MBD	2	VMC7g3	18,382,886	0.0048			18,319,787	0.00121	18,319,787	0.00226		
	4	Vchr4a	946,454	0.0081			169,985	0.00138	169,985	3.99E-04	169,985	9.98E-05
	4	VMC4D4	3,753,977			0.00582	3,338,266	0.00667	4,091,585	0.00341		
	8	VMC5H2	15,407,796			0.00142	15,994,442	0.00905	15,980,963	5.59E-04	15,980,990	1.29E-04
	9	VMC1C10	563,536			0.00672	201,966	1.72E-04	201,966	3.56E-04	624,691	0.00109
	9	FAM42	7,551,193			0.0079	7,607,178	0.00172	6,631,567	0.00162	6,631,567	1.03E-04
	10	FAM45	11,414,249			0.00229			10,520,743	0.00725	10,520,743	0.00484
	13	VVIN62	1,387,838			0.00107	942,614	0.00346	1,685,635	0.00516	1,552,906	0.00148
	19	B003	5,927,868			0.00811	6,177,682	0.00982	6,302,477	0.00787		
ShI	3	UDV-021-2	5,287,429	9.50E-04			5,345,915	8.94E-06	5,345,915	8.22E-04	5,586,982	0.00142
	3	RTAChr3-17a	18,564,018		0.00741		19,185,538	4.92E-04			18,786,805	0.00279
	7	FAM116	20,388,473	0.00541			20,565,562	0.00594	20,565,562	0.00663	20,565,562	0.00391

Trait	Chr	SSRs					SNPs					
		marker	position	P value			2010		2012		2013	
				2010	2012	2013	position	P value	position	P value	position	P value
FI	9	RTAChr9-13a	13,102,063			<i>0.00835</i>	14,035,676	0.00194	14,035,676	0.002		
	9	VVIQ52	21,558,043	<i>0.00958</i>								
	10	UDV-059	9,016,614			<i>0.00993</i>	8,454,460	4.71E-04	9,987,141	0.00657	8,808,953	<i>3.31E-05</i>
	16	VChr16b	17,432,683			<i>0.00646</i>					17,577,715	1.29E-04
	18	VMC3E5	321,045			<i>5.05E-04</i>	413,869	1.77E-04	389,123	4.53E-04	274,754	0.00208
	13	VVIM01	6,609,298			<i>0.00896</i>	6,870,567	0.00575			6,870,567	9.66E-05
Mst	14	VVIP05	2,861,324			<i>0.00238</i>					3,070,674	0.00176
	14	VVC34	13,207,677			<i>0.00554</i>						
	14	UDV-095	26,102,129			<i>2.81E-04</i>					26,080,069	2.83E-04
	14	VVIN70	29,853,616			<i>0.00178</i>					28,844,127	0.00573
	19	UDV-046	7,755,770			1.22E-04						
	19	UDV-062-2	21,594,816			<i>4.06E-04</i>					20,913,427	0.00105
MNS	16	VChr16b	17,432,683			<i>0.00338</i>	18,331,153	0.00263	16,506,010	0.007	16,505,997	0.00589
MSDW	15	RTAChr15-3b	2,650,923	<i>0.00931</i>					2,845,721	0.0076		
	15	VVIP33	13,060,296			<i>0.00522</i>	12,383,947	6.06E-04	12,566,432	0.00267	12,895,330	4.16E-04

See Table 1 for trait names. The SSR associations shown here were detected using the Q+K model at the significance threshold of 0.01. The SNP associations represent the maximum *P* value found in a radius of 1 Mb around the position of the corresponding SSR association. Bold *P* values highlight the significant associations at the Bonferroni corrected threshold (1.90E-03 for the SSRs and 1.4E-06 for the SNPs). Italic *P* values highlight the significant associations at 0.01 for the SSRs and at the equivalent one for the SNPs 7.30E-05 (both of them are 53 times more relaxed than the respective Bonferroni corrections).