# Génomique comparative et fonctionnelle de familles de gènes liés au métabolisme secondaire de la vigne (Vitis vinifera) et de ses proches parents

Gautier Arista

# THÈSE

Présentée à l'Université de Strasbourg
Pour l'obtention du titre de Docteur d'Université
Discipline : Aspects moléculaires et cellulaires de la biologie

**Gautier ARISTA**

# Génomique comparative et fonctionnelle de familles de gènes liés au métabolisme secondaire de la vigne (*Vitis vinifera*) et de ses proches parents.

Soutenue le 31 janvier 2017 devant le Jury :

| | |
|---|---|
| Rapporteurs : | P. FAIVRE-RAMPANT, Chargée de recherche INRA, Evry |
| | E. PAUX, Directeur de recherche INRA, Clermont-Ferrand |
| Examinateurs : | H. SCHALLER, Directeur de recherche CNRS, Strasbourg |
| | V. GEFFROY, Directrice de recherche INRA, Paris-Saclay |
| Encadrants : | P. HUGUENEY, Directeur de recherche INRA, Colmar |
| | C. RUSTENHOLZ, Maître de conférences Université de Strasbourg |

Financée par :

# Résumé

La vigne (*Vitis vinifera*) possède un métabolisme secondaire particulièrement riche donnant naissance à une large palette de molécules dont certaines sont impliquées dans les défenses contre les pathogènes et d'autres dans la grande diversité d'arômes qui fait la renommée des vins. L'analyse de la séquence de référence du génome de la vigne a permis de mettre en évidence une remarquable expansion de certaines familles de gènes liés au métabolisme secondaire par rapport aux autres plantes. Dans ce travail, j'ai étudié les familles gènes codant pour les *cytochromes P450*, dont certains sont impliqués dans la production d'arômes, les gènes codant pour les stilbènes synthases (*STS*), les endo-β-1,3-glucanases et les gènes de résistance de type *NBS* impliqués dans les défenses de la vigne. Ma thèse vise à proposer des hypothèses expliquant l'organisation structurale de ces familles de gènes et ainsi à mieux comprendre pourquoi certaines familles présentent une amplification dans le génome de la vigne.

Des approches bioinformatiques ont été utilisées afin d'étudier ces différentes familles de gènes. Les gènes *cytochromes P450* et gènes R de type *NBS* ont tout d'abord été annotés de manière manuelle dans le génome de référence de la vigne. L'expression des gènes endo-β-1,3-glucanases, *STS* et *cytochromes P450* a été analysée en utilisant une approche transcriptomique à grande échelle. Pour ce faire, un outil a été développé durant cette thèse pour estimer le niveau d'expression des gènes à partir de données RNA-Seq disponibles dans les banques de données publiques. Parallèlement, des données de reséquençage d'ADN de 56 cépages et espèces de vigne ont été analysées, afin de déterminer les variations structurales de type CNV au sein des familles de gènes à domaine *NBS* et de gènes *STS*.

Ces différents travaux ont permis de montrer que l'amplification des familles de gènes étudiées n'est pas spécifique du génome de référence mais est retrouvée dans l'ensemble du genre *Vitis*, mais également de mettre en évidence des variations structurales au sein des différents génomes étudiés. L'analyse de la famille *STS* a montré que ces gènes sont organisés en blocs de duplication, et que les gènes plus conservés sont aussi les plus exprimés. Nous avons également montré que les gènes à domaine *NBS* sont organisés en cluster, dont certains sont particulièrement soumis à variation. Ces travaux contribuent à une meilleure connaissance de facteurs de défense efficaces et durables ainsi que des gènes impliqués dans la synthèse d'arômes dans la vigne. Ces connaissances pourront bénéficier aux programmes de création variétale mis en œuvre à l'INRA de Colmar.



**Mots clés :** vigne, familles de gènes, génomique comparative, transcriptomique, CNV

# Abstract

Grapevine (*Vitis vinifera*) has a particularly rich secondary metabolism, giving rise to a wide range of molecules, some of which are involved in defences against pathogens and others in the great diversity of aromas that make wines famous. Analysis of grapevine reference genome has shown a remarkable expansion of certain families of genes linked to secondary metabolism in comparison with the other plants. In this work, I have analysed gene families coding for *cytochromes P450*, some of them being involved in the production of aromas, genes coding for stilbene synthases (*STS*), endo-β-1,3-glucanases and *NBS* type resistance genes involved in grapevine defences. My thesis intends to propose hypothesis to explain the structural organisation of these families and therefore better understand why some of these families are amplified in the grapevine genome.

Bioinformatic approaches have been used to study these different genes families. The *cytochromes P450* and R genes of *NBS* type were manually annotated to improve the knowledge of these families of genes. The expression of endo-β-1,3-glucanases, *STS* and *cytochromes P450* genes has been quantified using a large-scale transcriptomic approach. To this purpose, a tool has been developed during this thesis to estimate the level of genes expression from RNA-Seq data available in public databases. In the meantime, DNA resequencing data from 56 cultivars and grapevine species have been analysed to identify structural variations of CNV types within the genes with a *NBS* domain and the *STS* genes.

These works showed that the amplification of the gene families of interest was not specific to the reference genome but occurred at the scale of the *Vitis* genus, but also to highlighted structural variations in different genomes. Regarding the *STS* genes, blocks of duplication and more conserved and expressed genes were identified. For the genes with *NBS* domain, a clustered organisation has been highlighted with some clusters varying more than others in the studied genotypes. These works contribute to a better knowledge of gene families for efficient and durable defence against pathogens and optimal aromas synthesis in grapevine. This knowledge will benefit to breeding programs currently in progress at INRA Colmar.



**Key words:** grapevine, family of genes, comparative genomics, transcriptomics, CNV

*"Research is what I'm doing when I don't know what I'm doing"*

Wernher von Braun

# Remerciements

Je tiens tout d'abord à remercier Claire Parage qui, grâce à sa thèse, a permis de poser les bases de ce projet. Sans ses résultats, ce travail n'aurait jamais vu le jour. Je voudrais remercier l'INRA et la région Alsace qui ont financé cette thèse. Mais aussi, Camille Rustenholz et Philippe Hugueney qui ont décidé de me recruter après un entretien via webcam interposée, sans savoir si je portais un pantalon pour compléter le haut de mon costume et malgré la féroce concurrence d'une gagnante d'un concours de Miss pour le poste.

J'aimerais remercier Patricia Faivre-Rampant, Valérie Geffroy, Etienne Paux, Hubert Schaller, Camille Rustenholz, et Philippe Hugueney d'avoir accepté de faire partie de mon jury de thèse. Mais aussi, Raquel Tavares et Hadi Quesneville pour avoir participé à mon comité de thèse.

Je tiens tout particulièrement à remercier Camille Rustenholz qui, tout au long de ce projet, a fait preuve d'un encadrement exceptionnel. Merci pour tous ces moments, bons et moins bons, pour ces échanges et ces discussions plus ou moins enflammées, pour tes conseils, astuces et pistes pour me permettre d'avancer. Je ressors grandi de cette expérience et j'espère que toi aussi. Merci pour cette complicité et bonne entente dont, au final, peu de doctorants peuvent se vanter. Je ne peux qu'envier les ‘‘petits poussins'' qui restent sous ton aile (Guillaume et maintenant Sophie et Amandine) mais aussi les prochains avec qui tu partageras ton savoir, comme tu l'as fait avec moi, et qui réaliseront à quel point tu es formidable.

Merci à Philippe Hugueney, d'une part pour l'accueil au sein de l'équipe ‘‘Métabolisme Secondaire de la Vigne'' et, d'autre part, pour l'encadrement exemplaire, les discussions plus que fructueuses et le point de vue plus extérieur qui parfois manquait pour faire avancer ce projet. Merci aussi d'avoir instauré un groupe de tarot sur l'heure de midi, je ne compte plus les rois coupés, les petits volés et le nombre de rois que tu as mis dans tes écarts… Le tarot a permis de renforcer la cohésion et la bonne entente dans l'équipe et, bien que cela reste un plaisir, nous parvenons à rester raisonnable et à ne pas en abuser.

Bref, merci à vous deux pour l'aide inestimable apportée tout au long de la thèse et particulièrement sur les derniers moments qui se font toujours dans le rush malgré les précautions prises. Merci à toute l'équipe MSV pour l'accueil et la bonne ambiance au quotidien ainsi que les différents gâteaux et pâtisseries apparaissant aux pauses café. Tout particulièrement à Anne pour m'avoir permis de retourner en Belgique à plusieurs occasions.

Merci à Guillaume Barnabé pour sa collaboration dans ce travail mais aussi pour l'aide qu'il a apportée au fil du temps permettant la résolution de bugs (plus ou moins faciles à trouver, vive l'informatique), la découverte et l'utilisation de nouveaux outils et les échanges d'idées. Ce fût un honneur d'inaugurer pléthore de jeux de société avec David et toi à l'association et de jouer, rejouer et rerejouer, toujours dans l'ambiance et la bonne humeur, jusque pas d'heures. Vive Résistance même si nous sommes des espions…

Je voudrais aussi remercier Lauriane Renault et Sophie Blanc dont l'aide, bien que ponctuelle, n'en est pas moins appréciable. Toutes deux s'essayant à la bioinformatique mais qui, pour Sophie, est en train de devenir une nouvelle carrière. Je ne peux que t'encourager dans cette voie et te dire qu'avec du temps et de la patience, tu y arriveras. Mention spéciale à toi Sophie pour tous ces fous rires, ces moments complices qui suffisent à éclairer une grise journée et une pensée à tous ces Kinder Bueno que tu as éventrés.

Je tiens à remercier aussi Pere Mestre, Jean-François Chich, Tina Ilc et Danièle Werck-Reichhart pour m'avoir permis de travailler à leurs côtés en espérant qu'ils aient apprécié travailler avec moi autant que j'ai apprécié travailler avec eux.

Je voudrais continuer en remerciant tous les doctorants de l'INRA de Colmar. Aussi bien ceux qui ont terminé ou commencé leur thèse durant mon séjour. Merci pour tous ces moments et ces petites distractions parfois nécessaires. La bonne humeur générale régnant dans ce bureau efface facilement les quelques fois où la cohabitation n'a pas été facile. Je compte sur toi, Isabelle, pour faire preuve de self-control la prochaine fois que tu feras du shopping.

Une pensée pour Lise et David, le Ying et le Yang de la MSV enfin, de la ViVe maintenant. Vos aventures vont clairement me manquer. Lise, garde bien le trèfle à quatre feuilles que je t'ai offert, tu en as continuellement besoin. David, je crois que le proverbe ‘‘si tu n'existais pas il faudrait t'inventer’’ te résume à lui seul.

Enfin, je voudrais remercier les personnes travaillant à l'INRA de Colmar en général pour l'accueil et l'agréable séjour que j'ai pu passer. J'ai fait la connaissance de personnes exceptionnelles et cela fait toujours plaisir d'être capable de venir travailler avec le sourire car le cadre et l'atmosphère s'y prêtent. Merci à Régis, le cuistot du centre, qui nous chouchoute au jour le jour avec des plats exceptionnels et des desserts à tomber par terre. Ce n'est pas Sophie qui me contredira.

Un unique merci pour une personne unique, tu es rentrée dans ma vie durant cette thèse et j'espère que tu n'en sortiras jamais. Merci pour ton soutien et cette formidable aventure que nous vivons. Merci d'être toi, d'être là, tout simplement : je t'aime.

Je voudrais simplement terminer avec quelques citations de collègues qui résonneront dans ma tête pour encore quelque temps :

*''Donc tu fais de la bioinformatique, mais ça se passe comment en fait ? Tu mets quoi dans le manuscrit ? Parce que tu ne fais pas d'expériences si ?''*

En regardant mon écran sur lequel était ouvert un terminal Linux : *''Purée on dirait qu'on est dans Matrix !''*

*''Ah bon tu utilises un dictionnaire et tu aimes les livres ? Mais je pensais qu'un geek comme toi ça faisait tout sur ordinateur''*

# Abréviations

ADN : Acide DésoxyriboNucléique

bp : base pair

cDNA : complementary DeoxyriboNucleic Acid

CDS : Coding DNA Sequence

CNV : Copy Number Variation

CYP : Cytochrome P450

DNA : DeoxyriboNucleic Acid

FPKM : Fragments Per Kilo base of exon per Million fragments mapped

GEO : Gene Expression Omnibus

INRA : Institut National de la Recherche Agronomique

k : kilo

kb : kilo base pairs

Mb : mega base pairs

mRNA : messager RiboNucleic Acid

NBS : Nucleotide Binding Site

pb : paire de bases

QTL :Quantitative Trait Locus

RNA-Seq : RiboNucleic Acid Sequencing

SNP : Single Nucleotide Polymorphism

SRA : Sequence Read Archive

STS : STilbene Synthases

TF : Transcription Factor

UMR : Unité Mixte de Recherche

UV : Ultra Violet

# Table des matières

# Table des matières

**Avant-propos** ..... - 1 -

**Revue bibliographique** ..... - 3 -

**Revue bibliographique: Omic approaches to unravel to extraordinary diversity of grapevine** - 4 -

**Introduction** ..... - 4 -

**Genomics** ..... - 6 -

**Functional genomics** ..... - 11 -

**Proteomics** ..... - 21 -

**Metabolomics** ..... - 23 -

**Conclusion** ..... - 25 -

**References** ..... - 28 -

**Objectif de la thèse** ..... - 34 -

**Partie 1 : Création d'un outil d'analyse du transcriptome de la vigne** ..... - 36 -

**Introduction** ..... - 37 -

**Matériel et méthodes** ..... - 39 -

**Résultats** ..... - 41 -

**Conclusion** ..... - 55 -

**Perspectives** ..... - 56 -

**Références bibliographiques** ..... - 59 -

**Partie 2 : Analyse transcriptomique de la famille des gènes cytochromes P450 de la vigne** .... - 60 -

**Annotation, classification, genomic organization and expression of the *Vitis vinifera* CYPome** ..... - 62 -

**Abstract** ..... - 62 -

**Background** ..... - 63 -

**Results** ..... - 65 -

**Discussion** ..... - 76 -

**Material and methods** ..... - 79 -

**Acknowledgements** ..... - 84 -

**Accession Numbers** ..... - 84 -

**Authors' contributions** ..... - 84 -

**Supplemental information** ..... - 85 -

**References** ..... - 91 -

**Partie 3 : Étude des variations structurales de type CNV des gènes de résistance à domaine NBS chez la vigne** ..... - 97 -

**Evolutionary dynamics of the NBS resistance gene family from the grapevine reference genome to the *Vitis* genus** ..... - 99 -

**Abstract** ..... - 99 -

**Introduction** ..... - 100 -

**Materials and Methods** ..... - 102 -

**Results** ..... - 106 -

**Discussion** ..... - 118 -

**Supplementary data** ..... - 121 -

**References** ..... - 133 -

**Partie 4 : Génomique comparative et transcriptomique de la famille des gènes *STS* chez la vigne (*Vitis vinifera*) et ses proches parents** ..... - 137 -

**Comparative genomics and transcriptomics of the stilbene synthase gene family in grapevine (*Vitis vinifera*) and its wild relatives** ..... - 139 -

**Abstract** ..... - 139 -

**Introduction** ..... - 140 -

**Material & Methods** ..... - 142 -

**Results** ..... - 145 -

**Discussion** ..... - 153 -

**Conclusion** ..... - 156 -

**Supplementary data** ..... - 157 -

**References** ..... - 161 -

**Conclusion générale** ..... - 163 -

**Perspectives** ..... - 167 -

**Références bibliographiques générales** ..... - 174 -

# Avant-propos

La vigne est une plante millénaire d'abord trouvée sous forme sauvage avant d'être domestiquée, 3000 ans avant notre ère. La vigne cultivée (*Vitis vinifera*) a été introduite en France entre 1000 et 500 ans avant notre ère et est, depuis, devenue une part entière de l'économie et de la culture. La vigne représente un intérêt agronomique entre autre par la production de raisin, économique avec la production et l'exportation de vins et culturel dû aux nombreux cépages régionaux et caractéristiques à partir desquels de typiques vins sont produits. Cette plante est cependant sujette à de nombreuses maladies qui peuvent provoquer des dégâts très lourds sur les vendanges et les vignobles. Afin de lutter contre les pathogènes et leurs vecteurs, beaucoup de traitements phytosanitaires sont réalisés. Ainsi, bien que la viticulture ne représente qu'environ 3% de la surface cultivée en France, elle consomme près de 20% des produits phytosanitaires (Rapport de l'expertise par l'Inra et le Cernagref, 2013).

Dans le but de proposer des solutions pour une viticulture plus durable, de réduire la consommation de pesticides et fongicides et de limiter l'impact sur l'environnement, l'UMR 1131 ‘‘Santé de la Vigne et Qualité du Vin’’ conduit un programme de création variétale visant à obtenir de nouveaux cépages de vigne produisant des vins de qualité et étant dotés de résistances durables au mildiou (*Plasmopara viticola*) et à l'oïdium (*Erysiphe necator*). La compréhension des mécanismes de résistance de la vigne face à ses pathogènes est donc un enjeu important. Cela passe par l'étude des métabolites secondaires, le principal sujet d'étude de l'équipe ‘‘Métabolisme Secondaire de la Vigne’’ dans laquelle a été effectuée cette thèse. Ceux-ci sont effectivement impliqués aussi bien dans les défenses de la plante que dans la synthèse des arômes dans les baies et contribueraient à assurer la durabilité des résistances face aux pathogènes. Un exemple particulier de métabolite secondaire est le resvératrol dont les applications pharmaceutiques sont diverses et variées et qui joue un rôle dans le traitement et la prévention des maladies cardiovasculaires (Das & Das, 2010).

De par l'intérêt qu'elle suscite, de nombreux groupes de chercheurs étudient la vigne et différents de ses aspects. De la caractérisation des différentes espèces aux arômes spécifiques des divers cépages en passant par l'étude des génomes et les mécanismes de résistance, les disciplines impliquées sont diverses et variées. Cette plante a d'ailleurs été une des premières plante à fruits dont le génome a été séquencé (Jaillon *et al.*, 2007; Velasco *et al.*, 2007).

Dans cette première partie introductive est présentée une synthèse bibliographique reprenant une grande partie des travaux effectués sur vigne dans les différents domaines ''-omiques'' à savoir : génomique, transcriptomique, protéomique, métabolomique. Elle s'intéresse également aux différents outils disponibles et recense différents jeux de données disponibles dans les bases de données publiques avec un accent sur le RNA-Seq afin de faciliter leur accès par les chercheurs intéressés. Cette synthèse est rédigée en anglais car elle servira de base pour la rédaction d'un article, sous forme de revue bibliographique, qui sera soumis ultérieurement pour publication.

# Revue bibliographique

# Revue bibliographique: Omic approaches to unravel to extraordinary diversity of grapevine

## Introduction

Grapevine is culturally and economically the most valuable fruit crops in the world. It has been widely cultivated for thousands of years. Grapes can be used for fresh or dry consumption but are mainly transformed, through fermentation, into wine. Moderate consumption of wine has been associated to health benefits, due to its high polyphenol content. In particular, resveratrol has positive effects on vascular and cardiovascular functions (Li *et al.*, 2012; Gresele *et al.*, 2011), which led to the hypothesis of its involvement in the French paradox (Renaud and de Lorgeril, 1992). The history of viticulture has led to the generation of thousands of grape cultivars, which produce a tremendous diversity of grapes and wines. Indeed, today's grapevine varieties show a level of genetic diversity superior to human and comparable to maize (Myles et al., 2011). The molecular characterization of this enormous diversity has long been extremely difficult. However, the development of grapevine genomics through the availability of a reference genome has made this task considerably easier.

In 2007, Jaillon and his colleagues succeeded in the assembly of the first high quality sequence of a grapevine (*Vitis vinifera* spp *vinifera*) genome, using the nearly homozygous *Pinot Noir* PN40024. The current version of the reference genome sequence PN40024 12X.2 was then achieved with an increase in the sequencing depth (from 8X to 12X) and an improvement of the anchoring of the scaffolds (available at https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences). Since the first version V0 of the sequence annotation hosted at the Genoscope (gene ID starting with GSVIV), the Grape Genome Database hosted at CRIBI made available a V1 and a V2 version (gene ID VIT_XXsXXXgXXXXX) for which transcript isoforms were assembled (Vitulo *et al.*, 2014). These annotations were transferred on PN40024 12X.2 and are available at the URGI Grape Genome JBrowse (https://urgi.versailles.inra.fr/jbrowse/gmod_jbrowse/?data=myData/Vitis/data_gff).

To avoid further problems in feature identification and naming, a standard nomenclature, a gene naming system and a common annotation platform were initiated by Grimplet and coworkers in 2014. This has helped the whole grapevine research community to communicate efficiently within the framework of the International Grape Genome Program (IGGP).

Since its publication, the availability of the grapevine reference sequence has opened new perspectives for the scientific community working on grape and has raised grape to the status of model plant for other fruit crops. The availability of this tool was the major starting point for the boom in grape ‘‘-omic’’ approaches. By definition, an ‘‘-omic’’ approach is an exhaustive study of the genome and the annotated features (genomics), of the transcribed fraction of the genome (transcriptomics), of the proteins in various tissues (proteomics) and of the metabolites in various organs (metabolomics). In this review, we will focus on these four ‘‘-omic’’ approaches applied to grapevine. Through the analysis of the recent bibliography, we will report how these ‘‘-omic’’ approaches were used separately or in combination to explore the enormous genetic and phenotypic diversity of grapevine.

## Genomics

Grapevine genomics made the most out the first generation of sequencing technology as the reference sequence of the genome PN40024 (Jaillon *et al.*, 2007) and the sequence of the heterozygous genome of Pinot Noir (Velasco *et al.*, 2007) were generated at the apogee of whole genome shotgun strategy using Sanger sequencing technology. These sequences allowed a detailed view of grapevine genome structure. Jaillon and collaborators (2007) confirmed that the genome size was around 485 Mb, harbouring about 30,000 genes and 40% of repetitive elements in the nearly homozygous genome of PN40024. They also found that even if grapevine is considered to be a true diploid, it is actually a paleo-hexaploid with regions present in three ancestral homologous copies in the genome. Analysis of the sequence of the heterozygous Pinot Noir (Velasco *et al.*, 2007) allowed the identification of candidate genes influencing wine quality through secondary metabolites or genes involved in susceptibility to pathogens in a cultivated grape variety. A particular focus was made on the regulatory elements in the genome with a detailed description of the transcription factor families and the non-coding RNAs including microRNAs. These two *de novo* sequence assemblies allowed to increase our general knowledge about the structure of the grape genome, whereas groups using second generation sequencing generated more insights on the grape pan-genome, that means the whole set of genes in the *Vitis* genus (Morgante *et al.*, 2007). The pan-genome is composed of two sets of genes: i) the genes of the core-genome, that are common among the *Vitis* genomes and ii) the genes of the dispensable genome, that differs from one variety to another and are at the origin of cultivar-specific traits. More recently, the sequence of the genome of the Tannat variety was used to analyse the molecular bases of its intense pigmentation and its high amount of antioxidant compounds (Da Silva *et al.*, 2013). Not only expansion of gene families involved in polyphenols biosynthesis was found but also entire regions carrying cultivar-specific genes involved in the production of polyphenols were identified and showed to be lacking in the reference genome of PN40024.

The genome of the Sultanina variety was sequenced to study the difference between table and wine grapes (Di Genova *et al.*, 2014). Two hundred and forty novel genes lacking in the reference genome were identified, some of which potentially involved in embryo development and in the seedless trait important for table grape. Finally, these studies demonstrate the high plasticity of the grape dispensable genome in terms of gene content involved in cultivar-specific traits. The recent development of third generation sequencing technologies allowing the sequencing of 10 kb-sized reads or longer will surely confirm this specificity of the grape genome and will allow to dig into haplotype-specific traits as in the case of Cabernet Sauvignon (Chin *et al.*, 2016).

Another strategy to study the dispensable and core genomes is the use of resequencing data coming from whole genome sequencing at low depth (around 10-20X). This cost-efficient method is based on the alignment of the reads against the reference genome to identify large structural variations that could involve whole genes. Although it is less precise than *de novo* whole genome assembly, especially for identifying genes lacking from the reference genome, this approach has been applied successfully on many plant species (Causse *et al.*, 2013; Kim *et al.*, 2016; Guo *et al.*, 2014) and seems to be very promising on grape (Morgante, PAG Conference 2015 (https://pag.confex.com/pag/xxiii/webprogram/Paper14025.html), personal communication). Additionally, RNA-Seq analyses were used to assess the transcribed part of the grape dispensable genome. RNA-Seq was performed on a pool of 45 samples originating from different organs and developmental stages of the Corvina variety and allowed the identification of 180 Corvina-specific transcripts absent in the reference genome (Venturini *et al.*, 2013). Along with the sequencing of the Tannat variety, RNA-Seq was performed on berries and the *de novo* assembly of the transcriptome allowed to identify 902 Tannat-specific transcripts compared to PN40024, Pinot Noir and Corvina varieties (Da Silva *et al.*, 2013).

Even though RNA-Seq allows to detect only expressed genes, it is the best approach to identify alternative splicing, a process that confers transcriptome plasticity. Ten grapevine berry transcriptomes were studied (Potenza *et al.*, 2015) showing that over 40 % of multi-exonic genes are subjected to alternative splicing. Although around 70 % of splicing events have low expression, they were highly conserved in the ten studied cultivars. The assembly of the Corvina transcriptome allowed the identification of 19,517 novel isoforms of 9,463 genes annotated in the reference genome (Venturini *et al.*, 2013). Thus, even if genes are part of the core-genome of *Vitis*, they could potentially show cultivar-specific isoforms which may have an impact on cultivar-specific traits but this needs further validation.

In grapevine, structural genome variations have already been shown to impact the visible phenotype. The retrotransposon *Gret1* was identified as the cause of the loss of red colour in grape berries, making them white or either pink depending on insertion or excision (Kobayashi *et al.*, 2004). Along with the *de novo* assembly of the Sultanina variety, a detection of structural variations was performed and allowed to identify 13 putative candidate genes located in previously mapped QTLs for the seedlessness trait and showing structural variations or SNP (Di Genova *et al.*, 2014).

Polymorphisms like structural variations and SNPs are efficiently identified using second generation sequencing data mapped against the reference genome. The availability of thousands of markers allowed greater insight into the evolutionary history of the *Vitis* genus. Using resequencing data, Myles and coworkers (2010) developed a genotyping array of 9 k SNPs that they used to assess the genetic diversity and linkage disequilibrium in cultivated and wild *V. vinifera* plants (Myles *et al.*, 2011). They found that, even though grapevine has been cultivated for thousands of years, its high genetic diversity has been maintained despite the domestication and the breeding bottlenecks. By exploring the structuration of the *V. vinifera* population, they showed that only a small portion of the genetic combinations in the *Vitis* genus has been explored. They suggested that the devastation of European vineyards by mildews and phylloxera but also the repeated use of only a few elite cultivars in crosses, may explain the limitations in the exploitation of the *Vitis* genetic diversity. Wen and coworkers (2013) developed an original approach based on RNA-Seq to characterize 417 transcripts from 15 *Vitaceae* species. The sequences of these transcripts were used as markers to establish a robust phylogeny of these species. Genotyping By Sequencing (GBS) is a genotyping method under intensive development in plants that takes advantage of the flexibility and cost-efficiency of next generation sequencing to perform SNP detection at the whole genome scale (Elshire *et al.*, 2011). Using GBS, Migicovsky and coworkers (2016) evaluated the genetic ancestry of 64 hybrid cultivars. Their results suggest that the genomes of these hybrids are composed of a high proportion of wild *Vitis* to the detriment of *V. vinifera*, meaning that hybrid grape breeding is still in its infancy and could benefit from marker-assisted selection.

Genotyping and markers are not only useful to assess the genetic diversity of the grapevine species but can also be helpful to exploit this diversity for the breeding of new elite grapevine varieties (Myles, 2013). High resolution genotyping using the previously described 9k SNP array was used by (Mahanil *et al.*, 2012) to perform a QTL analysis to characterize Ren4 powdery mildew resistance and seedlessness traits. Narrow genomic regions were identified allowing their introgression through marker-assisted selection in breeding programs. Even though grapevine is highly heterozygous, GBS was efficiently used for SNP detection, genotyping and genetic mapping (Hyma *et al.*, 2015). GBS was successfully applied on a segregation population derived from the crossing of *V. riparia* Michx and *V. vinifera* cv Seyval for the identification of QTLs governing fruit quality traits. This approach greatly densified the genetic map previously established with SSR markers and allow a definition of QTLs compatible with further marker-assisted breeding (Yang *et al.*, 2016). Even if they are in their infancy in grapevine, Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) are believed to be the most promising approaches making advantage of genomics and next generation sequencing to characterize and exploit diversity (Myles, 2013). Fodor and coworkers (2014) made a proof-of-concept for the accuracy of GWAS and GS approaches on grapevine as a model for highly heterozygous perennial plants. Finally, they recommended using the combined GWAS-GS prediction model and a core-collection as training population for grapevine breeding.

**miRNA**

Because they are involved in the regulation of several biological processes, microRNAs (miRNAs) are widely studied. They are non-coding RNAs, about 21 nucleotides long, which target mRNAs by sequence complementarity leading to a decrease of expression by silencing and post-transcriptional regulation. In their review, Thomson et al. (2011) summarize techniques used to study miRNAs or their targets such as RNA-Seq, immunoprecipitation or reverse transcription. Each technique has strengths and weaknesses and a combination of several approaches is often needed to obtain trustable results. miRNAs were studied in *V. vinifera* using both microarrays and RNA-Seq (Mica *et al.*, 2009). It was possible to identify alternative splicing for miRNAs and highlight splicing events and patterns. Their expression was studied in different tissues allowing the identification of miRNAs precursors specifically expressed in particular tissues. The expression profiles of miRNAs were found to be different in different organs but also during berry development.

Small RNA-Seq was performed on grapevine flowers and berries to identify miRNAs (Wang *et al.*, 2011). 130 conserved miRNAs were confirmed by comparison with the database miRBase. 80 non-conserved or novel miRNAs were found and RT-PCR was performed on the novel miRNAs to confirm their existence in grapevine. Among these, 20 were found to be specifically expressed in berries and potentially involved in fruit development. The target genes of the identified miRNAs were then searched for and the result showed that miRNAs could have an impact on genes involved in plant development and stress response. The same research group conducted a similar study on *V. amurensis* miRNAs (Wang *et al.*, 2012). 126 conserved miRNAs were found, among which 72 were possibly specific to this species. Like in the previous study, miRNAs were found to affect grape development and the response to stresses.

In the light of the potential involvement of miRNAs on stress responses, many studies focused on analysing miRNAs under stress condition. In 2015, Sun *et al.* studied the difference in miRNAs content of *V. vinifera* cultivated under normal temperatures and at 4°C. In both conditions, 163 known and 67 putative novel miRNAs were identified. 44 miRNAs were differentially expressed in plants submitted to cold stress, among which 13 were upregulated. Analysis their target genes highlighted, in some cases, a negative correlation between these genes and miRNAs, but also suggested that miRNAs could regulate transcription factors like SBP, MYB, GRAS, bZIP. Similarly, the impact of water stress on miRNAs was investigated in grapevine (Pantaleo *et al.*, 2016). Virus-free plants and plants infected by a latent virus were both exposed to drought. The infected plants were shown to be more tolerant to stress than the virus-free ones. The authors suggested that miRNAs expressed in response to the water stress might be involved in drought stress tolerance. Further analyses are however needed to confirm this hypothesis.

# Functional genomics

## Gene families

The availability of the reference genome made the study of families of genes easier. Indeed, by screening the genome, it is possible to find, by similarity, genes that are likely to belong to the same family and evaluate precisely their number. Grapevine being used to produce wine, genes involved in aroma biosynthesis are of special interest. Terpenoids are important compounds for grape and wine flavours. The genes coding for the enzymes producing these compounds, the terpene synthases (TPS), were studied in *V. vinifera* by Martin and colleagues (2010). Analysis of the 12X reference genome revealed that the *TPS* gene family is one of the largest involved in secondary metabolism, with 69 putatively functional *TPS*, 20 partial genes and 63 probable pseudogenes. Gene annotation highlighted a cluster of 45 *TPS* on chromosome 18. Functional characterisation of 39 enzymes revealed that some of them are unique for a certain function, showing the diversification of this family.

Grapevine being susceptible to numerous diseases, mechanisms and genes involved in plant defences are of great interest. Stilbenes, like resveratrol, are secondary metabolites playing a major role in grapevine defences. The genes coding for the enzymes producing these compounds in grapevine, stilbene synthase (STS), were studied by Parage and coworkers (2012). The 12X reference genome was screened to identify genes of the *STS* family. After a refined and manual annotation, 48 *STS* genes were identified showing that this family, like the *TPS* genes family, is greatly amplified in grapevine compared to other plant. Evolutionary analysis showed that this family was subjected to purifying selection, resulting in a tendency to maintain all of the 48 gene copies. In parallel, Vannozzi *et al.* (2012) also studied the *STS* gene family with a special focus on gene expression under biotic and abiotic stresses. This study also includes chalcone synthases (CHS), which are enzymes closely related to STS and competing for the same substrate. *STS* gene expression was analysed in response to UV light, wound stress and downy mildew infection using a combination of microarray and RNA-Seq analyses. Three groups of genes could be identified regarding their response to UV, with high, intermediate and low response, respectively. Consistent with the fact that STS and CHS are competing enzymes, a difference of regulation between their genes was highlighted.

Similarly, the expression of genes coding for callose synthases was analysed following infection with *Plasmopara viticola* (Yu, 2013). 8 genes were identified in different *Vitis* species: *Muscadinia rotundifolia*, immune to the pathogen, *Vitis amurensis*, resistant to the pathogen, and *Vitis vinifera*, susceptible. Most genes were significantly induced during the infection. In susceptible cultivar, gene induction did not lead to callose production. In resistant cultivars, callose was accumulated and correlated with the resistance level of the corresponding *Vitis* species.

A list of genes families studied in grapevine is provided in Table 1, together with information about the size and role of these families, and the main results of the studies.

**Table 1: Inventory of gene families studied in grapevine.**

| Family of genes | Number of genes | Role of this family | Main results |
|---|---|---|---|
| SET DOMAIN GROUP (Aquea *et al.*, 2011) | 33 putative genes | Epigenetic regulators | - Deregulated during viral infection by Grapevine Leafroll Associated Virus 3<br>- Level of expression directly linked to changes observed in development<br>- Precise role unclear |
| Sirtuin/Sir2 (Silent information regulator 2) (Cucurachi *et al.*, 2012) | 2: VvSRT1, VvSTR2 | VvSRT2 possibly linked to photosynthetic or chloroplast activities | - VvSRT1 does not vary in leaves and berries<br>- VvSRT2 is most expressed in young leaves |
| Expansin (Dal Santo *et al.*, 2013) | 29 genes separated in four families | Cell wall modifications, cell expansion, most likely work by increasing movement among polymer compounds of cell walls | - Expansin-like B expressed in woody tissues suggesting a role in the formation of secondary cell walls<br>- The other families are involved in berry development |
| Defensin (Giacomelli *et al.*, 2012) | 79 putative genes: 46 complete genes<br>Cluster organisation | Expression in tissues associated with reproduction<br>Some specific genes expressed during berries ripening | - upregulation of some genes during *Botrytis cinerea* infection |
| Dehydrin (Yang *et al.*, 2012) | 4 genes: no expansion of the family compared to other plants | | - One gene with higher response to abiotic stress and *Erysiphe necator*<br>- One gene specifically express during late embryogenesis |
| Responsive to Dehydration 22 (RD22) (Matus *et al.*, 2014) | 19 genes coding for proteins containing a BURP domain | Evaluate the response of the abscisic acid to an abiotic stress | - Improve phylogenetic classification<br>- Large range of expression |
| FK506-binding proteins (Shangguan *et al.*, 2013) | 23 genes on 11 chromosomes | Receptors of FK506 and rapamycin<br>Acting as peptidylprolyl isomerase (PPIase) and protein folding chaperones | - Expression studied in stems, inflorescences, flowers and fruit tissues<br>- Five genes expressed in all tissues<br>- Some genes without any expression |
| Serine acetyltransferase (Tavares *et al.*, 2015) | 4 genes | Producing the O-acetylserine which is then transformed in cysteine in the sulphate reduction pathway | - Enzymes located in the cytosol, one of them is also found in plastids and another in mitochondria<br>- One gene strongly induced upon sulphate depletion and in leaves upon drought |

| Family of genes | Number of genes | Role of this family | Main results |
|---|---|---|---|
| Calcium-dependent protein kinases (CDPK) (Zhang *et al.*, 2015) | 19 genes divided in four groups | Role in plant development, stress response and hormone signalling | - Constructing phylogeny<br>- Most of the genes are induced under biotic and abiotic stresses in the Chinese grape *Vitis pseudoreticulata* |
| **Transcription factors** | | | |
| MADS-box (Grimplet *et al.*, 2016) | 90 genes among which 30 are new | Involved in developmental processes in flowers, fruits and seeds | - New genes specific to grapevine and belonging to the same group compared to other genes conserved in plants |
| AP2/ERF (Licausi *et al.*, 2010) | 149 genes | Important for plant development and stress response | - Specific groups of this family are showing an amplification in their number<br>- Potential new foles in fruit ripening<br>- Not fully understood yet |
| bZIP (Liu *et al.*, 2014) | 55 genes | Important for seed development, heat and drought responses | - Constructing phylogeny based on known genes in *Arabidopsis* |
| GRAS/SCL domain (Sun *et al.*, 2016) | 43 genes separated in six groups | Involved in plant development | - Constructing phylogeny<br>- Portion of genes specifically expressed in roots, stems or tendril<br>- Response specific during cold and water deficit stresses |
| WRKY genes (Wang *et al.*, 2014) | 59 putative genes separated in three groups | Zinc-finger transcription factors involved in plant development and stress response | - Expression profile specific according to the group in which WRKY genes were classified<br>- Genes from groups II-a and III involved in stress response (drought, powdery mildew infection, salicylic acid and ethylene treatments)<br>- *VvWRKY07*, *08* and *25* more induced than others |
| Dehydration responsive element-binding (DREB) (Zhao *et al.*, 2014) | 38 genes, on 15 chromosomes, divided in six groups according to *Arabidopsis* classification | Involved in the response to several abiotic stresses, hormone treatments and bud and berry growth | |

**Stress responses**

As previously shown, it is possible to study stress response of specific gene families. However, transcriptomics allow the analysis of the responses of an organism to specific conditions at the whole-genome scale. It is therefore possible to apply this approach to pathogen infection or stress response. Abbà and colleagues (2014) went a step further by improving the draft genome of the flavescence dorée phytoplasma using RNA-Seq. Indeed, by studying infected grapevine and using RNA-Seq with the pathogen genome as reference, it was possible to identify 3 new genes, 10 unannotated genes, and improve the annotation of 44 genes. Analysis of phytoplasma gene expression showed that the most expressed genes were related to translation and protein biosynthesis but also included many genes with unknown function. A mobile element was also shown to be highly expressed being suspected to contribute to the plasticity of the pathogen genome.

Many regions of the world are affected by climate changes, raising the question of grapevine tolerance to high temperature, drought or cold. Rienth and coworkers (2014), studied the effect of heat stress applied day and night to grape berries. These stresses were short but intense to simulate summer time. Using microarrays, the authors identified differences in transcripts related to malic acid and anthocyanin biosynthesis, which they attributed to the stress response, that were dependent on berry development. Moreover, some variations observed were different between day and night especially regarding genes involved in acidity and phenylpropanoid metabolism. Most heat stress-responsive genes were not dependant on the development stage of the berry, nor dependant on the time at which the stress was applied. With global warming, grapevine might be more and more exposed to drought. It is therefore important to study the tolerance mechanisms to this stress. Two grapevine genotypes, one tolerant and one sensitive to drought, were studied at the genomics and transcriptomics levels (Corso *et al.*, 2015). Both cultivars were exposed to progressive drought stress. A high accumulation of resveratrol in roots and flavonoids in leaves was noticed in the tolerant cultivar. The authors proposed that accumulation of resveratrol in roots could help the plant to cope with the oxidative stress associated with water deficit.

Climate changes will also be found under the form of decrease of temperatures and episodes of cold exposing the grapevine. Studying that stress and understanding its mechanisms is consequently important. *V. amurensis*, a wild Chinese cultivar tolerant to cold, and *V. vinifera*, widely cultivated but susceptible to cold, were studied (Xin *et al.*, 2013). Fewer genes were differentially expressed in *V. amurensis* (1314 genes) compared to *V. vinifera* (2307 genes) in response to cold stress. Only a small portion (408) of these transcripts were common between the two species. The differences observed suggest that specific pathways are triggered by cold in *V. amurensis*, as more genes related to metabolism, transport, signal transduction and transcription are upregulated in *V. amurensis*. The cold response of *V. amurensis* was analysed further using RNA-Seq by Xu and colleagues (2014). 6850 transcripts were differentially expressed in response to cold, 3676 being upregulated and 3174 downregulated. mRNAs were affected by alternative splicing, in particular by exon skipping, these events increasing with the cold exposure. 37 families of TFs, of genes involved in metabolism, biosynthesis of secondary metabolites and heat shock proteins were shown respond to cold.

Grapevine is susceptible to numerous diseases, whose control requires the best possible understanding of infection mechanisms. In addition, grapevine resistance and susceptibility phenotypes are highly diverse among *Vitis* species and cultivars, making the characterization of this diversity a difficult but necessary task. Downy mildew caused by the Oomycete *Plasmopara viticola* is a major disease affecting grapevine. Perazzolli and coworkers (2012) used RNA-Seq to analyse the effect of a downy mildew infection on grapevine, in combination with a preventive infection by *Trichoderma harzianum* T39. Asymptomatic infection by T39 can induce defences in grapevine, allowing a better reaction against *P. viticola*. Leaf samples were collected on both grapevine infected or not by T39 at 0 and 24 hours after inoculation with downy mildew. 7024 genes were differentially expressed showing the complexity of the resistance response. T39 treatment affected genes involved in microbial recognition but impacted also defence responses towards downy mildew, by activating partially the mechanisms associated to the resistance to *P. viticola*.

Powdery mildew caused by the fungi *Erysiphe necator* is another widely spread disease in the vineyard. Jones and colleagues (2014) first used a shotgun approach to sequence and assemble the genome of different *E. necator* isolates. Then, RNA-Seq and comparative genomics were used on infected grapevine to predict and annotate genes coding for the pathogen proteins, but these RNA-Seq data constitute a valuable resource to analyse the response of grapevine to powdery mildew infection. Finally, RNA-Seq was used to analyse the transcriptome of grapevine wood tissues infected with *Neofusicoccum parvum*, a fungus associated with grapevine trunk disease (Czemmel *et al.*, 2015). Using RNA-Seq, genes differentially expressed in the case of an infection were identified. As grapevine trunk disease is characterized by a long latent phase, it was of special interest to identify candidate genes affected in this phase in order to use them as potential markers for early detection of the disease. Four candidates were selected galactinol synthase, abscisic acid-induced wheat plasma membrane polypeptide-19 orthologue, embryonic cell protein 63 and BURP domain-containing protein. However, further experiments will be needed to confirm that these genes are specifically induced during the early stages of *N. parvum* infection and may be used as markers for grapevine trunk disease.

**Organ specific experiments**

RNA-Seq has been used to characterize the transcriptome of various grapevine organs, with a special focus on leaves, berries and flowers. Sex specification of *V. sylvestris* male and female flowers was studied (M., J., N., Ramos *et al.*, 2014) and compared to *V. vinifera*, whose flowers are hermaphrodite. *V. vinifera* flowers are fully functional, whereas *V. sylvestris* male flowers have reduced pistils and female flowers infertile pollen. For each type of flower, four development stages were analysed, as sex determination appears late in flower development. Clusters of genes differentially expressed between genders and between developmental stages have been proposed to be involved in sex differentiation.

Grape berries are of special interest for winemaking or as table grapes, which motivated extensive analysis of berry transcriptome. González-Agüero and colleagues (2013) have tried to identify reference genes in order to standardize RNA-Seq analyses of berries. A set 19 non-differentially expresses genes of was selected in 12 berry samples. Among them, *VvAIG1* (AvrRpt2-induced gene) and *VvTCPB* (T-complex 1 beta-like protein) had the most stable expression and could be considered as reference genes.

Based on the 8X reference genome, RNA-Seq was used to characterize berry development in *V. vinifera* cv. Corvina, through the analysis of three stages: post setting, véraison and ripening, were studied in (Zenoni *et al.*, 2010). 6695 genes were found to be expressed specifically depending on the development stage, showing the complexity of berry development. Four development stages: 3 weeks after flowering, early véraison, late véraison and harvesting, were studied using a more recent RNA-Seq technique (paired-end reads of 100 bp) in *V. vinifera* cv. Shiraz (Sweetman *et al.*, 2012). 4185 transcripts were upregulated in a single development stage. A coordination between organic acid, stilbene and terpenoid metabolisms was highlighted. However, further analyses will be needed to confirm that the changes observed are only due to berry development, and not to environmental conditions. Studying berries, just before and after véraison, in three different years allows the study of seasonal influence (Pilati *et al.*, 2007). Using microarrays, 1477 genes were shown to be consistently modulated over the studied years. During ripening, five functional categories were especially induced: cell wall organization and biogenesis, carbohydrate and secondary metabolisms and stress response. At the same time, photosynthesis was strongly repressed. This study also show, for the first time, the oxidative burst happening in grapevine during véraison. Finally, to specifically study flavours, skin and pulp of berries at a late development stages were analysed (Cramer *et al.*, 2014). Both skin and pulp exhibited massive transcriptional changes in late development stages. However, more transcripts involved in ethylene signalling, isoprenoid and fatty acid metabolism were expressed in skin, as well as those involved in the production of flavour and aroma compounds.

Grapes are used to make wine but are also eaten fresh, as table grape, or dried. An appreciated feature of table grape is to be seedless, which motivated the comparison of seeded and seedless grapevine performed by Nwafor and colleagues (2014). 80% of the transcripts had the same expression range between these two samples. Nevertheless, 1075 genes were differentially expressed, highlighting differences in pollen and ovule developmental pathways. For both table and wine grape, berry size is a major trait of importance. Even though gibberellins are known to increase grape berries size, the related mechanisms were only recently studied at the transcriptomic level (Chai *et al.*, 2014). The number of genes differentially expressed was increasing over time, most of the genes being down regulated. Genes involved in cell wall relaxation, through cell wall modification enzymes, cytoskeleton and membrane components and transporters, were proposed to play a major role in berries enlargement.

Finally, Fasoli and coworkers (2012), conducted a large-scale transcriptomics study on several organs like roots, leaves, pollen, in order to establish a grapevine gene expression atlas. Pollen and senescent leaves had a dedicated function and thus transcriptomes. The samples coming from other organs were pooled together. Using both microarrays and RNA-Seq, most samples could be categorised in two major classes regarding their maturity rather than their organ identity, namely, the vegetative/ green and mature/woody categories. A major reprogramming process was highlighted during maturation (*i.e.* the shift between these two categories), which was not observed in herbaceous annual species. This suggest that this reprogramming might be characteristic of perennial woody plants like grapevine. It is worth citing the tool developed using this data, which can be found at http://bar.utoronto.ca/efp_grape/cgi-bin/efpWeb.cgi. It consists of a browser where the expression of a given gene can be visually displayed in colours with a heatmap pattern, depending on its expression level.

To facilitate the retrieval and reuse of RNA-Seq data, an inventory of RNA-Seq experiments is presented in the Table 2. This table details the type of experiment, gives the access number (when possible) and the related publication.

**Table 2: inventory of RNA-Seq experiments conducted on grapevine.**

| Type of experiments | Accession number / website | Authors |
|---|---|---|
| Berries from *Vitis vinifera* cv. Centenial Seedless Treatment with gibberellin | SRA: SRP038904 | (Chai *et al.*, 2014) |
| Rootstock from M4 and 101.14 | SRA: SRA110531 | (Corso *et al.*, 2015) |
| Leaves from Cabernet Sauvignon infected with *Neofusicoccum parvum* | GEO: GSE58653 | (Czemmel *et al.*, 2015) |
| Berries, skin and seeds from *Vitis vinifera* cv. French Tannat | SRA: PRJNA203687 | (Da Silva *et al.*, 2013) |
| Berries from 'Ruby seedless' x 'Sultanina' crossing and parents | SRA: SRX366617 | (González-Agüero *et al.*, 2013) |
| Leaves from Carignan infected by *Erysiphe necator* (no non-infected controls) | GEO: GSE58958<br>SRA: SRP043708 | (Jones *et al.*, 2014) |
| Callus, leaf, root, stem from PN40024 | http://www.genoscope.cns.fr/externe/gmorse/raw_data/ | (Mica *et al.*, 2009) |
| Flowers, young berries, ripe berries | GEO: GSE58061 | (Nwafor *et al.*, 2014) |

| Type of experiments | Accession number / website | Authors |
|---|---|---|
| Leaves from *Vitis vinifera* cv. Pinot Noir inoculated with *Trichoderma harzianum T39* then infected with *Plasmopara viticola* | SRA: PRJNA168987 | (Perazzolli *et al.*, 2012) |
| Leaves from Summer Black (hybrid of *Vitis vinifera* and *Vitis labrusca*) | GEO: GSE74428, GSM1920330, GSM1920331, GSM1920332, GSM1920333 | (Pervaiz *et al.*, 2016) |
| Berries from *Vitis vinifera* cv. Pinot noir, Teroldego, Alicante Bouschet, Sangiovese, Moscato rosa, Lambrusco salamino, Cabernet franc, Chardonnay, Ansonica and Kozma Poloskei Muskotaly | SRA: PRJEB9534 | (Potenza *et al.*, 2015) |
| Flowers from *Vitis vinifera* cv. Touriga Nacional and *Vitis sylvestris* | GEO: GSE56844 | (M., J., Ramos *et al.*, 2014) |
| Berries from *Vitis vinifera* cv. Shiraz | On demand | (Sweetman *et al.*, 2012) |
| Leaves from *Vitis vinifera* cv. Pinot Noir under UV stress, wound stress and downy mildew infection | GEO: GSE37743 | (Vannozzi *et al.*, 2012) |
| Berries from *Vitis vinifera* cv. Corvina | SRA: SRA055265. | (Venturini *et al.*, 2013) |
| Mix from stems, leaves, tendrils and flowers for *Cissus microcarpa, Parthenocissus vitacea, Rhoicissus digitate, Ampelopsis arborea, Leea guineensis, Cayratia japonica, Cissus tuberosa, Tetrastigma lawsonii, Ampelopsis cordata, Vitis rotundifolia, Vitis tiliifolia, Pterisanthes eriopoda, Nothocissus spicifera, Cyphostemma sandersonii, Ampelocissus elegans* | SRA: PRJNA205096 | (Wen *et al.*, 2013) |
| Shoot from *Vitis amurensis* and *Vitis vinifera* cv. Muscat of Hamburg under cold stress | SRA: SRP018199 | (Xin *et al.*, 2013) |
| Leaves from *Vitis amurensis* | SRA: SRX314996, SRX315119, SRX315120 and SRX315121 | (Xu *et al.*, 2014) |
| Berries from *Vitis vinifera* cv. Corvina | SRA: SRA009962 | (Zenoni *et al.*, 2010) |
| | | |
| Small RNA-Seq | GEO: GSE85611 | (Paim Pinto *et al.*, 2016) |
| Small RNA-Seq | GEO: GSE63244 | (Pantaleo *et al.*, 2016) |
| Small RNA-Seq | GEO: GSE68970 | (Sun *et al.*, 2015) |
| Small RNA-Seq | GEO: GSE24531 | (Wang *et al.*, 2011) |
| Small RNA-Seq | GEO: GSE34169 | (Wang *et al.*, 2012) |
| Small RNA-Seq | SRA : SRR890731 | (Zhang *et al.*, 2014) |

## Proteomics

Proteomics is one of the various "-omic" approaches whose integration is a challenging task, but that may increase our comprehension of interplay between genome, transcriptome, proteome, metabolome etc. of living organisms (Bindschedler *et al.*, 2016). This integration will probably greatly benefit from the rapid evolution of proteomics tools, such as 2-Dimension Electrophoresis (DE), mass spectrometry techniques but also from the availability of fully sequenced genomes and bioinformatics tools. First proteomic studies were adapted to plants about 15 years ago (van Wijk, 2001; Cánovas *et al.*, 2004). The model plant *A. thaliana* was studied through proteomics (for example: Peck, 2005; Baginsky and Gruissem, 2006; Wienkoop *et al.*, 2010) but other plants of agronomical interest were also studied, like rice (Agrawal and Rakwal, 2006; Salekdeh *et al.*, 2002), wheat (Skylas *et al.*, 2001; Majoul *et al.*, 2003) or barley (Østergaard *et al.*, 2002).

Few articles were published on the proteomics of grapevine. Giribaldi and Giuffrida (2010) published a review on the proteomics of grapevine and wine. This article focused mainly on technological aspects of wine and berries like protein precipitation, presence of allergenic proteins or evolution of the protein pool in berries. Most of the studies concerning proteomics grapevine are related to responses to pathogens. A smaller set of publications concerns the variation of the protein pool of berries during ripening. A recent review (Shiratake and Suzuki, 2016) presents the state of the art on these subjects. Briefly, pathogens induce synthesis of proteins involved in hypersensitive response, such as PR10 proteins (or Ribosome Inactivating Protein) that are well known for their antifungal activity (Borad and Sriram, 2008) or NBS-LRR (Nucleotide Binding Site-Leucine Rich Repeat) proteins that are receptors involved in pathogen recognition and resistance (Cui *et al.*, 2014). Other publications also show an overexpression of proteins involved in oxidative stress response (Dadakova *et al.*, 2015) when plants are infected by *Botrytis cinerea* or by viruses (Giribaldi *et al.*, 2011). A kinetics of infection of grapevine by *Plasmopara viticola* (Milli *et al.*, 2012) showed that the first response, at 24 hours, consisted in the deregulation of proteins important for the photosynthesis. Resistance and response to stress is suppressed by the pathogen after 48 hours. After 96 hours though, sporulation of the pathogen starts and most of the proteins, modulated during the first response, are once again affected because the host is able to recover. These results also showed that the secondary metabolites are induced too late to significantly affect the progression of the infection.

Recently, a new and promising kind of study was undertaken by Nascimento-Gavioli and colleagues (2016) (in press). The authors studied the compatible and incompatible interaction of *Vitis* and *Plasmopara viticola*. The incompatible interaction Rpv1/Rpv3 was pyramided into compatible *Vitis*. The authors identified that the incompatible interaction resulted from the expression modulation of 41 proteins. For the first time, a study provides new insights at the protein level on the mechanisms of compatible and incompatible interaction between grapevine and downy mildew. Proteomics applied to berries describes proteins over or under expressed after véraison (Shiratake and Suzuki, 2016). Most of the identified proteins are involved in biotic or abiotic stress. The authors of the original studies suggest that these proteins can be markers of berry ripening. An integrative study, including transcriptomics, proteomics and metabolomics, was performed by Zamboni and coworkers (2010) on berry ripening. This study confirmed previous works but also showed that cell wall metabolism, photosynthesis and stress response are involved in ripening.

Despite great interest in proteomic studies, they rarely give clear and conclusive results, probably because of the complex interactions between metabolic pathways that are thus affected either by an infection or by a stress. Moreover, the great dynamics of protein expression in cells or organs and the limitations of proteomic techniques rarely allow determining causality of the observed phenomenon. In their review, George and Haynes (2014) discuss some proteomics limitations, such as the lack of available sequenced genomes for different grapevine cultivars apart from the reference genome PN40024. However, progresses in techniques and innovative and integrated approaches will probably improve our knowledge in grapevine proteomics.

## Metabolomics

In recent years, the development of metabolomics has provided new phenotyping tools for advanced understanding of primary and secondary metabolism in plants, as well as new ways to characterize plant diversity. Targeted and global non-targeted metabolomics analyses have therefore been widely used to better characterize grapevine development, responses to various stresses or environmental cues and phenotypic diversity. Gas chromatography coupled to mass spectrometry (GC-MS) has been used to obtain the metabolic profile of six stages of berry development, from flowering to ripe berries, in Cabernet Sauvignon and Merlot (Cuadros-Inostroza *et al.*, 2016). Sugars and amino acids levels were shown to have opposite behaviour during development, with a strong increase of soluble sugars and a parallel decrease of amino acid content upon berry ripening. However, some observed changes were stage and cultivar dependant suggesting differences in metabolic regulations between cultivars. Similarly, metabolomics analysis of Cabernet Sauvignon and Shiraz berry skins revealed similar developmental patterns of change in primary metabolites between the two cultivars. However, the extent of change in the major organic acid, sugars and flavonoids was greater in Shiraz compared to Cabernet Sauvignon. Part of these differences may be explained by a greater responsiveness of Shiraz to environmental cues, which may lead to changes in metabolic traits at fruit harvest. To extend that study, the same research group studied differences between red and white Muscat (Degu *et al.*, 2015). Sixty-one Muscat genotypes were studied and could be separated into six groups mainly depending on their anthocyanin composition. Three groups contained the red genotypes and three other groups contained the white ones. White genotypes were shown to have higher levels of flavonols and flavanols, which is explained by their lack of anthocyanins. The varietal affiliation was mainly based, for the red genotypes, on their anthocyanin profile, and for the white genotypes, on their level of quercetin 3-O-galactoside. Metabolic profiling better resolved the variability existing in the Muscat collection than SNP genetic mapping using a 20 k SNP array, confirming the potential of metabolomics to characterize grapevine diversity.

Several techniques such as metabolite profiling, expression analysis and biochemical studies were combined to identify the genes coding for the monoterpenol glucosyltranseferase, an enzyme that is glucosylating a diversity of terpenes in grape berries (Bönisch *et al.*, 2014). Terpenes are important volatile compounds in wine aromas and in grapevine, and are mainly found in glucosylated forms that are not volatiles. This glucosyltranseferase was shown to trasfer glucose on monoterpenols such as nerol, geraniol and citronellol. Selection of grapevine varieties with low monoterpenol glucosyltranseferase activity will lead to grapes enriched volatile terpenes for improved wine aromas.

Metabolomics has been used in integrated approaches in combination with transcriptomics and genomics to study the expression and regulation of genes producing metabolites of interest. Several groups have already tested this strategy by combining metabolomics techniques, RNA-Seq and SNP analysis (Degu *et al.*, 2014; Degu *et al.*, 2015; Domingos *et al.*, 2016).

## Conclusion

Overall, a lot of tools and techniques are available and in constant evolution. Here are presented ‘‘-omic’’ approaches used to study grapevine diversity and traits. Even though each ‘‘-omic’’ field is different, the nature of the results obtained with each one of them are different. It is often useful to combine several techniques to obtain a more complete view of the studied object as shown in several studied cited in this review. In particular, the work of Ghan and co-workers (2015) where five ‘‘-omic’’ technologies (microarrays, RNA-Seq, nano-liquid chromatography-MS, GC-MS, LC-MS) were consistently able to discriminate different grapevine cultivars (Cabernet Sauvignon, Merlot, Pinot Noir, Chardonnay, Semillon).

The data generated by these techniques, once analysed can be integrated into specialised databases like miRVine that is integrating all miRNAs identified in grapevine. Raw data generated in the scope of specific analysis are also often available in public databases. They are open for anyone to reuse either for a similar analysis, using a different technique, or for a different use, like the study of a specific family of genes. Ideally it would be of great interest to have a network between all these databases to make them communicate with each other and make it possible to have a look at all information at once. In Table 3 is presented an inventory of databases proposing data on grapevine.

Combining the different ‘‘-omic’’ technologies allows us to have a bigger picture of what, where and how things are happening. From the DNA sequence to the metabolite going through gene expression, protein structure, pathways, metabolite effects and a lot more, new questions can be asked and are open to discoveries. It makes no doubt that these fields will continue to be developed and will allow the research to go further and further.

**Table 3: inventory of different tools and databases available for grapevine.**

| Database name | Description | Data | Authors | Link |
|---|---|---|---|---|
| vitaceae.org | International Grape Genome Program web page | Practical information | Adam-Blondon, Burger, Cheng, Deluc, Pezzotti | http://vitaceae.org/index.php/International_Grape_Genome_Program |
| grapevine Jbrowse | Navigation on grapevine genome | Reference genome sequence, annotation and polymorphism from several groups | URGI | https://urgi.versailles.inra.fr/jbrowse/gmod_jbrowse/? |
| Grape Genome Browser | Navigation on grapevine genome | Reference genome sequence, annotation | Genoscope | http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/ |
| Gene Expression Omnibus | Functional genomics datasets | Microarray, protein arrays, SNP arrays, NGS, etc | NCBI | http://www.ncbi.nlm.nih.gov/geo/ |
| Sequence Read Archive | Storage of sequences | Raw and alignment | NCBI | http://www.ncbi.nlm.nih.gov/sra |
| VitisCyc | Gathering of metabolic pathways, reactions, compounds, etc | Browsers for genes, pathways, enzymes and compounds | (Naithani *et al.*, 2014) | http://pathways.cgrb.oregonstate.edu/metabolic.html |
| VTCdb | Transcriptional regulatory inference and co-expression network | microarrays | (Wong *et al.*, 2013) | http://vtcdb.adelaide.edu.au/Home.aspx |
| TreeTFDB | Transcription factors from *Jatropha curcas*, papaya, cassava, poplar, castor bean and grapevine | GeneChip, PlexDB and eFP browser | (Mochida *et al.*, 2013) | http://treetfdb.bmep.riken.jp/index.pl |
| SNiPlay | Detection, management and analysis of SNPs Design of Illumina chips or research of SNPs for comparison | Standard sequence, genotyping data, Sanger sequencing | (Dereeper *et al.*, 2011) | http://sniplay.southgreen.fr/cgi-bin/home.cgi |
| miRVine | Atlas of miRNAs found in grapevine in different organs, at different developmental stages | Small RNA-Seq | (Belli Kullan *et al.*, 2015) | https://mpss.danforthcenter.org/dbs/index.php?SITE=grape_sRNA_atlas |
| Biowine | Functional analysis of two cultivars : Nero d'Avola and Nerello Mascalese | RNA-Seq | (Pulvirenti *et al.*, 2015) | http://alpha.dmi.unict.it/biowine/ |
| GrapeGenDB | Retrieving annotation | GrapeGen Affymetrix custom array (GrapeGena520510F) | http://bioinfogp.cnb.csic.es/tools/GrapeGendb/ | http://bioinfogp.cnb.csic.es/tools/GrapeGendb/ |

| Database name | Description | Data | Authors | Link |
|---|---|---|---|---|
| PLEXdb | Gene expression ressources for plants and plant pathogens | Microarrays | http://www.plexdb.org/index.php | http://www.plexdb.org/ |
| FlagDB++ | Integrative database around plant genomes | Annotations, cDNA, repeat elements, predicted CDS, gene families, protein motifs | (Dèrozier *et al.*, 2011) | http://tools.ips2.u-psud.fr/projects/FLAGdb++/HTML/index.shtml |

# References

**Abbà, S., Galetto, L., Carle, P., Carrère, S., Delledonne, M., Foissac, X., Palmano, S., Veratti, F. and Marzachì, C.** (2014) RNA-Seq profile of flavescence dorée phytoplasma in grapevine. *BMC Genomics*, **15**, 1088.

**Agrawal, G.K. and Rakwal, R.** (2006) Rice proteomics: a cornerstone for cereal food crop proteomes. *Mass Spectrom Rev*, **25**, 1–53.

**Aquea, F., Vega, A., Timmermann, T., Poupin, M.J. and Arce-Johnson, P.** (2011) Genome-wide analysis of the SET DOMAIN GROUP family in Grapevine. *Plant Cell Rep.*, **30**, 1087–1097.

**Baginsky, S. and Gruissem, W.** (2006) Arabidopsis thaliana proteomics: from proteome to genome. *J Exp Bot*, **57**, 1485–1491.

**Belli Kullan, J., Lopes Paim Pinto, D., Bertolini, E., et al.** (2015) miRVine: a microRNA expression atlas of grapevine based on small RNA sequencing. *BMC Genomics*, **16**, 393.

**Bindschedler, L. V, Panstruga, R. and Spanu, P.D.** (2016) Mildew-Omics: How Global Analyses Aid the Understanding of Life and Evolution of Powdery Mildews. *Front. Plant Sci.*, **7**, 123.

**Bönisch, F., Frotscher, J., Stanitzek, S., Rühl, E., Wüst, M., Bitz, O. and Schwab, W.** (2014) A UDP-Glucose:Monoterpenol Glucosyltransferase Adds to the Chemical Diversity of the Grapevine Metabolome. *Plant Physiol.*, **165**, 561–581.

**Borad, V. and Sriram, S.** (2008) Pathogenesis-related proteins for the plant protection. *Asian J Exp Sci*.

**Cánovas, F.M., Dumas-Gaudot, E., Recorbet, G., Jorrin, J., Mock, H.-P. and Rossignol, M.** (2004) Plant proteome analysis. *Proteomics*, **4**, 285–298.

**Causse, M., Desplat, N., Pascual, L., et al.** (2013) Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics*, **14**, 791–805.

**Chai, L., Li, Y., Chen, S., Perl, A., Zhao, F. and Ma, H.** (2014) RNA sequencing reveals high resolution expression change of major plant hormone pathway genes after young seedless grape berries treated with gibberellin. *Plant Sci.*, **229**, 215–24.

**Chin, C., Peluso, P., Sedlazeck, F.J., et al.** (2016) Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing.

**Corso, M., Vannozzi, A., Maza, E., et al.** (2015) Comprehensive transcript profiling of two grapevine rootstock genotypes contrasting in drought susceptibility links the phenylpropanoid pathway to enhanced tolerance. *J. Exp. Bot.*, **66**, 5739–5752.

**Cramer, G.R., Ghan, R., Schlauch, K.A., et al.** (2014) Transcriptomic analysis of the late stages of grapevine (Vitis vinifera cv. Cabernet Sauvignon) berry ripening reveals significant induction of ethylene signaling and flavor pathways in the skin. *BMC Plant Biol.*, **14**, 370.

**Cuadros-Inostroza, A., Ruíz-Lara, S., González, E., Eckardt, A., Willmitzer, L. and Peña-Cortés, H.** (2016) GC–MS metabolic profiling of Cabernet Sauvignon and Merlot cultivars during grapevine berry development and network analysis reveals a stage- and cultivar-dependent connectivity of primary metabolites. *Metabolomics*, **12**, 39.

**Cucurachi, M., Busconi, M., Morreale, G., Zanetti, A., Bavaresco, L. and Fogher, C.** (2012) Characterization and differential expression analysis of complete coding sequences of Vitis vinifera L. sirtuin genes. *Plant Physiol. Biochem.*, **54**, 123–132.

**Cui, H., Tsuda, K. and Parker, J.E.** (2014) Effector-Triggered Immunity: From Pathogen Perception to Robust Defense. *Annu. Rev. Plant Biol.*

**Czemmel, S., Galarneau, E.R., Travadon, R., McElrone, A.J., Cramer, G.R. and Baumgartner, K.** (2015) Genes expressed in grapevine leaves reveal latent wood infection by the fungal pathogen Neofusicoccum parvum. *PLoS One*, **10**, 1–21.

**Dadakova, K., Havelkova, M., Kurkova, B., Tlolkova, I., Kasparovsky, T., Zdrahal, Z. and Lochman, J.**

(2015) Proteome and transcript analysis of Vitis vinifera cell cultures subjected to Botrytis cinerea infection. *J. Proteomics*, **119**, 143–153.

**Dal Santo, S., Vannozzi, A., Tornielli, G.B., Fasoli, M., Venturini, L., Pezzotti, M. and Zenoni, S.** (2013) Genome-Wide Analysis of the Expansin Gene Superfamily Reveals Grapevine-Specific Structural and Functional Characteristics. *PLoS One*, **8**.

**Degu, A., Hochberg, U., Sikron, N., et al.** (2014) Metabolite and transcript profiling of berry skin during fruit development elucidates differential regulation between Cabernet Sauvignon and Shiraz cultivars at branching points in the polyphenol pathway. *BMC Plant Biol.*, **14**, 1–20.

**Degu, A., Morcia, C., Tumino, G., et al.** (2015) Metabolite profiling elucidates communalities and differences in the polyphenol biosynthetic pathways of red and white Muscat genotypes. *Plant Physiol. Biochem.*, **86**, 24–33.

**Dereeper, A., Nicolas, S., Cunff, L. Le, Bacilieri, R., Doligez, A., Peros, J.-P., Ruiz, M. and This, P.** (2011) SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics*, **12**, 134.

**Dèrozier, S., Samson, F., Tamby, J.-P., et al.** (2011) Exploration of plant genomes in the FLAGdb++ environment. *Plant Methods*, **7**, 8.

**Domingos, S., Fino, J., Paulo, O.S., Oliveira, C.M. and Goulao, L.F.** (2016) Molecular candidates for early-stage flower-to-fruit transition in stenospermocarpic table grape (Vitis vinifera L.) inflorescences ascribed by differential transcriptome and metabolome profiles. *Plant Sci.*, **244**, 40–56.

**Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E.** (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, 1–10.

**Fasoli, M., Dal Santo, S., Zenoni, S., et al.** (2012) The Grapevine Expression Atlas Reveals a Deep Transcriptome Shift Driving the Entire Plant into a Maturation Program. *Plant Cell*, **24**, 3489–3505.

**Fodor, A., Segura, V., Denis, M., et al.** (2014) Genome-wide prediction methods in highly diverse and heterozygous species: Proof-of-concept through simulation in grapevine. *PLoS One*, **9**.

**Genova, A. Di, Almeida, A.M., Muñoz-Espinoza, C., et al.** (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.*, **14**, 7.

**George, I.S. and Haynes, P.A.** (2014) Current perspectives in proteomic analysis of abiotic stress in Grapevines. *Front. Plant Sci.*, **5**, 686.

**Ghan, R., Sluyter, S.C. Van, Hochberg, U., et al.** (2015) Five omic technologies are concordant in differentiating the biochemical characteristics of the berries of five grapevine (Vitis vinifera L.) cultivars. *BMC Genomics*, **16**, 946.

**Giacomelli, L., Nanni, V., Lenzi, L., et al.** (2012) Identification and Characterization of the Defensin-Like Gene Family of Grapevine. *Mol. Plant-Microbe Interact.*, **25**, 1118–1131.

**Giribaldi, M. and Giuffrida, M.G.** (2010) Heard it through the grapevine: proteomic perspective on grape and wine. *J Proteomics*, **73**, 1647–1655.

**Giribaldi, M., Purrotti, M., Pacifico, D., et al.** (2011) A multidisciplinary study on the effects of phloem-limited viruses on the agronomical performance and berry quality of Vitis vinifera cv. Nebbiolo. *J Proteomics*, **75**, 306–315.

**González-Agüero, M., García-Rojas, M., Genova, A. Di, Correa, J., Maass, A., Orellana, A. and Hinrichsen, P.** (2013) Identification of two putative reference genes from grapevine suitable for gene expression analysis in berry and related tissues derived from RNA-Seq data. *BMC Genomics*, **14**, 878.

**Gresele, P., Cerletti, C., Guglielmini, G., Pignatelli, P., Gaetano, G. de and Violi, F.** (2011) Effects of resveratrol and other wine polyphenols on vascular function: An update. *J. Nutr. Biochem.*, **22**, 201–211.

**Grimplet, J., Adam-Blondon, A.-F., Bert, P.-F., et al.** (2014) The grapevine gene nomenclature system. *BMC Genomics*, **15**, 1077.

**Grimplet, J., Martínez-zapater, J.M. and Carmona, M.J.** (2016) Structural and functional annotation of the MADS-box transcription factor family in grapevine. *BMC Genomics*, 1–23.

**Guo, L., Gao, Z. and Qian, Q.** (2014) Application of resequencing to rice genomics, functional genomics and evolutionary analysis. *Rice (N. Y).*, **7**, 4.

**Hyma, K.E., Barba, P., Wang, M., Londo, J.P., Acharya, C.B., Mitchell, S.E., Sun, Q., Reisch, B. and Cadle-Davidson, L.** (2015) *Heterozygous Mapping Strategy (HetMappS) for High Resolution Genotyping-By-Sequencing Markers: A Case Study in Grapevine*,.

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

**Jones, L., Riaz, S., Morales-Cruz, A., Amrine, K.C., McGuire, B., Gubler, W.D., Walker, M.A. and Cantu, D.** (2014) Adaptive genomic structural variation in the grape powdery mildew pathogen, *Erysiphe necator*. *BMC Genomics*, **15**, 1081.

**Kim, T.-S., He, Q., Kim, K.-W., et al.** (2016) Genome-wide resequencing of KRICE_CORE reveals their potential for future breeding, as well as functional and evolutionary studies in the post-genomic era. *BMC Genomics*, **17**, 408.

**Kobayashi, S., Goto-yamamoto, N. and Hirochika, H.** (2004) Mutations in Grape Skin Color. , **304**, 8602.

**Li, H., Xia, N. and Förstermann, U.** (2012) Cardiovascular effects and molecular targets of resveratrol. *Nitric Oxide*, **26**, 102–110.

**Licausi, F., Giorgi, F.M., Zenoni, S., Osti, F., Pezzotti, M. and Perata, P.** (2010) Genomic and transcriptomic analysis of the AP2/ERF superfamily in Vitis vinifera. *BMC Genomics*, **11**, 719.

**Liu, J., Chen, N., Chen, F., Cai, B., Dal Santo, S., Tornielli, G.B., Pezzotti, M. and Cheng, Z.-M.M.** (2014) Genome-wide analysis and expression profile of the bZIP transcription factor gene family in grapevine (Vitis vinifera). *BMC Genomics*, **15**, 281.

**Mahanil, S., Ramming, D., Cadle-Davidson, M., Owens, C., Garris, A., Myles, S. and Cadle-Davidson, L.** (2012) Development of marker sets useful in the early selection of Ren4 powdery mildew resistance and seedlessness for table and raisin grape breeding. *Theor. Appl. Genet.*, **124**, 23–33.

**Majoul, T., Bancel, E., Tribo"\i, E., Hamida, J. Ben and Branlard, G.** (2003) Proteomic analysis of the effect of heat stress on hexaploid wheat grain: Characterization of heat-responsive proteins from total endosperm. *Proteomics*, **3**, 175–183.

**Martin, D.M., Aubourg, S., Schouwey, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T. and Bohlmann, J.** (2010) Functional annotation, genome organization and phylogeny of the grapevine (Vitis vinifera) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.*, **10**, 226.

**Matus, J.T., Aquea, F., Espinoza, C., et al.** (2014) Inspection of the grapevine BURP superfamily highlights an expansion of RD22 genes with distinctive expression features in berry development and ABA-mediated stress responses. *PLoS One*, **9**.

**Mica, E., Piccolo, V., Delledonne, M., et al.** (2009) High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera. *BMC Genomics*, **10**, 558.

**Migicovsky, Z., Sawler, J., Money, D., et al.** (2016) Genomic ancestry estimation quantifies use of wild species in grape breeding. *BMC Genomics*, **17**, 478.

**Milli, A., Cecconi, D., Bortesi, L., et al.** (2012) Proteomic analysis of the compatible interaction between Vitis vinifera and Plasmopara viticola. *J. Proteomics*, **75**, 1284–1302.

**Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., Shinozaki, K. and Tran, L.S.P.** (2013) TreeTFDB: An integrative database of the transcription factors from six economically important tree crops for functional predictions and comparative and functional genomics. *DNA Res.*, **20**, 151–162.

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

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

**Myles, S., Boyko, A.R., Owens, C.L., 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, G.Y., Buckler, E. and Ware, D.** (2010) Rapid genomic characterization of the genus Vitis. *PLoS One*, **5**.

**Naithani, S., Raja, R., Waddell, E.N., Elser, J., Gouthu, S., Deluc, L.G. and Jaiswal, P.** (2014) VitisCyc: a metabolic pathway knowledgebase for grapevine (Vitis vinifera). *Front. Plant Sci.*, **5**, 644.

**Nascimento-Gavioli, M.C.A., Agapito-Tenfen, S.Z., Nodari, R.O., Welter, L.J., Sanchez Mora, F.D., Saifert, L., Silva, A.L. da and Guerra, M.P.** (2016) Proteome of Plasmopara viticola-infected Vitis vinifera provides insights into grapevine Rpv1/Rpv3 pyramided resistance to downy mildew. *J Proteomics*.

**Nwafor, C.C., Gribaudo, I., Schneider, A., Wehrens, R., Grando, M.S. and Costantini, L.** (2014) Transcriptome analysis during berry development provides insights into co-regulated and altered gene expression between a seeded wine grape variety and its seedless somatic variant. *BMC Genomics*, **15**, 1030.

**Østergaard, O., Melchior, S., Roepstorff, P. and Svensson, B.** (2002) Initial proteome analysis of mature barley seeds and malt. *Proteomics*, **2**, 733–739.

**Paim Pinto, D.L., Brancadoro, L., Dal Santo, S., Lorenzis, G. De, Pezzotti, M., Meyers, B.C., Pè, M.E. and Mica, E.** (2016) The Influence of Genotype and Environment on Small RNA Profiles in Grapevine Berry. *Front. Plant Sci.*, **7**.

**Pantaleo, V., Vitali, M., Boccacci, P., Miozzi, L., Cuozzo, D., Chitarra, W., Mannini, F., Lovisolo, C. and Gambino, G.** (2016) Novel functional microRNAs from virus-free and infected Vitis vinifera plants under water stress. *Sci. Rep.*, **6**, 20167.

**Parage, C., Tavares, R., Rety, S., et al.** (2012) Structural, Functional, and Evolutionary Analysis of the Unusually Large Stilbene Synthase Gene Family in Grapevine. *Plant Physiol.*, **160**, 1407–1419.

**Peck, S.C.** (2005) Update on proteomics in Arabidopsis. Where do we go from here? *Plant Physiol*, **138**, 591–599.

**Perazzolli, M., Moretto, M., Fontana, P., Ferrarini, A., Velasco, R., Moser, C., Delledonne, M. and Pertot, I.** (2012) Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. *BMC Genomics*, **13**, 660.

**Pervaiz, T., Haifeng, J., Salman Haider, M., Cheng, Z., Cui, M., Wang, M., Cui, L., Wang, X. and Fang, J.** (2016) Transcriptomic Analysis of Grapevine (cv. Summer Black) Leaf, Using the Illumina Platform. *PLoS One*, **11**, e0147369.

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

**Potenza, E., Racchi, M.L., Sterck, L., Coller, E., Asquini, E., Tosatto, S.C.E., Velasco, R., Peer, Y. Van de and Cestaro, A.** (2015) Exploration of alternative splicing events in ten different grapevine cultivars. *BMC Genomics*, **16**, 706.

**Pulvirenti, A., Giugno, R., Distefano, R., et al.** (2015) A knowledge base for Vitis vinifera functional analysis. *BMC Syst. Biol.*, **9 Suppl 3**, S5.

**Ramos, M.J., Coito, J., Silva, H., Cunha, J., Costa, M.M. and Rocheta, M.** (2014) Flower development and sex specification in wild grapevine. *BMC Genomics*, **15**, 1095.

**Ramos, M.J.N., Coito, J.L., Silva, H.G., Cunha, J., Costa, M.M.R. and Rocheta, M.** (2014) Flower development and sex specification in wild grapevine. *BMC Genomics*, **15**, 1095.

**Renaud, S. and Lorgeril, M. de** (1992) Wine, alcohol, platelets, and the French paradox for coronary heart disease. *Lancet*, **339**, 1523–1526.

**Rienth, M., Torregrosa, L., Luchaire, N., Chatbanyong, R., Lecourieux, D., Kelly, M.T. and Romieu, C.** (2014) Day and night heat stress trigger different transcriptomic responses in green and ripening grapevine (vitis vinifera) fruit. *BMC Plant Biol.*, **14**, 108.

**Salekdeh, G.H., Siopongco, J., Wade, L.J., Ghareyazie, B. and Bennett, J.** (2002) Proteomic analysis of rice leaves during drought stress and recovery. *Proteomics*, **2**, 1131–1145.

**Shangguan, L., Kayesh, E., Leng, X., Sun, X., Korir, N.K., Mu, Q. and Fang, J.** (2013) Whole genome identification and analysis of FK506-binding protein family genes in grapevine (Vitis vinifera L.). *Mol. Biol. Rep.*, **40**, 4015–4031.

**Shiratake, K. and Suzuki, M.** (2016) Omics studies of citrus, grape and rosaceae fruit trees. *Breed Sci*, **66**, 122–138.

**Silva, C. Da, Zamperin, G., Ferrarini, A., 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.

**Skylas, D.J., Copeland, L., Rathmell, W.G. and Wrigley, C.W.** (2001) The wheat-grain proteome as a basis for more efficient cultivar identification. *Proteomics*, **1**, 1542–1546.

**Sun, X., Fan, G., Su, L., Wang, W., Liang, Z., Li, S. and Xin, H.** (2015) Identification of cold-inducible microRNAs in grapevine. *Front. Plant Sci.*, **6**, 1–13.

**Sun, X., Xie, Z., Zhang, C., Mu, Q., Wu, W., Wang, B. and Fang, J.** (2016) A characterization of grapevine of GRAS domain transcription factor gene family. *Funct. Integr. Genomics*.

**Sweetman, C., Wong, D.C., Ford, C.M. and Drew, D.P.** (2012) Transcriptome analysis at four developmental stages of grape berry (Vitis vinifera cv. Shiraz) provides insights into regulated and coordinated gene expression. *BMC Genomics*, **13**, 691.

**Tavares, S., Wirtz, M., Beier, M.P., Bogs, J., Hell, R. and Amâncio, S.** (2015) Characterization of the serine acetyltransferase gene family of Vitis vinifera uncovers differences in regulation of OAS synthesis in woody plants. *Front. Plant Sci.*, **6**, 74.

**Thomson, D.W., Bracken, C.P. and Goodall, G.J.** (2011) Experimental strategies for microRNA target identification. *Nucleic Acids Res.*, **39**, 6845–6853.

**Vannozzi, A., Dry, I.B., Fasoli, M., Zenoni, S. and Lucchin, M.** (2012) Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. *BMC Plant Biol.*, **12**, 130.

**Velasco, R., Zharkikh, A., Troggio, M., et al.** (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, **2**.

**Venturini, L., Ferrarini, A., Zenoni, S., et al.** (2013) De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. *BMC Genomics*, **14**, 41.

**Vitulo, N., Forcato, C., Carpinelli, E., 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.

**Wang, C., Han, J., Liu, C., Kibet, K., Kayesh, E., Shangguan, L., Li, X. and Fang, J.** (2012) Identification of microRNAs from Amur grape (vitis amurensis Rupr.) by deep sequencing and analysis of microRNA variations with bioinformatics. *BMC Genomics*, **13**, 122.

**Wang, C., Wang, X., Kibet, N.K., Song, C., Zhang, C., Li, X., Han, J. and Fang, J.** (2011) Deep sequencing of grapevine flower and berry short RNA library for discovery of novel microRNAs and validation of precise sequences of grapevine microRNAs deposited in miRBase. *Physiol. Plant.*, **143**, 64–81.

**Wang, M., Vannozzi, A., Wang, G., Liang, Y.-H., Tornielli, G.B., Zenoni, S., Cavallini, E., Pezzotti, M. and Cheng, Z.-M. (Max)** (2014) Genome and transcriptome analysis of the grapevine (Vitis vinifera L.) WRKY gene family. *Hortic. Res.*, **1**, 16.

**Wen, J., Xiong, Z., Nie, Z.L., et al.** (2013) Transcriptome Sequences Resolve Deep Relationships of the Grape Family. *PLoS One*, **8**, 1–9.

**Wienkoop, S., Baginsky, S. and Weckwerth, W.** (2010) Arabidopsis thaliana as a model organism for plant proteome research. *J Proteomics*, **73**, 2239–2248.

**Wijk, K.J. van** (2001) Challenges and prospects of plant proteomics. *Plant Physiol*, **126**, 501–508.

**Wong, D.C.J., Sweetman, C., Drew, D.P. and Ford, C.M.** (2013) VTCdb: a gene co-expression database for the crop species Vitis vinifera (grapevine). *BMC Genomics*, **14**, 882.

**Xin, H., Zhu, W., Wang, L., et al.** (2013) Genome wide transcriptional profile analysis of Vitis amurensis and Vitis vinifera in response to cold stress. *PLoS One*, **8**, e58740.

**Xu, W., Li, R., Zhang, N., Ma, F., Jiao, Y. and Wang, Z.** (2014) Transcriptome profiling of Vitis amurensis, an extremely cold-tolerant Chinese wild Vitis species, reveals candidate genes and events that potentially connected to cold stress. *Plant Mol. Biol.*, **86**, 527–541.

**Yang, S., Fresnedo-Ramírez, J., Sun, Q., et al.** (2016) Next generation mapping of enological traits in an F2 interspecific grapevine hybrid family. *PLoS One*, **11**, 1–19.

**Yang, Y., He, M., Zhu, Z., Li, S., Xu, Y., Zhang, C., Singer, S.D. and Wang, Y.** (2012) Identification of the dehydrin gene family from grapevine species and analysis of their responsiveness to various forms of abiotic and biotic stress. *BMC Plant Biol.*, **12**, 140.

**Yu** (2013) Callose Synthase Family Genes Involved in the Grapevine Defense Response to Downy Mildew Disease. *Phytopathol.*, 56–64.

**Zamboni, A., Carli, M. Di, Guzzo, F., et al.** (2010) Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks. *Plant Physiol*, **154**, 1439–1459.

**Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., Bellin, D., Pezzotti, M. and Delledonne, M.** (2010) Characterization of Transcriptional Complexity during Berry Development in Vitis vinifera Using RNA-Seq. *Plant Physiol.*, **152**, 1787–1795.

**Zhang, K., Han, Y.-T., Zhao, F.-L., Hu, Y., Gao, Y.-R., Ma, Y.-F., Zheng, Y., Wang, Y.-J. and Wen, Y.-Q.** (2015) Genome-wide Identification and Expression Analysis of the CDPK Gene Family in Grape, Vitis spp. *BMC Plant Biol.*, **15**, 164.

**Zhang, Z., Qi, S., Tang, N., et al.** (2014) Discovery of Replicating Circular RNAs by RNA-Seq and Computational Algorithms. *PLoS Pathog.*, **10**.

**Zhao, T., Xia, H., Liu, J. and Ma, F.** (2014) The gene family of dehydration responsive element-binding transcription factors in grape (Vitis vinifera): Genome-wide identification and analysis, expression profiles, and involvement in abiotic stress resistance. *Mol. Biol. Rep.*, **41**, 1577–1590.

# Objectif de la thèse

La vigne synthétise une palette de métabolites secondaires très large et diverse selon les cépages. L'équipe dans laquelle a été réalisée cette thèse se focalise sur l'étude des métabolites secondaires impliqués dans les défenses de la vigne face à ses pathogènes mais aussi sur les arômes et les précurseurs d'arômes dans les raisins. Plusieurs travaux, déjà publiés ou en cours de publication, ont mis en évidence une amplification particulière de certaines familles de gènes dans le génome de la vigne, par comparaison à d'autres génomes de plantes (Matus *et al.*, 2014; Velasco *et al.*, 2007). Par exemple, la famille des gènes codant pour les terpènes synthases (*TPS*) compte 152 gènes dans le génome de la vigne mais moins de 50 dans la plupart des autres plantes telles que *Arabidopsis thaliana*, le riz, la tomate ou le peuplier (Martin *et al.*, 2010). C'est également le cas des familles de gènes codant pour certains *cytochromes P450* (CYP82 avec 69 gènes, CYP71 avec 51 gènes et CYP76 avec 42 gènes), potentiellement impliqués dans la production d'arômes et les gènes codant pour les stilbènes synthases (*STS* avec 48 gènes), jouant un rôle dans la défense de la vigne. De plus, ces familles de gènes sont souvent organisées en groupe de plusieurs membres à proximité les uns des autres appelés clusters. Des pressions de sélection négative s'exercent notamment sur la famille des gènes *STS* ce qui tendrait à maintenir l'organisation de cette famille en clusters comptant de nombreux gènes (Parage *et al.*, 2012).

Ma thèse vise à proposer des hypothèses expliquant l'organisation structurale de ces familles de gènes et ainsi à mieux comprendre pourquoi certaines familles présentent une amplification dans le génome de la vigne. Une hypothèse serait que les gènes d'une même famille ne s'exprimeraient pas dans les mêmes conditions et donc qu'ils seraient soumis à des régulations différentes conférant une grande possibilité d'adaptation pour la vigne. Un second objectif de ma thèse est d'estimer le niveau de conservation des gènes de ces familles dans les génomes d'autres cépages et d'autres espèces du genre *Vitis* dans le but d'appréhender la dynamique évolutive de ces familles.

Pour aborder ces différentes questions, des approches bioinformatiques ont été utilisées. Les gènes *cytochromes P450* et gènes R de type *NBS* ont été annotés de manière manuelle dans le génome de référence de la vigne, afin obtenir les annotations de bonne qualité nécessaires à la fiabilité des études transcriptomiques et de variations structurales. L'expression des gènes endo-β-1,3-glucanases, *STS* et *cytochromes P450* a été quantifiée afin d'étudier leur implication éventuelle dans les mécanismes de résistances et de production d'arômes. Pour ce faire, un outil a été développé durant cette thèse pour estimer le niveau d'expression des gènes à partir de données RNA-Seq disponibles dans les banques de données publiques. Parallèlement, des données de reséquençage d'ADN, de 56 cépages et espèces de vigne, ont été analysées afin de déterminer les variations structurales de type CNV au sein des familles de gènes à domaine *NBS* et des gènes *STS*.

# Partie 1 : Création d'un outil d'analyse du transcriptome de la vigne

# Introduction

Le développement des nouvelles technologies de séquençage a conduit à une augmentation importante des données transcriptomiques disponibles chez la vigne. Grâce à l'accessibilité des bases de données publiques, l'obtention et la réutilisation de données transcriptomiques publiées est grandement facilitée, ce qui offre un large panel d'expériences dans différents tissus de vigne soumis à diverses conditions. En particulier, le RNA-Seq s'est largement répandu avec différentes techniques utilisées comme la séquence par synthèse (Illumina), le pyro-séquençage (454), la détection de proton (Ion torrent) ou le séquençage en temps réel (PacBio) (Han *et al.*, 2015). Les champs d'application du RNA-Seq vont de la quantification d'expression à la détection d'allèles en passant par l'expression différentielle et un large panel d'outils sont disponibles et ont été rapportés par Han *et al.* (2015).

Bien que l'intérêt d'une étude puisse résider dans un nombre limité ou une famille particulière de gènes, l'exhaustivité des données RNA-Seq permet d'analyser l'expression des gènes à l'échelle du génome entier et ceci, dans un grand nombre conditions expérimentales. Cela permet l'étude des gènes qui seraient co-exprimés avec les gènes d'intérêt et des réseaux de co-expression de manière générale. Traiter des données RNA-Seq nécessite des outils particuliers et cela a été montré par Hong et ses collègues (2013). En effet, un outil de reconstruction de réseaux de co-expression spécifique pour le RNA-Seq a été développé et a montré des performances supérieures aux autres méthodes existantes.

Des outils d'analyses transcriptomiques ont été développés depuis quelques années chez la vigne. L'eFP browser de vigne proposant une vue d'ensemble de l'expression d'un gène donné dans les différents organes de la plante à partir de données de puces et RNA-Seq (Fasoli *et al.*, 2012). La base de données BIOWINE permet l'analyse de l'expression de gènes dans les cultivars Nero d'Avola et Nerello Mascalese avec la possibilité d'observer l'expression différentielle entre différentes conditions expérimentales (Pulvirenti *et al.*, 2015). Plus récemment, l'outil VESPUCCI propose une visualisation de l'expression et une reconstruction de réseaux ainsi que des informations sur les gènes analysés (Moretto *et al.*, 2016). Ces outils ne permettent cependant pas à l'utilisateur de sélectionner facilement et à sa guise les expériences et les résultats associés pour réaliser des analyses complémentaires.

Une réflexion a été engagée afin de développer un outil facile d’accès et d’utilisation pour le plus grand nombre, permettant de visualiser directement l’expression des gènes d'intérêt dans une sélection d’expériences et l'obtention de ces données pour un usage ultérieur. Afin d’exploiter au mieux les données RNA-Seq disponibles chez la vigne, nous avons décidé de développer un pipeline d’analyse à l'échelle du génome entier et de calculer l’expression normalisée de tous les gènes annotés sur le génome de référence de la vigne. Même si seules quelques familles de gènes (endo-β-1,3-glucanases, cytochromes P450, stilbène synthases) ont été étudiées au cours de cette thèse, l'outil développé peut être utilisé pour analyser l'expression de n'importe quel gène d'intérêt chez la vigne. L’intérêt particulier qui a été porté aux résultats de ce pipeline par différents chercheurs de l'unité, intéressés par le potentiel que présentent ces informations sur l’expression de tous les gènes annotés du génome de la vigne, nous a conforté dans l’idée de la création d’un outil dédié. Nous avons donc décidé, en plus du pipeline, de développer une base de données interrogative sous la forme d’un site internet actuellement en développement. Cela permettra d’obtenir, par simple recherche avec l’identifiant d’un ou de plusieurs gènes, les résultats d’expression dans une sélection des expériences disponibles.

Les méthodes utilisées pour développer cet outil sont détaillées ci-après ainsi que les résultats et un exemple d’utilisation.

## Matériel et méthodes

Les données RNA-Seq de vigne ont été téléchargées à partir de la base de données publique SRA du NCBI (http://www.ncbi.nlm.nih.gov/sra). Dans un premier temps, les 81 jeux de données sélectionnés ont été générés dans le cadre de différents projets et travaux : Da Silva *et al.* 2013; Jones *et al.* 2014; Perazzolli *et al.* 2012; Ramos *et al.* 2014; Sweetman *et al.* 2012; Vannozzi *et al.* 2012; Venturini *et al.* 2013 et du projet Vitaroma. La technique de séquençage utilisée est Illumina et la longueur des reads varie de 38 à 100 pb en paired- ou en single-end selon les expériences. Ces données représentent aussi bien de l'expression dans des fleurs, de jeunes baies, des baies mûres ou des feuilles. Concernant les feuilles, celles-ci peuvent être saines, soumises à un stress biotique ou à un stress abiotique. Différents génotypes sont utilisés dans ces expériences. Un tableau récapitulatif présentant les expériences plus en détail est présenté en Annexe 1. Une fois téléchargées sous le format SRA, les données sont converties au format fastq, en utilisant la commande fastq-dump disponible dans le package SRA Toolkit Package version 2.3.4 (http://www.ncbi.nlm.nih.gov/books/NBK158900), permettant d'obtenir les reads bruts.

Les reads sont ensuite alignés sur le génome de référence de la vigne PN40024 12x.1 en utilisant la version 2013-11-27 de GSNAP (Wu and Nacu, 2010) avec les paramètres : -B 4, -N 1, -n 3, --nofails et le protocole de qualité adéquat, illumina ou Sanger, en fonction des différentes expériences. Cette étape permet l'obtention d'alignements pour lesquels les reads sont triés selon une valeur d'edit distance (nombre de modifications nécessaires dans la séquence d'intérêt pour être identique au génome de référence). Concrètement, le génome de référence étant *Vitis vinifera* PN40024 (Jaillon *et al.*, 2007), si un autre *Vitis vinifera* ou *Vitis sylvestris* est aligné sur le génome de référence, le seuil sera fixé à 95% d'identité ce qui correspond à une edit distance de 5 si les reads ont une longueur de 100 pb. En utilisant ces critères, si un reads s'aligne à plusieurs positions, uniquement le meilleur alignement (plus haut pourcentage d'identité) sera conservé.

Un second tri est ensuite effectué afin de ne conserver que les reads s'alignant de manière unique. En effet, les alignements multiples de reads dans les portions dupliquées du génome occasionneraient un potentiel biais de sur-expression par rapport aux portions de faible complexité. Dans le cas de reads paired-end, ceux qui ne sont pas retrouvés par paires sont aussi éliminés. Ces opérations de triage sont réalisées en utilisant un script Perl développé dans l'équipe.

Afin quantifier l'expression des gènes du génome de référence, la première version de l'annotation de la vigne a été téléchargée à partir de la base de données ''Grape Genome Database'' hébergée au CRIBI. À cette annotation automatique, ont été ajoutées les annotations, vérifiées manuellement, des gènes codant pour les stilbène synthases, terpène synthases, O-methyl transférases, cytochromes P450 et gènes de résistance de type NBS. Afin d'éviter les ambiguïtés, les annotations automatiques présentant un chevauchement avec les annotations manuelles ont été écartées de l'analyse.

Les reads alignés sur chaque annotation ont été comptés en utilisant la commande htseq-count de la suite HTSeq version 0.6.0 (Anders *et al.*, 2015) avec les paramètres : -m intersection-nonempty et -s no. Les FPKM (Fragments Per Kilo base of exon per Million fragments mapped), représentant une valeur d'expression normalisée permettant la comparaison de l'expression des gènes entre eux mais aussi entre les différentes expériences, ont ensuite été calculés en utilisant un script R développé dans le cadre de la thèse.

# Résultats

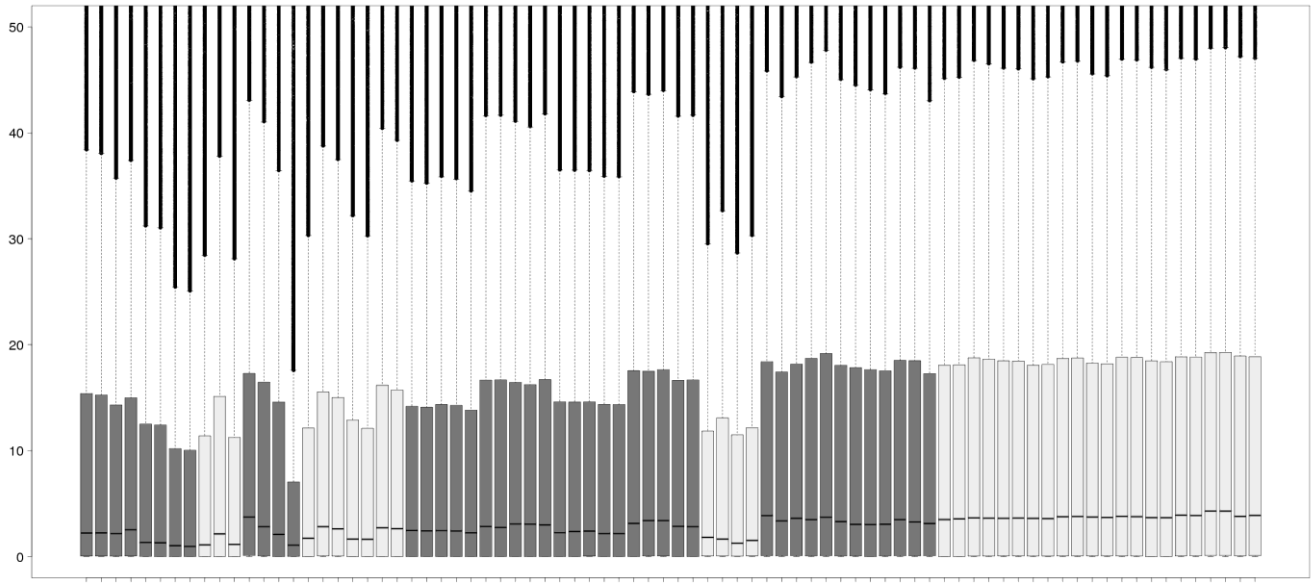



**Figure 1 : Vérification de la normalisation des données RNA-Seq.** Boîtes à moustaches des différentes expériences analysées. En abscisses se trouvent les différentes expériences et en ordonnées sont les valeurs de FPKM. La médiane est représentée par une barre noire dans les boîtes. L'alternance de couleur matérialise les différents jeux de données avec, de gauche à droite les données de Vitaroma, Venturini, Da Silva, Vannozzi, Perazzolli, Sweetman, Jones, Ramos.

La Figure 1 représente la distribution des données dans les différentes expériences RNA-Seq analysées. Les médianes, varient entre 0,9 et 4,3. Il est cependant possible d'observer un léger effet dû aux conditions expérimentales. Par exemple, un léger écart entre deux groupes de médianes correspondants à deux groupes d'expériences générées dans le cadre de projets différents.

Ces données d'expression, pour tous les gènes, dans toutes les expériences étudiées, forment une base de données qui sera utilisée pour le développement de l'outil disponible sur internet et permettant la récupération du profil d'expression pour un ou plusieurs gènes donnés.

Les résultats d'expression dans les différentes expériences RNA-Seq analysées ont été utilisés dans le cadre d'un travail sur les endo-β-1,3-glucanases de vigne afin d'identifier les gènes candidats dont l'expression est induite lors d'une infection par le mildiou ou l'oïdium. En effet, ces enzymes participent à l'hydrolyse des membranes cellulaires de champignons infectant la vigne et participent donc à la résistance de la plante face à de nombreux pathogènes fongiques. Trois enzymes ont pu être identifiées dont une montrant une forte activité antimicrobienne malgré une faible activité *in vitro*. Ce travail, présenté dans l'article ci-après, est le fruit d'une collaboration avec d'autres chercheurs de l'INRA de Colmar. Mon implication a consisté à récupérer l'expression de gènes, à partir d'une liste de gènes candidats, dans toutes les conditions disponibles permettant de mettre en évidence que certaines enzymes sont spécifiquement plus exprimées lors d'une infection au mildiou ou à l'oïdium

## Conclusion

Contrairement à la base de données BIOWINE (Pulvirenti *et al.*, 2015), à l'eFP browser de vigne (Fasoli *et al.*, 2012) et à VESPUCCI (Moretto *et al.*, 2016), l'outil développé dans cette thèse se focalise sur les données RNA-Seq et propose un large panel d'expériences dans diverses conditions expérimentales comme décrit en Annexe 1. Cet outil, simple, pratique et convivial présente un intérêt particulier pour la communauté de chercheurs travaillant sur vigne, comme en témoigne l'intérêt qu'il a déjà suscité auprès des chercheurs de l'unité durant son développement. En effet, il permet une sélection des expériences individuellement, une visualisation numérique des valeurs d'expression et leur exportation pour un usage externe. La Table 1 reprend quelques caractéristiques des différents outils disponibles afin de les comparer avec notre nouvel outil.

De par la démocratisation du RNA-Seq, le nombre de jeux de données disponibles dans les banques de données publiques va continuer d'augmenter dans le futur. Le maintien de cet outil à jour ne fera qu'augmenter son intérêt grâce à la possibilité pour n'importe quel utilisateur d'obtenir un profil d'expression dans un nombre croissant d'expériences.

**Table 1 : Caractéristiques techniques et pratiques des différents outils.** Comparaison de VESPUCCI, BIOWINE et de l'outil développé durant la thèse.

| Critère | VESPUCCI | BIOWINE | Outil développé |
|---|---|---|---|
| RNA-Seq | < 15 % | 100 % | 100 % |
| Exhaustivité des données RNA-Seq | Non (peu de données RNA-Seq) | Non (limité à deux cultivars) | Oui (pas de limites) |
| Choix des comparaisons d'expériences | Non | Oui | Oui |
| Valeur chiffrée d'expression | Non | Non | Oui |

## Perspectives

À ce jour, le pipeline n'est pas entièrement automatisé, l'objectif serait qu'à partir d'un seul script, avec les paramètres adéquats, l'intégralité des opérations soit réalisée de manière automatique jusqu'à arriver au résultat final, c'est-à-dire l'obtention des FPKM pour tous les gènes d'intérêt. Le site internet est actuellement sous la forme d'un prototype et n'est pas encore accessible au public. L'objectif sera de le rendre complètement opérationnel à court terme. Le fonctionnement de cet outil est brièvement présenté dans les Figures 2 et 3.

D'autres méthodes de normalisation sont disponibles et ont été rapportées par Huang *et al.* (2015). Il serait intéressant de proposer les résultats en utilisant d'autres outils de normalisation en plus des FPKM, comme par exemple en utilisant les packages R edgeR (Robinson *et al.*, 2009) et DESeq2 (Love *et al.*, 2014).

Enfin, le maintien de cet outil à jour, en ajoutant les nouveaux jeux de données RNA-Seq au fur et à mesure de leur disponibilité, sera un travail sur le long terme, mais apportera beaucoup pour la pérennité de cet outil.


**Figure 3 : exemple d'utilisation du site internet.** Dans le cadre de gauche sont entrés les identifiants des gènes pour lesquels un résultat est demandé. Une requête pour dix gènes peut être effectuée. Dans le menu déroulant à gauche se trouvent toutes les expériences disponibles à sélectionner pour obtenir les résultats correspondants. Il est possible de les sélectionner individuellement.

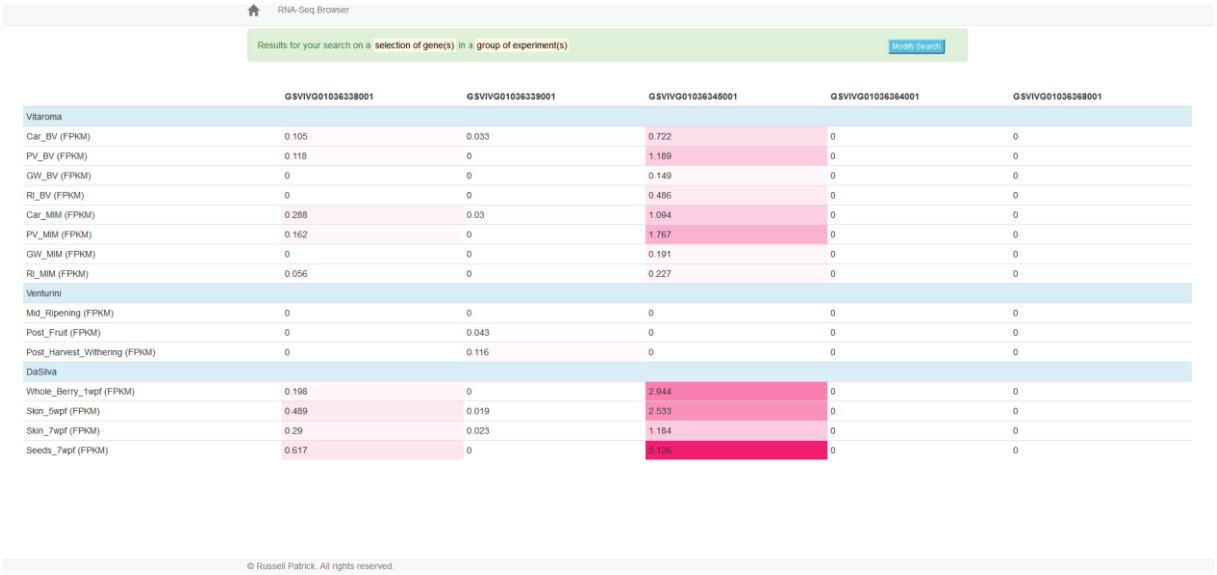



**Figure 4 : résultats de l'interrogation de la base de données.** En haut est disponible l'option pour lancer une nouvelle recherche. La première colonne présente le nom des expériences ainsi que le premier auteur de la publication où sont décrites les données ou le projet ayant généré les données. Les autres colonnes représentent les résultats avec en en-tête les identifiants des gènes. Chaque ligne représente une expérience différente. Les valeurs d'expression sont colorées selon leur valeur de FPKM allant du blanc, pas d'expression, au rose foncé, haute expression. L'échelle de couleur est ajustée en fonction des résultats affichés.

**Annexe 1** : tableau récapitulatif des expériences RNA-Seq analysées.

| Groupe / Projet | Espèce / Variété | Reads | Conditions étudiées |
|---|---|---|---|
| Da Silva | Tannat | Illumina<br>2x100 pb | Baies 1 semaine après floraison<br>Peaux 5 et 7 semaines après floraison<br>Graines 7 semaines après floraison |
| Jones | Carignan | Illumina<br>2x100 pb | Feuilles infectées par l'oïdium 12, 24, 72 et 144 hpi |
| Perazzolli | Pinot Noir | Illumina<br>2x100 pb | Feuilles contrôles<br>Feuilles traitées au T39<br>Feuilles infectées par le mildew 24 hpi<br>Feuilles traitées au T39 et infectées par le mildew 24 hpi |
| Ramos | Vitis sylvestris | Illumina 51 pb | Fleurs aux stades développementaux B, D, G et H |
| Sweetman | Shiraz | Illumina 100 pb | Baies vertes<br>Baies avant véraison<br>Baies après véraison<br>Baies mûres |
| Vannozzi | Pinot Noir | Illumina<br>2x39 pb | Feuilles contrôles<br>Feuilles blessées 24 et 48 hpi<br>Feuilles traitées sous UV 24 et 48 hpi<br>Feuilles infectées par le mildew 24 et 48 hpi |
| Venturini | Corvina | Illumina<br>2x51 pb | Baies vertes<br>Baies mi-maturité<br>Baies mi-pourries (2 mois après récolte) |
| Vitaroma | Carménère<br>Gewurztraminer<br>Petit Verdot<br>Riesling | Illumina<br>2x38 pb | Baies vertes<br>Baies mi-maturité |

pb = paire de bases. hpi = heures post inoculation.

# Références bibliographiques

**Anders, S., Pyl, P.T. and Huber, W.** (2015) HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

**Fasoli, M., Dal Santo, S., Zenoni, S., et al.** (2012) The Grapevine Expression Atlas Reveals a Deep Transcriptome Shift Driving the Entire Plant into a Maturation Program. *Plant Cell*, **24**, 3489–3505.

**Giorgi, F.M., Fabbro, C. Del and Licausi, F.** (2013) Comparative study of RNA-seq- and Microarray-derived coexpression networks in Arabidopsis thaliana. *Bioinformatics*, **29**, 717–724.

**Han, Y., Gao, S., Muegge, K., Zhang, W. and Zhou, B.** (2015) Advanced applications of RNA sequencing and challenges. *Bioinform. Biol. Insights*, **9**, 29–46.

**Hong, S., Chen, X., Jin, L. and Xiong, M.** (2013) Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.*, **41**, 1–15.

**Huang, H.C., Niu, Y. and Qin, L.X.** (2015) Differential expression analysis for RNA-Seq: An overview of statistical methods and computational software. *Cancer Inform.*, **14**, 57–67.

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

**Jones, L., Riaz, S., Morales-Cruz, A., Amrine, K.C., McGuire, B., Gubler, W.D., Walker, M.A. and Cantu, D.** (2014) Adaptive genomic structural variation in the grape powdery mildew pathogen, *Erysiphe necator*. *BMC Genomics*, **15**, 1081.

**Love, M.I., Huber, W. and Anders, S.** (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

**Moretto, M., Sonego, P., Pilati, S., et al.** (2016) VESPUCCI: Exploring Patterns of Gene Expression in Grapevine. *Front. Plant Sci.*, **7**, 1–11.

**Perazzolli, M., Moretto, M., Fontana, P., Ferrarini, A., Velasco, R., Moser, C., Delledonne, M. and Pertot, I.** (2012) Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. *BMC Genomics*, **13**, 660.

**Pulvirenti, A., Giugno, R., Distefano, R., et al.** (2015) A knowledge base for Vitis vinifera functional analysis. *BMC Syst. Biol.*, **9 Suppl 3**, S5.

**Ramos, M.J., Coito, J., Silva, H., Cunha, J., Costa, M.M. and Rocheta, M.** (2014) Flower development and sex specification in wild grapevine. *BMC Genomics*, **15**, 1095.

**Robinson, M.D., McCarthy, D.J. and Smyth, G.K.** (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

**Silva, C. Da, Zamperin, G., Ferrarini, A., 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.

**Sweetman, C., Wong, D.C., Ford, C.M. and Drew, D.P.** (2012) Transcriptome analysis at four developmental stages of grape berry (Vitis vinifera cv. Shiraz) provides insights into regulated and coordinated gene expression. *BMC Genomics*, **13**, 691.

**Vannozzi, A., Dry, I.B., Fasoli, M., Zenoni, S. and Lucchin, M.** (2012) Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. *BMC Plant Biol.*, **12**, 130.

**Venturini, L., Ferrarini, A., Zenoni, S., et al.** (2013) De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. *BMC Genomics*, **14**, 41.

**Wu, T.D. and Nacu, S.** (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

# Partie 2 : Analyse transcriptomique de la famille des gènes cytochromes P450 de la vigne

Les cytochromes P450, ayant pour la majorité une fonction d'oxygénase, contrôlent beaucoup de fonctions différentes dans une plante. Certains gènes sont impliqués dans les défenses contre des pathogènes tandis que d'autres sont exprimés lors du développement des baies suggérant un rôle dans la production d'arômes. Les gènes des *cytochromes P450* sont organisés en différentes familles, physiquement proches, dont la fonction est généralement similaire. L'identité de séquence entre les différents gènes *P450* est cependant très variable allant de 10% à presque 100% d'identité.

Ce travail s'inscrit dans le cadre du projet ANR InteGrape, mené en collaboration avec l'équipe "Cytochromes P450 pour la biosynthèse des biopolymères, la signalisation et l'adaptation" de l'Institut de Biologie Moléculaire des Plantes de Strasbourg et l'équipe de Gabriel Marais au Laboratoire de Biométrie et Biologie Évolutive (UMR CNRS 5558, Lyon).

Le projet a pour but de comprendre l'impact des nombreux gènes impliqués dans la synthèse des terpènes sur la richesse des arômes et des vins. En effet, la vigne possède un grand nombre de gènes *TPS* et *cytochromes P450* impliqués dans la synthèse de ces terpènes. Le projet InteGrape a pour but de mieux comprendre les conséquences de l'amplification, même partielle, de ces familles de gènes sur la production et la diversification des arômes, ainsi que les mécanismes évolutifs liés à cette amplification

Les résultats provenant du pipeline décrit dans la partie 1 ont également été utilisés de manière exhaustive sur la famille des gènes *cytochromes P450* afin de caractériser leur expression. Certains représentants de cette famille sont plus exprimés en conditions de stress et pourraient jouer un rôle dans la production de composants spécifiques importants dans la défense de la plante. De plus, l'annotation et l'étude de l'organisation des gènes de cette famille ont montré une forte tendance des gènes *P450* à être organisés en clusters, comme cela a été observé pour les gènes codant pour les stilbènes synthases.

Les résultats obtenus sont présentés dans l'article ci-après. Le travail sur les P450 s'inscrit également dans le cadre de la thèse de Tina Ilc (IBMP, Strasbourg), soutenue en décembre 2015. En plus de l'analyse transcriptomique, j'ai contribué à l'identification et l'annotation des gènes codant pour les *cytochromes P450* dans le génome de référence de la vigne.

# Annotation, classification, genomic organization and expression of the *Vitis vinifera* CYPome

Tina Ilc[1], Gautier Arista[2], Raquel Tavares[3], Nicolas Navrot[1], Frédéric Choulet[4], Marc Fischer[2], Philippe Hugueney[2], Danièle Werck-Reichhart[1], Camille Rustenholz[2]

[1] Institute of Plant Molecular Biology, Centre National de la Recherche Scientifique, Université de Strasbourg, Strasbourg, France

[2] Université de Strasbourg, INRA, SVQV UMR-A 1131, F-68000 Colmar, France

[3] Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Université de Lyon 1, Lyon, France

[4] Laboratoire Structure et Evolution du Génome du Blé, Institut National de la Recherche Agronomique, Université Blaise Pascal, Clermont-Ferrand, France

## Abstract

Cytochromes P450 are enzymes that control a wide range of functions in plants, from hormonal signaling and biosynthesis of structural polymers, to defense or communication with other organisms, and represent one of the largest gene/protein families in the plant kingdom. The manual annotation of cytochrome P450 genes in the genome of *Vitis vinifera* PN40024 revealed 579 P450 sequences, including 279 complete genes. Most of the P450 sequences in grapevine genome are organized in physical clusters, resulting from tandem or segmental duplications. Although most of these clusters are small, some P450 families, such as CYP76 and CYP82, underwent multiple duplications and formed large clusters of homologous sequences. Analysis of gene expression revealed highly specific expression patterns, which are often shared within the genes in large physical clusters. Some of these genes are induced upon biotic stress, which points to their role in plant defense, whereas others are specifically activated during grape berry ripening and might be responsible for the production of berry-specific metabolites, such as aroma compounds. Our comprehensive gene annotation and expression analysis provide groundwork for further functional characterization of this major gene family in grapevine.

## Background

Grapevine (*Vitis vinifera L.*) is one of the oldest (Zohary and Spiegel-Roy, 1972) and economically the most important (Anon, 2015) fruit crops in the world. The majority of grapes produced worldwide are used in winemaking. Modern cultivated grapevine has been shaped by thousands of years of selection for traits such as berry size, sugar content or skin color (Myles *et al.*, 2011), but today viticulture is facing new challenges. In addition to pathogen pressure, it has to deal with climate change (Duchêne and Schneider, 2005; Duchêne *et al.*, 2010; Hannah *et al.*, 2013), and shift of consumer preference towards higher quality wines with a lower environmental impact (Bisson *et al.*, 2002; Borneman *et al.*, 2013). Traditional breeding is extremely difficult to apply in grapevine because of its long lifecycle, reduced fitness of progeny and complexity of quality traits (Gray *et al.*, 2014). Sequencing of the grapevine genome in 2007 (Jaillon *et al.*, 2007) and advances in the 'omics' techniques (Langridge and Fleury, 2011) set the stage for more efficient breeding solutions. The next crucial step towards improved grapevine varieties is the identification of genes underlying important traits, such as interactions with pathogens, fruit development and quality.

Many developmental as well as ecological functions in plants are controlled by cytochrome P450 oxygenases (Nelson and Werck-Reichhart, 2011; Bak *et al.*, 2011). These enzymes catalyze regio- and stereospecific insertion of an oxygen atom into small, hydrophobic substrates that range from terpenoids and fatty acids to amino acids and their derivatives, such as phenolic compounds. In the model plant *Arabidopsis thaliana* they control processes as diverse as plant growth and branching (Helliwell *et al.*, 1998; Booker *et al.*, 2005), flower (Anastasiou *et al.*, 2007) and fruit development (Ito and Meyerowitz, 2000), formation of lignin and surface biopolymers (Ehlting *et al.*, 2006; Wellesen *et al.*, 2001), emission of volatiles (Lee *et al.*, 2010) or plant-pathogen and plant-insect interactions (Nafisi *et al.*, 2007; Hansen, 2001). In crop plants, P450s have played major roles in shaping agriculturally-relevant traits, such as fruit size (Proc Natl Acad Sci U S A. 2013 Oct 15;110(42):17125-30). This makes cytochromes P450 attractive targets for further crop improvement.

Cytochromes P450 in plants evolved into many distinct families, which are usually defined as genes with 40% or higher protein sequence identity. Within one P450 family the biochemical function is often conserved across the plant kingdom. For example, enzymes from the CYP97 family are involved in carotenoid hydroxylation, CYP79s in the *N*-hydroxylation of amino acid to aldoximes, CYP75s in the hydroxylation of flavonoids, and CYP704s in fatty acid

hydroxylation to form the precursors to structural polymers sporopollenin and cutin (Hamberger and Bak, 2013). Members of other families, however, have divergent functions: some members of CYP72 family are involved in iridoid biosynthesis, whereas others oxidize triterpene substrates (Hamberger and Bak, 2013). These differences stem from different evolutionary pressures on genes with different functions. Families with essential functions, such as hormone metabolism or synthesis of biopolymers, are usually maintained at low copy number and high purifying selection, whereas families with adaptive functions expanded or "bloomed" in certain taxa (Feyereisen, 2011). A well-documented example is a bloom of CYP76M subfamily in rice (*Oryza sativa*), which consists of 11 genes and 2 pseudogenes. At least 4 members of this subfamily are involved in the biosynthesis of diterpenoid antifungal compounds (Swaminathan *et al.*, 2009; Wang *et al.*, 2012). They are clustered close together in the genome, which is another common feature of recently duplicated P450s and probably results from sequential tandem duplications (Feyereisen, 2011). Interestingly, CYP76 members from other plants, for example *Arabidopsis thaliana* or *Catharanthus roseus*, have a different biochemical function, namely oxidation of monoterpenols or their iridoid derivatives (Hofer *et al.*, 2014; Miettinen *et al.*, 2014). Recently expanded P450 families might therefore have interesting ecological functions, but those are more difficult to predict compared to functions of conserved P450 families. In addition, function of many P450 families is still unknown or poorly understood.

Previous annotation of P450s has highlighted some potentially interesting gene families in the highly heterozygous *V. vinifera* cv. Pinot Noir genome (Velasco *et al.*, 2007; Nelson, 2009; Nelson *et al.*, 2008). In this work we performed the first complete manual annotation of P450s in the nearly homozygous *V. vinifera* reference genome PN40024 (Jaillon *et al.*, 2007). We discuss the structural organization of the genes with particular focus on gene clusters. We evaluate phylogenetic relationships between those genes to identify recently expanded gene families likely linked to adaptive traits or domestication. Finally, we investigate spatio-temporal gene expression patterns, with particular focus on berry development and pathogen response to identify P450s with potential roles in these important physiological processes. This work will support further functional characterization of cytochrome P450 genes in grapevine.

# Results

## Gene annotation, classification and phylogeny

A similarity search of the *V. vinifera* PN40024 genome with known P450 sequences revealed 579 putative P450 sequences. We manually curated the sequences obtained with a gene prediction algorithm, and validated the annotation with grapevine unigenes and RNAseq reads (see Material and methods). We distributed them into four categories: genes, partial genes, putative pseudogenes and pseudogenes. This led to the identification of 279 full-length genes, which is fewer than 315 genes reported for the heterozygous Pinot Noir genome on the Cytochrome P450 homepage (http://drnelson.uthsc.edu/CytochromeP450.html), and suggests that some sequences previously annotated as different genes are probably allelic variants. The number of cytochromes P450 in grapevine is comparable to their number in other plants (e.g. 273 in *Arabidopsis thaliana* and *Solanum lycopersicum*, 309 in *Oryza sativa*). Twenty sequences were annotated as partial genes, lacking a segment of the sequence due to gaps in the genome assembly. Eleven putative pseudogenes only contain one nonsense mutation or frame shift, which could originate from sequencing errors or be genuine and still be functional genes in some varieties. Finally, the 269 pseudogenes are fragments, either containing multiple stop codons or frameshift mutations, or sequences not aligning to the whole length of homologous P450 genes.

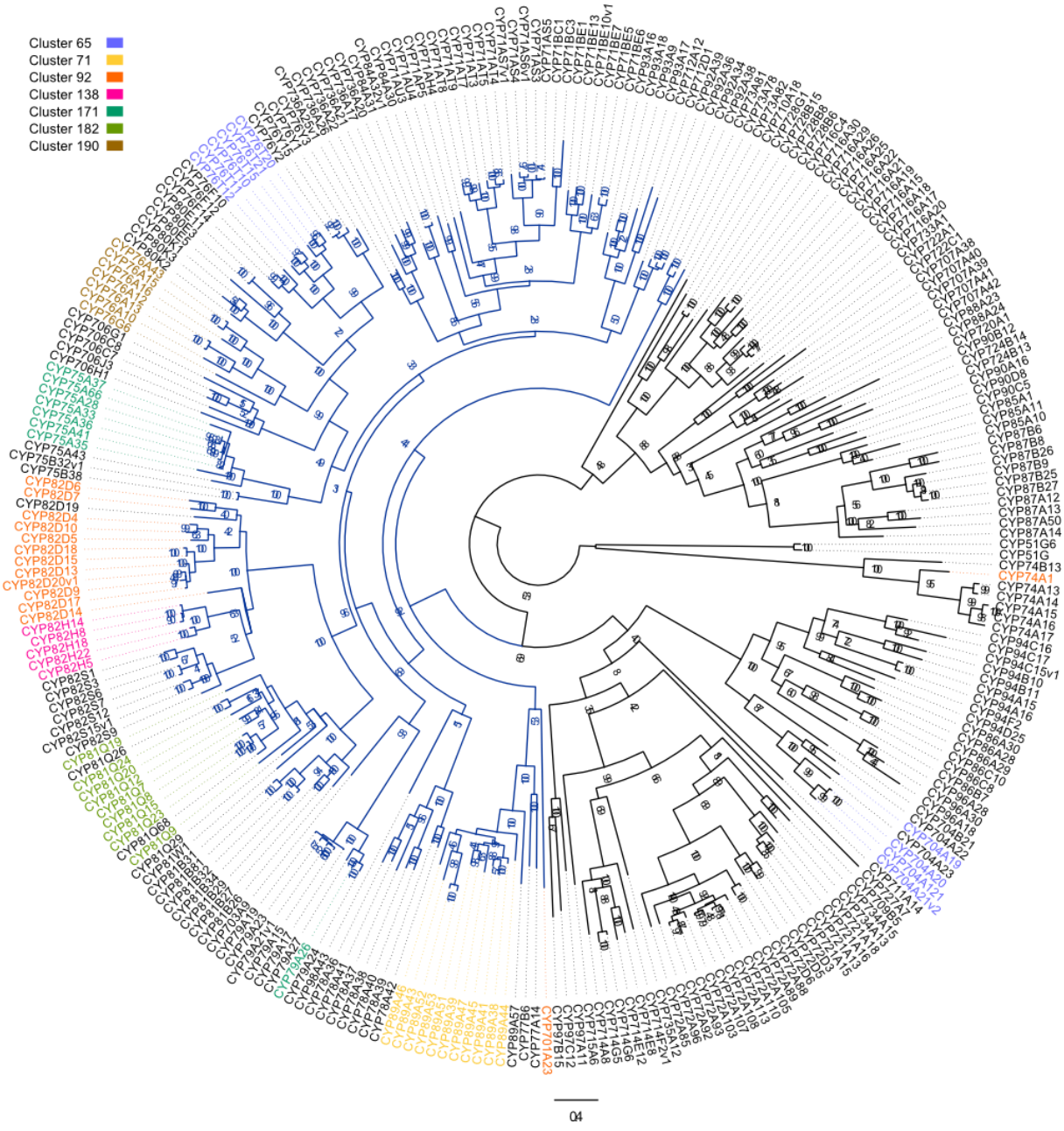



**Figure 1: Molecular phylogenetic analysis of grapevine cytochrome P450.** The alignment of full length cytochrome P450 protein sequences was used to generate a maximum likelihood tree. The dark blue clade is the clan 71, which often contains genes involved in specialized metabolism. The highlighted genes belong to the seven largest physical clusters.

The cytochrome P450 families usually have conserved or similar functions across species. Grapevine P450s can be assigned to 48 families based on sequence identities. A phylogenetic analysis of either the full-length sequences or of a subset of conserved P450 sites confirmed this classification for most of the families. One exception is the CYP90B gene, which is clustered with CYP720 and CYP724 as previously observed (Nelson *et al.*, 2004). Other exceptions are the families CYP76 and CYP80, which form a monophyletic group Figure 1, see Material and methods).

We thus investigated the phylogeny of these two families in the broader context of selected angiosperm species Figure S1). CYP80 clearly groups with CYP76 sequences, but the phylogenetical relationships of the clades CYP80, CYP76A/G and the rest of CYP76 sequences (labeled core CYP76) remain uncertain. Within the CYP76A/G clade, a eudicot duplication gave rise to the two subfamilies CYP76A and CYP76G. Within the large "core CYP76" clade the uncertain position of both the monocot and *Amborella trichopoda* CYP76s could be due to a problem of long-branch attraction. A specific core eudicot duplication gave rise to CYP76F/B/X on one side and CYP76T/C/E on other side. These tree topologies were obtained both with the full-length alignment and the partial alignment of conserved sites. Although species-specific "blooms" appeared in the whole CYP76/80 family, they are particularly abundant in the "core CYP76" clade. Different subfamilies "expanded" in different species.

Comparison of P450 family sizes between species (Figure S2) allowed us to identify families that potentially expanded in grapevine and might have a role in the production of species-specific specialized metabolites: an expansion of the CYP75 family, involved in anthocyanin biosynthesis, is already well documented (Falginella *et al.*, 2010), whereas the function of CYP82, the largest P450 family in grapevine with 25 members, is currently unknown in this species. Other families that are larger in grapevine than in most other species are: CYP76, CYP79, CYP80, CYP81, CYP87, CYP89 and CYP716.

**Structural organization of the P450s in the PN40024 genome**

The 579 cytochrome P450 sequences are distributed on all the 19 chromosomes. Some chromosomes, namely 18, 19 and 6, carry a high number of P450s, whereas others, for example chromosome 5, carry very few (Figure S3). 24 P450 sequences (7 genes, 6 partial genes, 11 pseudogenes) are located on the "Unknown" chromosome, which is composed of scaffolds that could not be anchored on any of the 19 chromosomes. Since the genome is not completely homozygous (estimated homozygosity is 93% (Jaillon *et al.*, 2007)), the "Unknown" chromosome may also contain eventual allelic variants of genes that are placed on the 19 chromosomes.

We further investigated the distribution of cytochrome P450 sequences in clusters or groups in close physical proximity (separated by less than 200 kb and 8 non-P450 genes (Richly *et al.*, 2002; Yang *et al.*, 2008)). Our results show that P450 sequences are organized in clusters and not randomly distributed in the genome (bootstrap test, *p*-value < 0.0001). A large majority of cytochrome P450 sequences (452 or 78%) are part of one of the 85 clusters and only 22% (127 P450 sequences) are isolated in the grape genome. The largest number of clusters (40%) are only composed of 2 P450 sequences, whereas the largest cluster counts 35 P450 sequences. On average, there are 5 P450s per cluster and the median is 3 P450s per cluster (Figure S4). The clusters are not enriched neither in complete genes nor pseudogenes, compared to isolated annotations (data not shown). Some chromosomes, such as 16 and 18, are enriched in clustered P450s, whereas others, such as chromosomes 4 and 11, are enriched in isolated P450 (Figure 2, Figure S3).

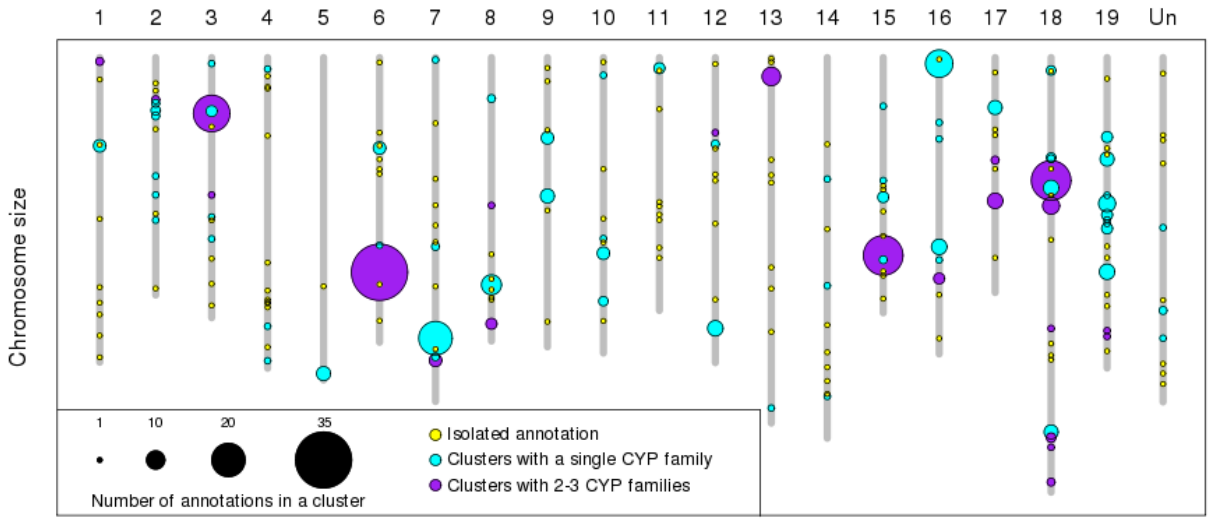



**Figure 2: Physical map of cytochrome P450 sequences on the 19 *V. vinifera* chromosomes.** Yellow circles represent isolated annotations, light blue circles represent physical clusters composed of members of only one P450 family and the purple circles represent physical clusters composed of members of 2–3 P450 families. The circle size is proportional to the number of sequences in the cluster. The numbers 1–19 are chromosome numbers and "Un" is "Unknown chromosome" which contains sequences with unknown chromosome location.

Cytochrome P450 families group genes with higher sequence similarity (≥40% protein sequence identity) and often a similar function. A majority of physical clusters are composed of members of only one P450 family (63 clusters, 74%) and the remaining clusters are composed of up to 3 P450 families. The 4 largest clusters are composed of several P450 families, whereas the clusters with single P450 families are smaller (Figure 2, Figure S4). Most of the largest P450 families (CYP82, CYP71, CYP81, CYP76, CYP72 and others) are organized in clusters (Table S1).

Clustering by P450 family already indicates that more similar P450 sequences cluster in closer physical proximity. But many P450 families are dispersed among several clusters. We thus wished to explore whether the closest paralogs belong to the same or different (Figure 3). The majority of clustered P450 genes (86%) have their closest paralog (the best BLAST hit) in the same cluster. The second and third closest paralogs (second and third best BLAST hit) are in the same cluster for 58% and 49% of the clustered P450 genes. The sequence similarities within the same cluster are thus higher than between clusters.

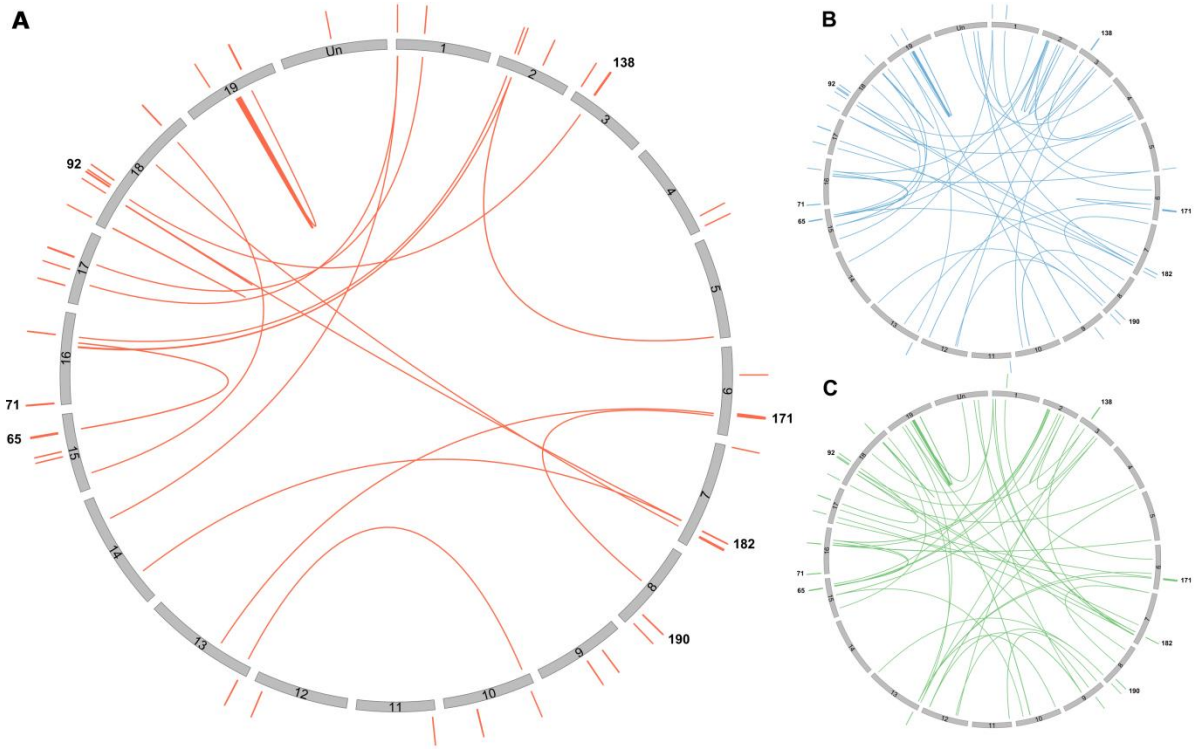



**Figure 3: Similarity of the P450 genes between and within clusters.** For each circle, the grey bars correspond to the 19 grape chromosomes and the "Unknown chromosome". The lines connect complete P450 genes according to their similarity. The lines outside the circles show the similarity between genes of the same cluster, whereas the lines in the circle connect similar genes of different clusters. Only P450 genes that from clusters composed of at least two complete genes are illustrated here. The seven largest clusters are labeled with numbers corresponding to Table 1. The lines are connecting the genes corresponding to the best BLAST hit (A), second best hit (B) or third best blast hit (C).

**Large P450 clusters in *V. vinifera* genome formed *via* different mechanisms**

To investigate the mechanisms underlying the formation of large physical clusters of cytochrome P450 genes, we further analyzed the sequence similarity within clusters, taking into account not only the coding P450 sequences, but also the surrounding non-coding-sequences. This allowed us to infer the mechanism of cluster formation. We focused on the 7 largest physical clusters, which comprise of 11 to 35 P450 sequences (Table 1). Together, these 7 clusters contain 23% of all P450 genes, and a similar fraction of total P450 sequences. Most of the sequences in these clusters are part of "clan 71", which is a large clade of plant cytochromes P450 often involved in the biosynthesis of specialized metabolites (Figure 1).

Analysis of similarity blocks within these clusters showed they differ remarkably in their structures (Figure S5). One of the largest physical clusters, cluster 65, is characterized by low similarities, both among the P450 sequences and surrounding non-coding regions. The similarity blocks of two other large physical clusters, **71** and **171**, are restricted to P450 sequences and do not extend to the intergenic regions. Single gene duplications were thus probably the main mechanism of formation of these two clusters. The similarity blocks of physical clusters **138** and **182** extend to the non-coding regions around the cytochromes P450 annotations. This suggests the duplication events leading to formation of these clusters happened relatively recently. High similarity between the non-coding regions, which include the promoter regions, should result in similar expression profiles. Cluster **138** has the highest fraction (73%) of pseudogenes of all the seven large clusters. In physical clusters **92** and **190**, the similarity blocks extend over even longer regions that include 3–4 cytochrome P450 sequences and their intergenic regions (Figure 4). In addition, the type of annotation (gene or pseudogene) was also maintained in the same order between duplicated blocks. This suggests these two clusters formed through very recent segmental duplications.

**Table 1: Description of the seven largest physical P450 clusters in the *V. vinifera* genome.** Label – sequential number of each cluster in the genome; Chr – chromosome number; Location – chromosome coordinates; Total seq. – number of P450 sequences in each cluster, including complete and partial genes, putative pseudogenes and pseudogenes with their family distribution; Expressed seq. – number of expressed P450 sequences in the cluster; Complete genes – number of complete P450 genes in the cluster; co-expression – expression pattern of the cluster ("-" signifies low co-expression within the cluster); Organization – description of structural organization and mechanism of formation of each cluster.

| Label | Chr | Location | Total seq. | Expressed seq. | Complete genes | Co-expression | Organization |
|---|---|---|---|---|---|---|---|
| 65 | 15 | 15572751.. 15909327 | 20 CYP76 4 CYP704 | 20 | 10 | Flowers | Low similarity among members |
| 71 | 16 | 401789.. 596606 | 16 CYP89 | 14 | 11 | All leaves | Single gene duplications |
| 92 | 18 | 9625486.. 9912876 | 22 CYP82 1 CYP74 1 CYP704 | 21 | 14 | Leaves and ripe berries | Duplicated blocks with co-expression; some single gene duplications |
| 138 | 3 | 4387722.. 4512089 | 22 CYP82 | 18 | 5 | Young berries | Small duplicated blocks, a few are co-expressed, single gene duplications |
| 171 | 6 | 16790972.. 17446396 | 21 CYP75 14 CYP79 | 29 | 8 | - | Single gene duplications |
| 182 | 7 | 22260680.. 22372250 | 20 CYP81 | 17 | 9 | Berries | Small duplicated blocks, a few are co-expressed, single gene duplications |
| 190 | 8 | 18038159.. 18121816 | 11 CYP76 | 10 | 7 | Flowers | Duplicated blocks with co-expression; some single gene duplication |

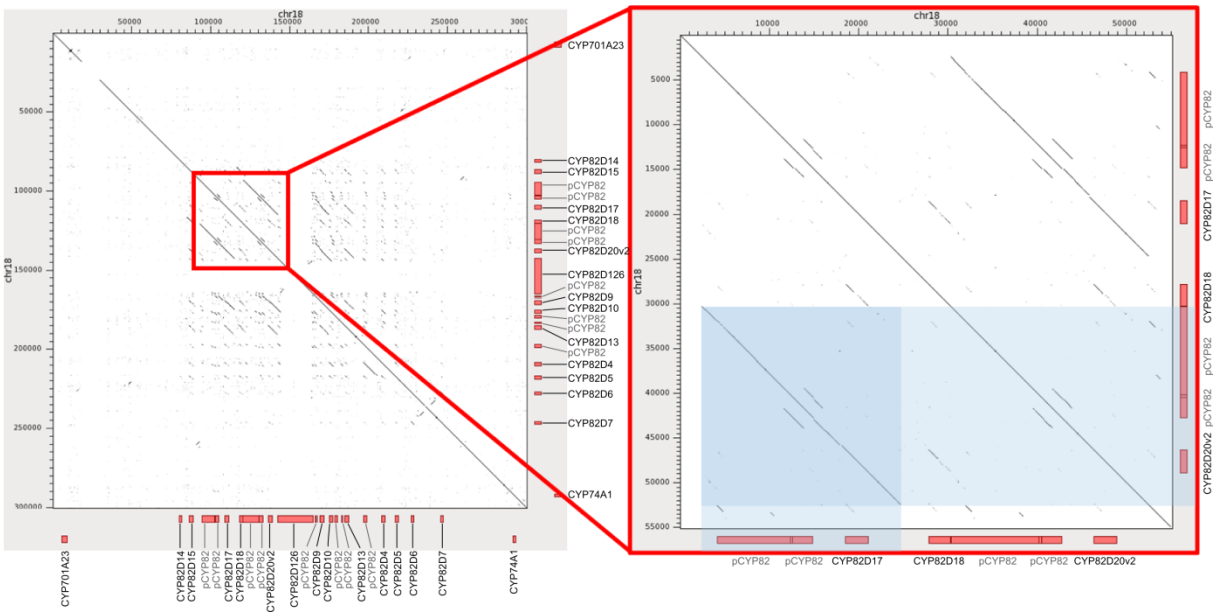



**Figure 4: Dot matrix of segmental duplications in the physical cluster 92.** Physical cluster 92 is located on chromosome 18 and comprises 22 CYP82 sequences, one CYP74 sequence and one CYP704 sequence. The dots and the black lines represent the sequence similarities in cluster **92** compared to itself. The red rectangles on the sides of the graph represent cytochrome P450 sequences. Complete genes are labeled with their name and pseudogenes are labeled with "p" and the P450 family. A) The similarities for the whole cluster **92**. B) A zoom of the red squared region which contains two 20-kbp-sequence blocks with very high similarity. Analysis of gene expression showed that CYP82D17 and CYP82D20v2 are co-expressed (expression cluster **O**, expression in leaves), and so are the first and the third pseudogene in the enlarged segment (expression cluster **J**, expression in ripe berries).

## Expression profiles of grapevine P450s

To identify P450 genes with potential roles in pathogen resistance or biosynthesis of berry metabolites we analyzed the expression of the 579 P450 sequences. Pseudogenes were included in the analysis of expression to account for sequences that might be functional in other varieties, as well as for recently pseudogenized sequences that may still be expressed to some extent. We used 73 RNA-Seq datasets (Table S2), which describe gene expression in different tissues (flowers, berries, leaves), different stages of berry development, and pathogen infection. We grouped the 73 experiments in 5 categories: flowers, green berries, ripe berries, leaves (control) and leaves under biotic stress. The latter category includes leaves infected with the powdery mildew pathogen *Erysiphe necator* and the downy mildew pathogen *Plasmopara viticola*.

To enable a meaningful comparison of gene expression between different experiments we calculated fragments per kilobase of transcript per million mapped reads (FPKM) for each P450 sequence in the 5 categories of experiments. We grouped the expression levels into 4 classes (no expression, low, average or high expression). The majority of P450 sequences (494 or 85%) were expressed in at least one experiment.

Of the remaining 85 non-expressed P450 sequences, only 4 were complete genes. Expression of complete P450 genes (mean FPKM = 11, median FPKM = 0.6) was higher compared to pseudogenes and putative pseudogenes (mean FPKM = 1.6, median FPKM = 0) or partial genes (mean FPKM = 1.7, median = 0.1). In each of the 5 categories, an average of 12% of the genes were not expressed, 54% had low, 20% average and 14% high expression. Interestingly, in leaves exposed to biotic stress, the fraction of non-expressed genes drops from 28% to only 9%, whereas the fraction of the highly expressed genes increases from 15 to 21%. This indicates a major shift in expression caused by biotic stress. Mean expression per category for all P450 sequencing is available in the **Appendix I**.

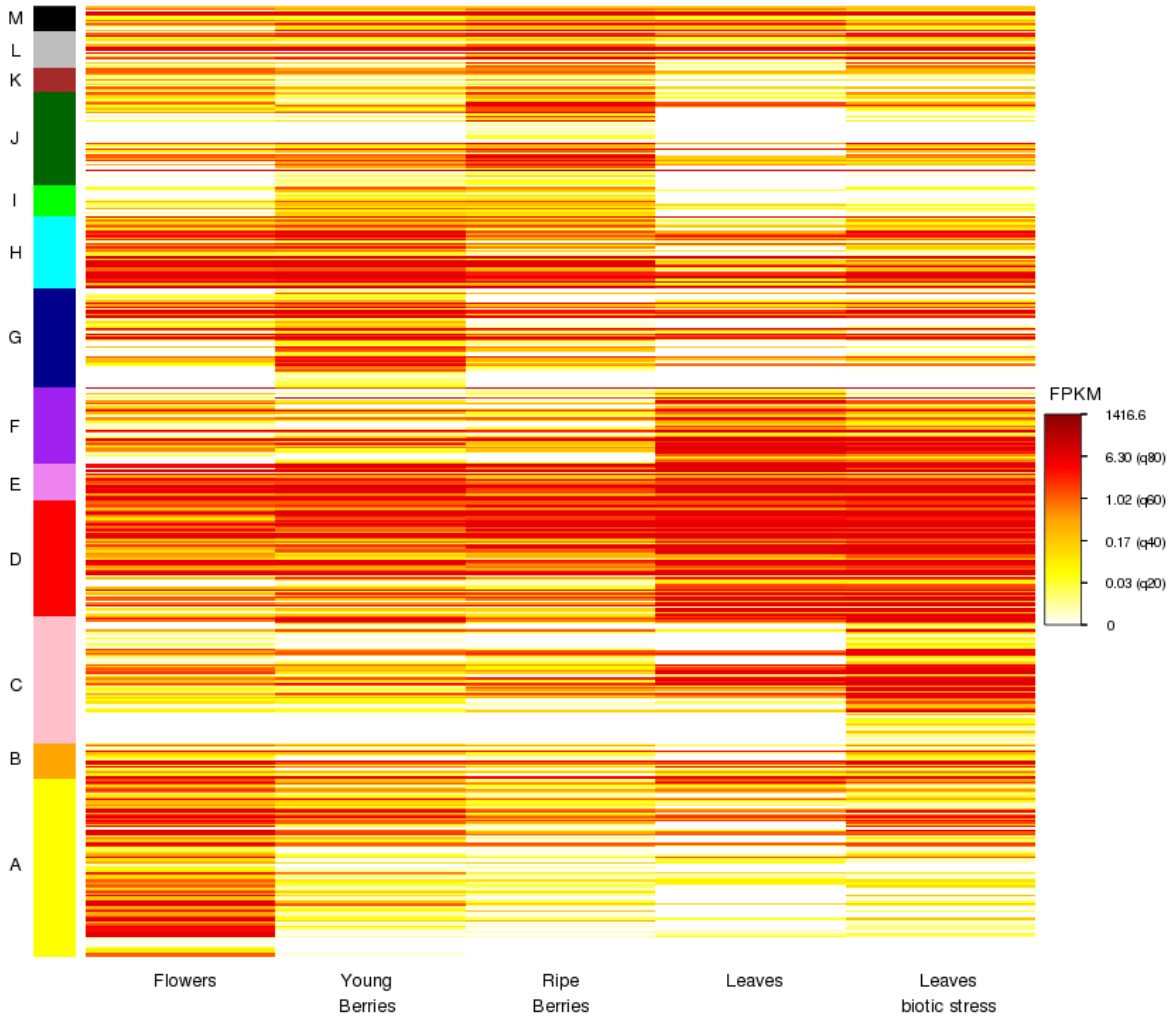



**Figure 5: Heat map of the P450 sequences, clustered according to their expression profile.** The expression levels were averaged over the experiments classified in one of the five experimental categories: flowers, young berries, ripe berries, leaves and leaves under biotic stress. This heatmap includes the 494 expressed cytochrome P450 sequences. The color bars on the left are showing the 13 expression clusters, which are designated by the letters on the side. The 20, 40, 60 and 80% quantiles were used to design the FPKM color scale so that the numbers of expression values in each interval are similar.

We analyzed the expression patterns by clustering the expression profiles of the 494 expressed cytochrome P450 sequences. A Pearsons' correlation coefficient cut-off of 0.795 resulted in 13 expression clusters, shown in Figure 5. Five expression clusters (**A**, **C**, **F**, **G** and **J**) were up-regulated in one of the five experimental categories. These five clusters contained 297 P450s, showing that 60% of the expressed P450s were specifically up-regulated rather than constitutively expressed in the investigated conditions. The largest expression cluster was **A**, which consisted of 92 sequences with preferential expression in flowers. Other large expression clusters were cluster **C** with 66 sequences preferentially expressed in leaves subjected to biotic stress, cluster **D** with 60 sequences expressed in both leaves conditions and cluster **G** with 51 sequences preferentially expressed in young berries. The clusters **B**, **E** and **L** consisted of 57 sequences with constitutive expression.

The analysis of gene expression highlighted some of the P450s with potential role in biotic stress response. From the 166 sequences preferentially expressed in leaves (clusters **C**, **D** and **F**), 76 were upregulated (more than 2 fold) and 20 were downregulated (more than 2 fold) upon biotic stress, confirming an expression shift upon pathogen infection. The P450 family with the highest number of upregulated genes was CYP82 (11 sequences, 6 complete genes), followed by CYP71 and CYP81 with 10 and 8 sequences, respectively (4 and 3 complete genes, respectively). The CYP87 family is also well represented among the upregulated genes with 7 sequences (4 complete genes) out of which 3 complete genes are in the 10 most expressed genes upon pathogen infection.

Another major shift in the P450 expression pattern occurs during the grape berry ripening. From 115 sequences preferentially expressed in berries (clusters **G**, **J** and **I**), 52 were upregulated in green berries (cluster **G**) and 48 in ripe berries (cluster **J**). Interestingly, the most represented families cluster **J** were also CYP82 with 10 sequences (3 complete genes) and CYP71 with 5 sequences (3 complete genes). This cluster in addition featured the P450 gene with the highest expression in all experiments: CYP78A41 with a FPKM value of 1417. In cluster **G**, the most represented family is CYP716 with 7 sequences (5 complete genes).

We more thoroughly investigated co-expression of cytochrome P450 sequences within the 7 largest physical clusters (Table 1). In the previous section, we identified 4 large clusters with high similarity, not only among the coding, but also the non-coding regions. These non-coding regions presumably include promoter sequences, so the genes in these clusters are expected to have the same expression pattern. Indeed, the 20 kb duplicated block within cluster **92** (Figure 4b) retained the same expression profile after the segmental duplication. The duplicated segment consists of three P450 sequences: two pseudogenes and one gene. The first pseudogene in both blocks retained a very low level of expression in ripe berries, the second pseudogene in both blocks was not expressed at any experimental conditions, whereas the two complete genes (CYP82D17 and CYP82D20v2) were co-expressed in leaves and slightly induced upon pathogen infection. Eight out of 24 sequences in this cluster shared this same expression pattern (expression clusters **C** and **D**), whereas 6 other P450 sequences in the same cluster were up-regulated in ripe berries (cluster **J**). Interestingly, cluster **138** is also composed of CYP82 sequences, but these sequences were preferentially expressed in young berries. Another large physical cluster with coordinated expression pattern was cluster **71**, which comprises 11 CYP89 genes and 5 pseudogenes. Five of these genes were specifically expressed in leaves (cluster **D**). Eight CYP76 sequences in the physical cluster **190**, on the other hand, were co-expressed in flowers.

## Discussion

We produced a reliable and validated manual annotation of cytochromes P450 in the genome of the nearly homozygous grapevine (*V. vinifera)* accession PN40024 (Jaillon *et al.*, 2007). Cytochrome P450 superfamily in *Vitis vinifera* contains both very similar and very divergent genes (sequence identity ranges from 10% to almost 100%), and often form clusters in very close physical proximity, which makes it challenging for automated annotation algorithms. Manual curation is therefore necessary to produce a reliable annotation, suitable for demanding downstream applications such as phylogenetic or gene expression analysis. Grapevine P450s have been previously manually annotated (Cytochrome P450 homepage, http://drnelson.uthsc.edu/vitis.htm) in the highly heterozygous genome of Pinot noir cultivar (Velasco *et al.*, 2007). Our annotation represents an improvement over the existing dataset for several reasons. The assembly of PN40024 genome is of better quality compared to the Pinot noir genome: it contains fewer gaps and a higher fraction of anchored contigs. The homozygosity of the genome not only enabled a better quality of the assembly, but also assured that most of the annotated sequences are individual loci and not allelic variants. This can partially explain a lower number of cytochrome P450 genes in our annotation—279—compared to the 315 genes reported on the Cytochrome P450 homepage. Additionally, the annotation on the Cytochrome P450 homepage classifies the sequences in only two categories, genes and pseudogenes, whereas we employed a more stringent classification into genes, partial genes, putative pseudogenes and pseudogenes. Lastly, we report the exact genomic coordinates of the P450 sequences, which facilitate comparison to annotations of other genes, and provide insights into structural organization of the grapevine CYPome.

Several gene families involved in the biosynthesis of specialized metabolites, such as terpene synthase genes (TPS) and stilbene synthase genes (STS), have expanded in grapevine genome compared to other species (Martin *et al.*, 2010; Parage *et al.*, 2012; Jaillon *et al.*, 2007). Although the total number of cytochrome P450 genes in grapevine is comparable to other species, individual P450 families experienced similar expansions. These expended families, similarly to TPS and STS families, form large physical clusters of more than 10 homologous sequences. One of such families is CYP75, which together with CYP79 family members forms the largest physical cluster of 35 P450 sequences on chromosome 6. Expansion of CYP75 genes in grapevine was previously documented, but the presence of another P450 family, CYP79, in the same cluster was not reported (Falginella *et al.*, 2010).

Clustered genes with low or no homology sometimes participate in the same biosynthetic pathway (Takos and Rook, 2012; Nützmann and Osbourn, 2014), but this is unlikely in the case of CYP75 and CYP79, since both families have well established roles in different biosynthetic pathways: CYP79 genes code for amino acid *N*-hydroxylases (Hamberger and Bak, 2013), whereas CYP75A genes code for flavonoid 3',5'-hydroxylases (Ayabe and Akashi, 2006), crucial enzymes in the biosynthesis of blue anthocyanins in the grape skin (Falginella *et al.*, 2012; Falginella *et al.*, 2010). We can, however, not exclude the recruitment of some of these genes in other pathways. Interestingly, the sequencing of the genome of the grapevine cultivar Tannat, characterized by its very deep color, revealed an even higher number of CYP75 genes compared to the PN40024 accession (Da Silva *et al.*, 2013). Copy number of genes in a cluster can therefore vary between cultivars and influence varietal characteristics. Other expanded P450 families in the grapevine genome that form large clusters are CYP82, CYP76, CYP81 and CYP89.

Analysis of gene expression across several tissues and conditions provides a first hint to the putative P450 functions in grapevine. Pathogen infection causes a major shift in the P450 expression, inducing members from families CYP71, CYP81, CYP82 and CYP87. Their homologs in other species have been shown to participate in biosynthesis of highly specialized defense compounds (Table S1). Interestingly, CYP736A25v1, which was shown to be upregulated upon infection with the Pierce disease pathogen *Xylella fastidiosa* (Cheng *et al.*, 2010), is also upregulated upon infection with powdery mildew and downy mildew pathogens. Two other sequences from the CYP736 family are also induced by biotic stress but their expression level is lower. Another large shift in expression occurs in developing grape berries. The most upregulated P450 families in the ripe-berry expression cluster are CYP82 and CYP71 (the two largest P450 families in grapevine). These P450s are likely to participate in the biosynthesis of defense compounds or compounds important for the organoleptic properties of wine (aroma, colour, taste, mouthfeel). The most up-regulated P450 gene in ripe berries and the P450 gene with the overall highest expressions is CYP78A41. A member of the same P450 family in tomato (*S. lycopersicum*) was selected during domestication to increase fruit size (Chakrabarti *et al.*, 2013). High expression of CYP78A41 in grape berries points to a similar event in grapevine domestication. However, other P450 families not only upregulated in ripe berries could play an important role in aroma biosynthesis as for example CYP76 family for which CYP76F14 was identified as playing a major role in the production of wine lactone (Ilc *et al.*, 2016).

The phylogenetic and structural data suggest that some P450 families underwent multiple tandem or segmental duplications, which resulted in large physical clusters of homologous sequences. Most of these P450 families are involved in biosynthesis of highly specialized metabolites in other plant species. These genes are often expressed in specific conditions and tissues, such as leaves upon pathogen infection. Our work thus lays the ground for discovery of interesting novel P450 functions in grapevine.

# Material and methods

## Gene annotation

We annotated the cytochromes P450 using the 12x version of the assembly of the *Vitis vinifera* cv PN40024 genome (Grimplet *et al.*, 2012; Jaillon *et al.*, 2007). Four publically available datasets of cytochromes P450 were used to perform similarity searches in the PN40024 genome. 947 protein sequences of grape P450s were downloaded from the NCBI Protein database (http://www.ncbi.nlm.nih.gov/protein, Feb 2014). Three datasets were downloaded from David Nelson's website (http://drnelson.uthsc.edu/CytochromeP450.html, Feb 2014), which stores manually curated annotations of cytochromes P450 for many species: 702 P450 protein sequences of *Vitis vinifera* cv Pinot Noir clone ENTAV115 (28, http://drnelson.uthsc.edu/vitis.htm); 416 P450 protein sequences of *Vitis vinifera* cv PN40024 from the 8x assembly version of the genome (10, http://drnelson.uthsc.edu/Vitis.additionalP450s.htm); and 288 P450 protein sequences of *Arabidopsis thaliana* (35, http://drnelson.uthsc.edu/Arabidopsis.Blast.file.html). The four datasets were masked for repeat sequences using the online tool "Repeat Masking" from Censor (http://www.girinst.org/censor/index.php).

The four masked datasets were used to perform four independent TBLASTN analyses (Altschul *et al.*, 1997) against the PN40024 12x sequence with an e-value cutoff of 1e-3. The TBLASTN outputs were parsed using a homemade script. The hits from the three grape datasets were kept if they were at least 50 amino acids long with at least 70% sequence identity. The hits from the *Arabidopsis* dataset were kept if they were at least 50 amino acids long with an identity percentage of at least 50%. The software Exonerate (version 2.2.0, build October 2008, 37) was used to predict gene structures using the protein2genome parameter and the same cutoff of sequence identity as above. A homemade script was used to reformat the output files from exonerate into files in the gff format. These gff files were imported to the Artemis genome browser (Rutherford *et al.*, 2000) to perform the manual curation of the structures suggested by Exonerate. The parsed hits identified through TBLASTN were used to improve or to complete the Exonerate annotations. Every annotation starting with a start codon, ending with a stop codon and with correct exon-intron borders (GT-AG or sometimes GC-AG) was considered as a complete "gene".

Every annotation showing the previously described gene structure but with a single point mutation creating a frameshift, a premature stop codon or a wrong exon-intron border was considered as a "putative pseudogene" also marked "pseudogene?" because it may result from a mistake in the genome assembly. Every annotation interrupted by a gap in the genomic sequence or including one was considered as a "partial" annotation. All the other annotations with wrong gene structure but showing a significant similarity level with a cytochrome P450 from one of the four datasets were annotated as "pseudogenes". The genome annotation V1 stored in Grape Genome Database hosted at CRIBI (39; http://genomes.cribi.unipd.it/DATA/GFF/V1.phase.gff3) and a set of expertized and functional grape cytochromes P450 were used to guide the manual curation.

To validate the gene structure, two transcript datasets were used. First, the *Vitis vinifera* unigene set build #15 from the NCBI database was downloaded (ftp://ftp.ncbi.nih.gov/repository/UniGene/Vitis_vinifera/Vvi.seq.uniq.gz). The 32,193 unigenes were mapped on the PN40024 12x sequence using GMAP version 2013-11-27 (Wu and Watanabe, 2005) using the default parameters except for the format parameter which was set to "gff3_match_cdna". The second transcript dataset was locally assembled using six RNA-Seq experiments ((Perazzolli *et al.*, 2012), SRR519450, SRR519456, SRR520380 and SRR520385; (Sweetman *et al.*, 2012), all four samples; (Vannozzi *et al.*, 2012), SRR493740-SRR493746; (Da Silva *et al.*, 2013), SRR866544, SRR866570, SRR866571 and SRR866576; (Venturini *et al.*, 2013), SRR522472, SRR522477 and SRR522478; and eight unpublished RNA-Seq datasets acquired by INRA. The software Tophat2 v2.0.11 (Kim *et al.*, 2013) was used to map the RNA-Seq reads against the PN40024 12x sequence using the following parameters: -p 5 -N 5 --read-edit-dist 5. The software Cufflinks v2.2.1 (Trapnell *et al.*, 2010) was used to assemble the transcripts from all the RNA-Seq experiments. First the cufflinks command was used with the -p 5 parameter and then the cuffmerge command with the -p 15 parameter and using the fasta file of the PN40024 12x sequence for the -s parameter. This assembly led to 32,219 transcripts and to a gtf file showing their mapped location in the PN40024 12x sequence. The two transcript datasets were formatted in gff format compatible with the Artemis Browser so that the predicted gene structures of the cytochromes P450 could be compared with the transcripts and edited if needed.

The command maskFastaFromBed v2.19.1 from the bedtools package (Quinlan and Hall, 2010) was used to mask the regions of the PN40024 12x sequence where we annotated cytochrome P450 exons after having reformatted the gff file of the annotations into a bed file. We performed TBLASTN analyses of the four grape cytochrome P450 datasets against the masked PN40024 12x sequence and parsing analyses using the same parameters and cutoffs than previously described. This step allowed to identify the region of the grape genome for which a cytochrome P450 similarity was missed during the manual curation.

To validate the set of complete genes of cytochromes P450 that we annotated, a BLAST against non-redundant sequence database (NR) was performed and only the genes for which the best hit was a cytochrome P450 were kept. For the pseudogenes, a BLASTX was performed against the set of complete P450 genes that we annotated and we kept only the ones that aligned over at least 30% of the query length with the percentage identity of 50%.

The cytochrome P450 annotations were transferred to the improved version of the PN40024 12x assembly when it was released (PN40024 12X.2; https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences) using a homemade script. The presence of physical clusters of cytochrome P450s in the grape genome was tested based on the following definition of a cluster. Two consecutive P450 annotations are part of a cluster if they are separated by 200kb and 8 non-P450 genes at the most (Richly *et al.*, 2002; Yang *et al.*, 2008). The two annotations also have to be located on the same scaffold which guaranties a precise estimation of the intergenic distances. A bootstrap test was performed to check whether the cytochromes P450 were more clustered than what is randomly expected. A homemade script was developed with R version 3.0.2 (R Development Core Team, 2011). Ten thousand sampling without replacement of 579 (number of P450 annotations) or 279 features (number of complete P450 genes) were performed on the genome annotation V1 stored in Grape Genome Database hosted at CRIBI counting 29,971 features. The percentage of features organized in clusters was computed using the same protocol as for cytochromes P450. The p-value was calculated by counting each time a percentage equal of greater than the percentage of P450 in clusters divided by 10000 (number of iterations).

Sequence similarity within and between clusters (Figure 3) was analyzed by performing a BLASTP search of translated complete P450 genes against themselves. Only the genes that aligned over at least 70% of the query length with the percentage identity of 40% were kept. The Circos software (Krzywinski *et al.*, 2009) was used to draw the figure. Clusters that contained less than two complete genes were excluded from this analysis (i.e. clusters that contained partial genes, pseudogenes and putative pseudogenes with less than 2 complete genes).

The dotter software version 4.23 (Sonnhammer and Durbin, 1995) was used to draw the sequence similarity graphs of the cluster 190 with its fasta sequence and annotations in a gff format as an input.

**Sequence classification**

Cytochrome P450 genes, partial genes and putative pseudogenes were aligned to the P450 sequences from the heterozygous Pinot Noir genome, retrieved from the cytochrome P450 homepage (http://drnelson.uthsc.edu/CytochromeP450.html). In the case of protein sequence identity above 95%, the original name was kept. New sequences were assigned a family based on the best hit among already named grapevine P450s. 22 sequences were given a new CYP name, and genes previously annotated as members of CYP81V subfamily were re-classified to CYP81Q subfamily.

**Phylogeny**

Sequences from non-Vitis species were retrieved from the cytochrome P450 homepage (http://drnelson.uthsc.edu/CytochromeP450.html). Pseudogenes and incomplete genes were excluded from the analysis. 279 *Vitis vinifera* CYP (Figure 1) and 191 CYP76, 80 and 706 protein sequences from *Aquilegia caerulea, Nelumbo nucifera, Mimulus guttatus, Solanum lycopersicum, Amborella trichopoda, Oryza sativa, Brachypodium distachyon, Arabidopsis thaliana, Medicago trunculata, Populus trichocarpa* and *Vitis vinifera* (Figure S1) were aligned with MUSCLE (Edgar, 2004) implemented in Seaview (Galtier *et al.*, 1996; Gouy *et al.*, 2010). Conserved sites were selected in the alignment using Gblocks (Castresana, 2000) using the less stringent option parameters. Maximum likelihood phylogenies were obtained from the full-length alignments and from the subset of more conserved sites alignments (all Vitis CYP: 166 sites and 11 species CYP alignment: 278 sites) using RAxML (v 8.2.4) (Stamatakis, 2014) via the CIPRES Science Gateway (Miller *et al.*, 2010) and PhyML (implemented in Seaview v 4.5.4) (Guindon *et al.*, 2010).

Bootstrap values are shown on the nodes of the Vitis all CYP phylogeny. Nodes with bootstrap values below 60 were manually suppressed from the 11 species CYP phylogeny and are shown as trifurcations (unsolved topologies). The trees were visualized and colored using Figtree (http://tree.bio.ed.ac.uk/software/figtree). The species cladogram in (Figure S1) was inferred from the APGIII system (The angiosperm phylogeny group, 2009).

**Gene expression**

We retrieved raw grape RNA-Seq data from NCBI SRA public database (http://www.ncbi.nlm.nih.gov/sra). 73 sequence files generated in the framework of 7 different experiments (Jones *et al.*, 2014; Ramos *et al.*, 2014; Perazzolli *et al.*, 2012; Da Silva *et al.*, 2013; Venturini *et al.*, 2013; Sweetman *et al.*, 2012) and eight unpublished RNA-Seq datasets acquired by INRA were used. The data were formatted in the fastq format using the fastq-dump command from the SRA Toolkit package version 2.3.4 (http://www.ncbi.nlm.nih.gov/books/NBK158900).

Alignments of these reads against the PN40024 12x sequence were then performed using GSNAP version 2013-11-27 (Wu and Nacu, 2010) with the following parameters: -B 4, -N 1, -n 3, --nofails and the quality protocol according to the experiment. These files were parsed to keep the best, unique and paired (if paired-end reads) alignments using a homemade script.

The number of fragments aligned on each annotation from the genome annotation V1 stored in Grape Genome Database hosted at CRIBI and the cytochromes P450 was counted using the command htseq-count from the HTSeq framework version 0.6.0 (Anders *et al.*, 2015) with the following parameters: -m intersection-nonempty and -s no. Using a homemade script, FPKMs (Fragments Per Kilo base of exon per Million fragments mapped) were calculated for every annotation.

Using all non-zero FPKM values, the 33th and 66th quantiles were calculated to assign the expression values to one of the four levels of expression chosen: no, low, average and high expression. The experiments were grouped into five categories regarding the conditions under which the samples were obtained. These categories are: flowers, young berries, ripe berries, leaves (control) and leaves under biotic stress. An average expression per category was then calculated for each gene and assigned to one of the four levels of expression regarding its value: no expression if the average was zero, low expression between zero and the 33% quantile, average expression between 33% and 66% quantile and high expression for averages higher than the 66% quantile.

The average expression values for each P450 annotation were used to perform a clustering analysis using HCE version 3.5 (Seo *et al.*, 2006) with a complete linkage method and a Pearson's correlation as distance measure. The cut-off to define the clusters was set at a Pearson's correlation coefficient of 0.795.

## Acknowledgements

We thank David R Nelson for assigning names to newly discovered sequences and updating the names for previously known sequences; and Adrian Arellano Davin for editing the CYP76 phylogenetic tree. We also thank Etienne Paux for his helpful comments on the study design.

## Accession Numbers

Cytochrome P450 annotation will be submitted to public databases.

## Authors' contributions

TI contributed to manual curation and phylogenetic analysis, performed an expert check of the manual curation and gene classification, and wrote the manuscript. GA contributed to manual curation, performed the analysis of gene expression and wrote the manuscript. RT performed the phylogenetic analysis. NN contributed to manual curation and wrote the manuscript. FC participated in the study design. MF contributed to manual curation. PH and DWR acquired funding, contributed to study design and supervision and edited the manuscript. CR contributed to study design, performed the analysis of structural organization, coordinated the work and wrote the manuscript.

# Supplemental information

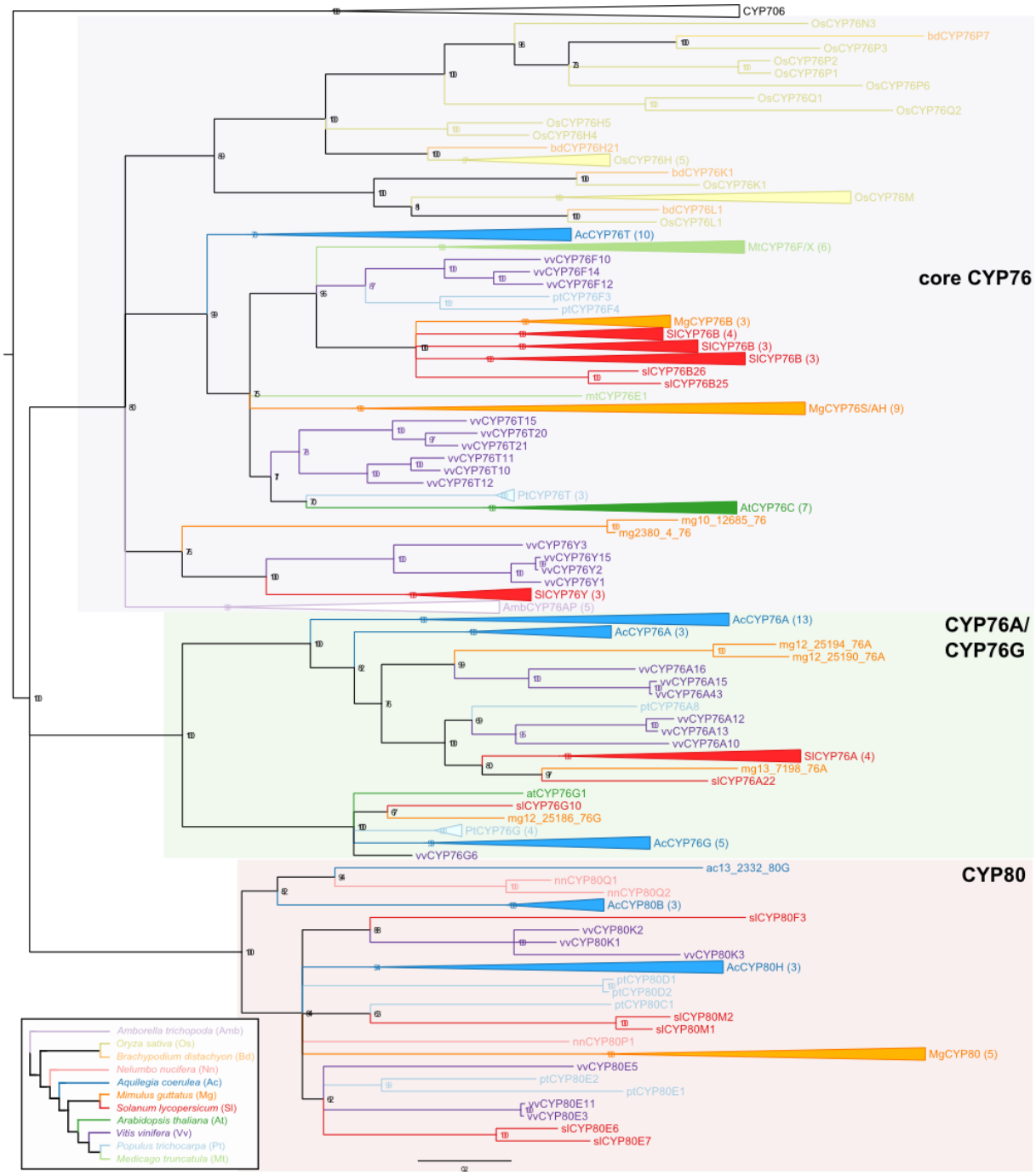



**Figure S1: Phylogeny of CYP80 and CYP76 in angiosperms.** Maximum likelihood tree of full length CYP76 and CYP80 protein sequences from a selection of angiosperms, rooted with CYP706 from all the included species. Nodes with bootstrap values below 60 are collapsed to trifurcations. Species specific clades with more than two members (except *V. vinifera*) are collapsed to triangles. The label of the triangle gives the subfamily and the number of members contained in the clade.

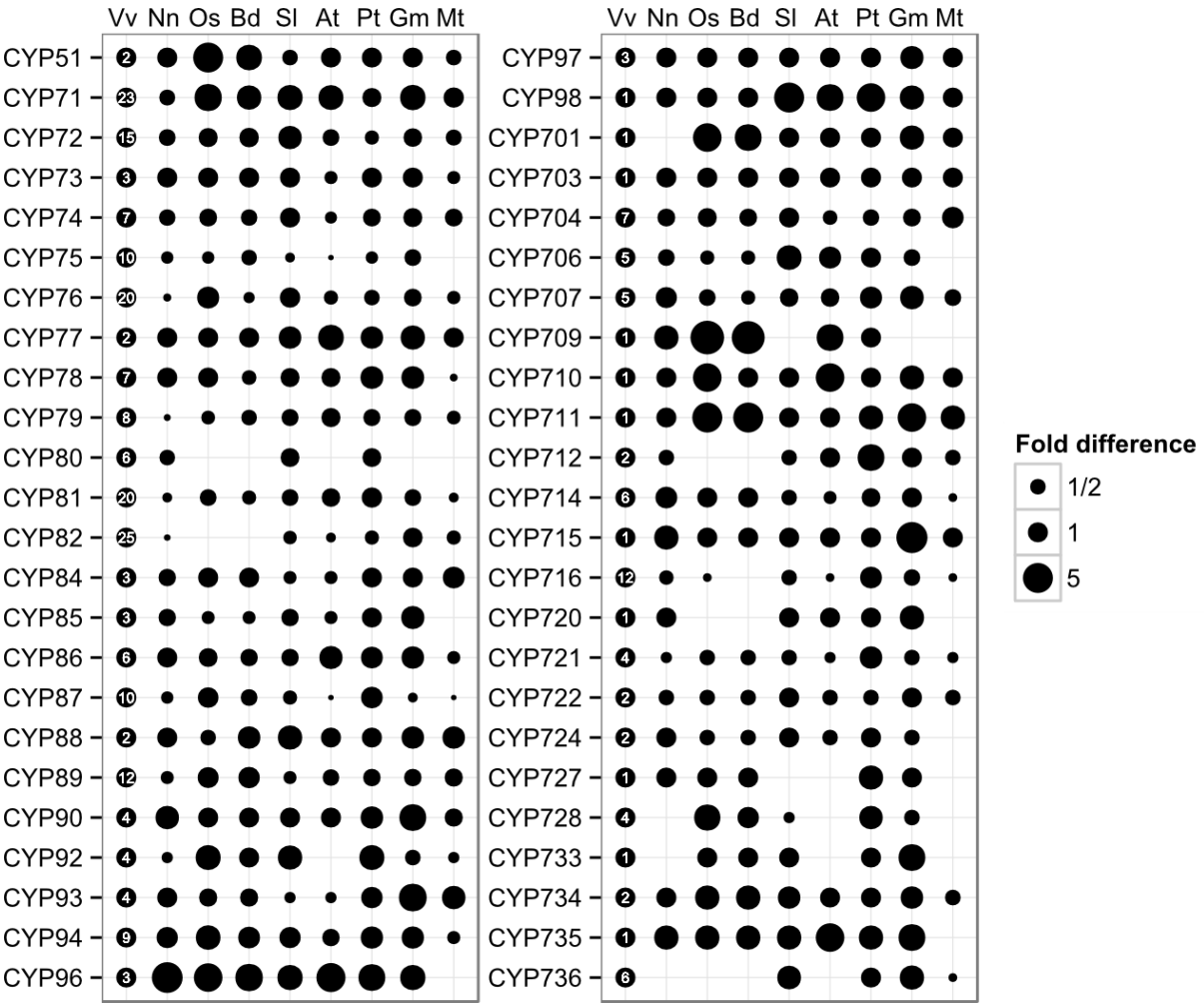



**Figure S2: Comparison of the number of P450 genes per family between species.** Dot size is proportional to the relative family size (number of genes per family) in a given species compared to *Vitis vinifera* (Vv = *Vitis vinifera*, Nn = *Nelumbo nucifera*, Os = Oryza sativa, Bd = *Brachypodium distachyon*, Sl = *Solanum lycopersicum*, At = *Arabidopsis thaliana*, Pt = *Populus trichocarpa*, Gm = *Glycine max*, Mt = *Medicago truncatula*). The numbers in the first column are the absolute family sizes (numbers of genes per family) in *Vitis vinifera*. The number of genes per family was retrieved from the cytochrome P450 homepage. Pseudogenes and families not present in *V. vinifera* (CYP83, CYP99, CYP702, CYP705, CYP708, CYP718 and CYP729) were excluded from the count.

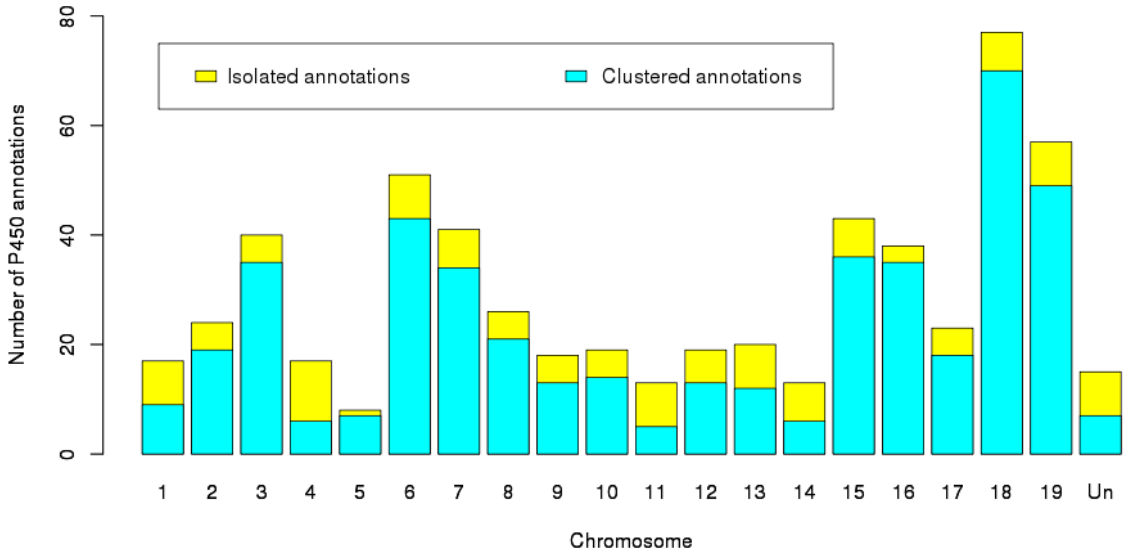



**Figure S3: Distribution of the *V. vinifera* P450s per chromosome.** The blue bar corresponds to clustered annotations and the yellow bar to the isolated annotations. The "Unknown chromosome" is labeled as "Un".

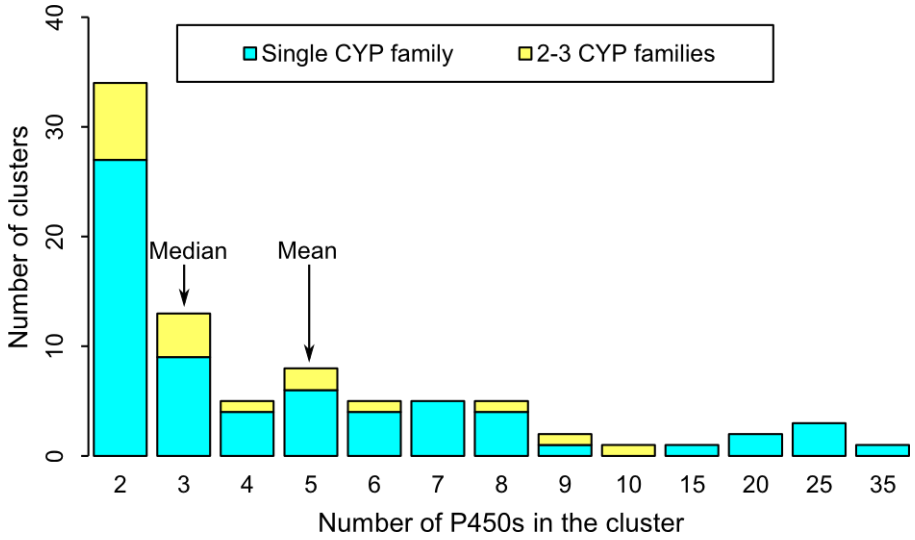



**Figure S4: Distribution of the P450 sequences per physical cluster.** Median and average values are labeled with arrows. The clusters composed of a single P450 family are represented in blue and those composed of 2 or 3 P450 families in orange.

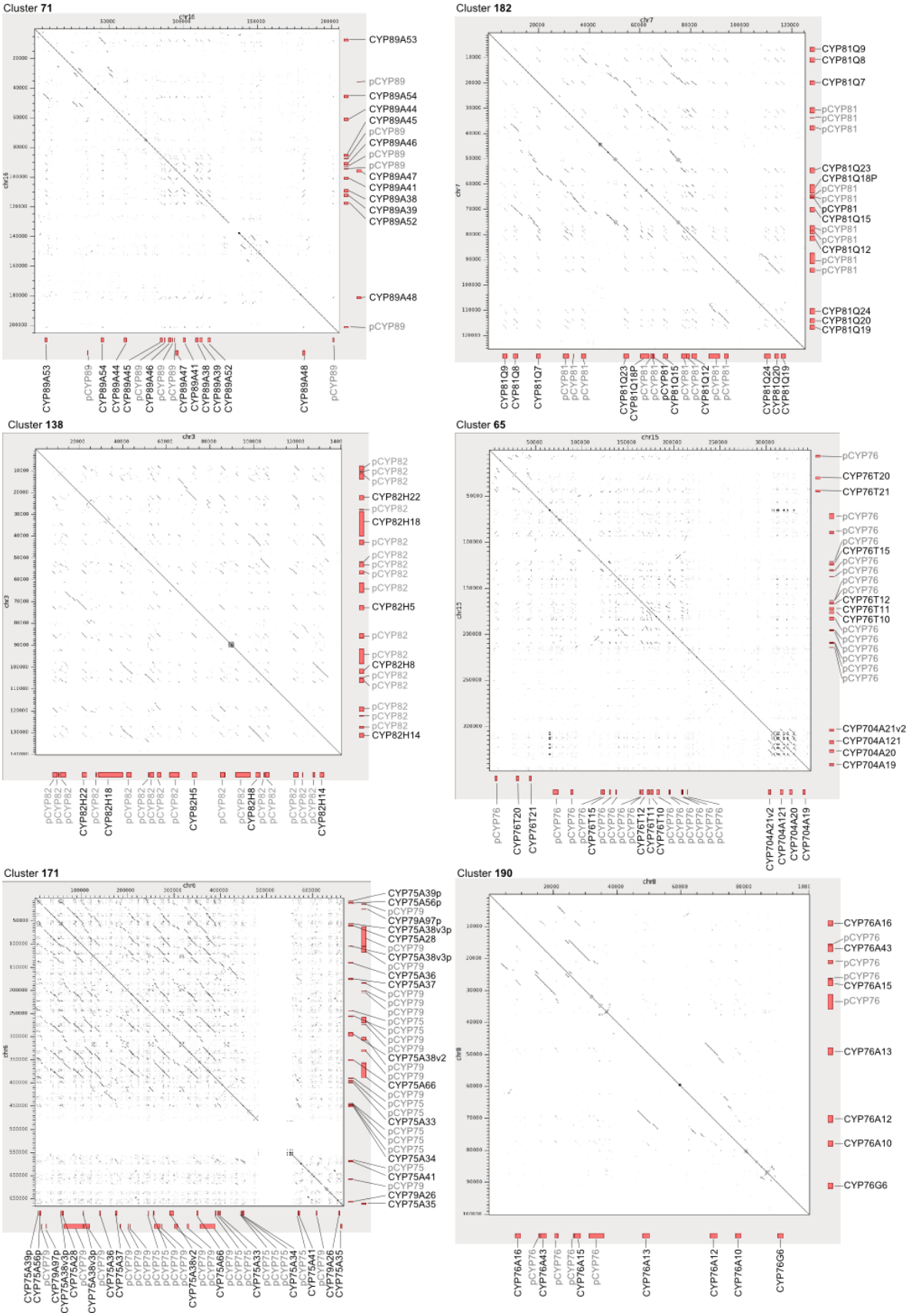



**Figure S5: Dot matrix plots of the largest physical clusters.** The dots and the black lines represent the sequence similarities in cluster 92 compared to itself. The red rectangles on the sides of the graph represent cytochrome P450 sequences. Complete genes are labeled with their name and pseudogenes are labeled with "p" and the P450 family.

**Table S1: List of P450 families with majority of its members grouped in physical clusters.**

| Family | Family size (total sequences) | Number of clustered members | Examples of functions in other organisms |
|---|---|---|---|
| CYP82 | 69 | 66 | Biosynthesis of homoterpenes in *A. thaliana* (Lee *et al.*, 2010), opioids in *Papaver somniferum* (Beaudoin and Facchini, 2013; Dang and Facchini, 2014; Farrow *et al.*, 2015). |
| CYP71 | 51 | 45 | Biosynthesis of monoterpenoids in mint species *Mentha x piperita* and *Mentha x spicata* (Haudenschild *et al.*, 2000), cyanogenic glucosides in cassava (*Manihot esculenta*) (Jørgensen *et al.*, 2011), furanocoumarins in several species (Larbat *et al.*, 2009), artemisinin in *Artemisia annua* (Teoh *et al.*, 2006), flavonoids in soybean (*Glycine max*). (Ayabe and Akashi, 2006) |
| CYP81 | 50 | 42 | Biosynthesis of indole glucosinolates in *A. thaliana* (Pfalz *et al.*, 2009), isoflavonoid phytoalexins in *Medicago truncatula, G. echinata* and*, Lotus japonicus* (Ayabe and Akashi, 2006), sesamin in *Sesamum spp.* (Ono *et al.*, 2006) |
| CYP76 | 42 | 42 | Biosynthesis of monoterpene volatiles in *A. thaliana* (Boachon *et al.*, 2015), monoterpene indole alkaloids in *Catharanthus roseus* (Collu *et al.*, 2001; Miettinen *et al.*, 2014), sesquiterpene volatiles in sandalwood (*Santalum album*) (Diaz-Chavez *et al.*, 2013), phytoalexins in rice (*Oryza sativa*) (Swaminathan *et al.*, 2009; Wang *et al.*, 2012), tanshinones in Chinese sage (*Salvia miltiorrhiza*)(Guo *et al.*, 2013), pigment betalain in beetroot (*Beta vulgaris*) (Hatlestad *et al.*, 2012).Metabolism of xenobiotics in *A. thaliana* (Hofer *et al.*, 2014). |
| CYP72 | 36 | 33 | Biosynthesis of monoterpene indole alkaloids in *C. roseus* (Irmler *et al.*, 2000; Miettinen *et al.*, 2014), glycyrrhizin in licorice (*Glycyrrhiza*) (Seki *et al.*, 2011), saponins in *M. trucantula* (Biazzi *et al.*, 2015). |
| CYP79 | 26 | 25 | Biosynthesis of glucosinolates in *A. thaliana* (Mikkelsen *et al.*, 2000; Wittstock, 2000; Hansen, 2001), cyanogenic glucosides in cassava (*M. esculenta*) (Andersen, 2000). |
| CYP89 | 25 | 21 | Chlorophyll degradation in *A. thaliana* (Christ *et al.*, 2013). |
| CYP75 | 24 | 23 | Biosynthesis of flavonoids in *Petunia x hybrida*, *A. thaliana, Gentiana triflora, C. roseus, etc*. (Ayabe and Akashi, 2006) |
| CYP716 | 23 | 12 | Biosynthesis of saponins in *M. trucantula* (Carelli *et al.*, 2011) and *Maesa lanceolata* (Moses *et al.*, 2014). |
| CYP706 | 21 | 19 | Biosyntheis of sesquiterpenoids in cotton (*Gossypium arboreum*) (Luo *et al.*, 2001). |
| CYP87 | 20 | 15 | Biosynthesis of saponins in *Maesa lanceolata* (Moses *et al.*, 2014). |
| CYP714 | 16 | 13 | Degradation of hormones (gibberelin) in rice (*O. sativa*) (Magome *et al.*, 2013). |
| CYP736 | 13 | 11 | Unknown. Pathogen response in grapevine *V. vinifera* (Cheng *et al.*, 2010). |
| CYP728 | 11 | 9 | Unknown. |
| CYP80 | 10 | 9 | Alkaloid biosynthesis in barberry *(Berberis stolonifera)* (Kraus and Kutchan, 1995) and California poppy (*Eschscholzia californica*) (Pauli and Kutchan, 1998). |
| CYP96 | 9 | 9 | Biosynthesis of cuticular wax in *A. thaliana* (Greer *et al.*, 2007). |
| CYP721 | 8 | 8 | Unknown. |
| CYP74 | 7 | 7 | Biosynthesis of hormones (jasmonates) and C6 volatiles in *A. thaliana* and other plants (Laudert *et al.*, 1996; Park *et al.*, 2002; Hughes *et al.*, 2009). |
| CYP92 | 7 | 5 | Unknown. |
| CYP93 | 7 | 6 | Biosynthesis of flavonoids in soybean (*G. max*), *Glycyrrhiza echinata*, *Gerbera hybrid*, *Antirrhinum majus*, *Torrenia hybrid*, etc. (Ayabe and Akashi, 2006) |
| CYP712 | 6 | 5 | Unknown. |

**Table S2: Description of RNA-Seq experiments used for analysis of gene expression.**

| Genotype | Tissue | Conditions | Reference |
|---|---|---|---|
| Carignan | Leaves | Infection with powdery mildew pathogen *Erysiphe nectator* | (Jones *et al.*, 2014) |
| Pinot noir | Leaves | Infection with downy mildew pathogen *Plasmopara viticola* | (Perazzolli *et al.*, 2012) |
| Touriga Nacional | Flowers | Development | (Ramos *et al.*, 2014) |
| Tannat | Berries | Development | (Da Silva *et al.*, 2013) |
| Shiraz | Berries | Development | (Sweetman *et al.*, 2012) |
| Corvina | Berries | Development | (Venturini *et al.*, 2013) |
| 4 cultivars | Berries | Development | Unpublished |

# References

**Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J.** (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

**Anastasiou, E., Kenz, S., Gerstung, M., MacLean, D., Timmer, J., Fleck, C. and Lenhard, M.** (2007) Control of Plant Organ Size by KLUH/CYP78A5-Dependent Intercellular Signaling. *Dev. Cell*, **13**, 843–856.

**Anders, S., Pyl, P.T. and Huber, W.** (2015) HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

**Andersen, M.D.** (2000) Cytochromes P-450 from Cassava (Manihot esculenta Crantz) Catalyzing the First Steps in the Biosynthesis of the Cyanogenic Glucosides Linamarin and Lotaustralin. *J. Biol. Chem.*, **275**, 1966–1975.

**Anon** (2015) Food and Agriculture Organisation fo the United Nations. Food and Agricultural commodities production / Commodities by region.

**Ayabe, S. and Akashi, T.** (2006) Cytochrome P450s in flavonoid metabolism. *Phytochem. Rev.*, **5**, 271–282.

**Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S. and Werck-Reichhart, D.** (2011) Cytochromes P450. *Arab. B.*, **9**, e0144.

**Beaudoin, G. a W. and Facchini, P.J.** (2013) Isolation and characterization of a cDNA encoding (S)-cis-N-methylstylopine 14-hydroxylase from opium poppy, a key enzyme in sanguinarine biosynthesis. *Biochem. Biophys. Res. Commun.*, **431**, 597–603.

**Biazzi, E., Carelli, M., Tava, A., Abbruscato, P., Losini, I., Avato, P., Scotti, C. and Calderini, O.** (2015) CYP72A67 catalyses a key oxidative step in Medicago truncatula hemolytic saponin biosynthesis. *Mol. Plant*, **8**, 1493–1506.

**Bisson, L.F., Waterhouse, A.L., Ebeler, S.E., Walker, M.A. and Lapsley, J.T.** (2002) The present and future of the international wine industry. *Nature*, **418**, 696–699.

**Boachon, B., Junker, R.R., Miesch, L., et al.** (2015) CYP76C1 (Cytochrome P450)-Mediated Linalool Metabolism and the Formation of Volatile and Soluble Linalool Oxides in Arabidopsis Flowers: A Strategy for Defense against Floral Antagonists. *Plant Cell*, **27**, 2972–90.

**Booker, J., Sieberer, T., Wright, W., et al.** (2005) MAX1 Encodes a Cytochrome P450 Family Member that Acts Downstream of MAX3/4 to Produce a Carotenoid-Derived Branch-Inhibiting Hormone. *Dev. Cell*, **8**, 443–449.

**Borneman, A.R., Schmidt, S.A. and Pretorius, I.S.** (2013) At the cutting-edge of grape and wine biotechnology. *Trends Genet.*, **29**, 263–271.

**Carelli, M., Biazzi, E., Panara, F., et al.** (2011) Medicago truncatula CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. *Plant Cell*, **23**, 3070–81.

**Castresana, J.** (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.

**Chakrabarti, M., Zhang, N., Sauvage, C., et al.** (2013) A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 17125–30.

**Cheng, D.W., Lin, H., Takahashi, Y., Walker, M.A., Civerolo, E.L. and Stenger, D.C.** (2010) Transcriptional regulation of the grape cytochrome P450 monooxygenase gene CYP736B expression in response to Xylella fastidiosa infection. *BMC Plant Biol.*, **10**, 135.

**Christ, B., Sussenbacher, I., Moser, S., Bichsel, N., Egert, A., Muller, T., Krautler, B. and Hortensteiner, S.** (2013) Cytochrome P450 CYP89A9 Is Involved in the Formation of Major Chlorophyll Catabolites during Leaf Senescence in Arabidopsis. *Plant Cell*, **25**, 1868–1880.

**Collu, G., Unver, N., Peltenburg-Looman, A.M.G., Heijden, R. van der, Verpoorte, R. and Memelink, J.**

(2001) Geraniol 10-hydroxylase, a cytochrome P450 enzyme involved in terpenoid indole alkaloid biosynthesis. *FEBS Lett.*, **508**, 215–220.

**Dang, T.-T.T. and Facchini, P.J.** (2014) CYP82Y1 Is N-Methylcanadine 1-Hydroxylase, a Key Noscapine Biosynthetic Enzyme in Opium Poppy. *J. Biol. Chem.*, **289**, 2013–2026.

**Diaz-Chavez, M.L., Moniodis, J., Madilao, L.L., et al.** (2013) Biosynthesis of Sandalwood Oil: Santalum album CYP76F Cytochromes P450 Produce Santalols and Bergamotol. *PLoS One*, **8**, e75053.

**Duchêne, E., Huard, F., Dumas, V., Schneider, C. and Merdinoglu, D.** (2010) The challenge of adapting grapevine varieties to climate change. *Clim. Res.*, **41**, 193–204.

**Duchêne, E. and Schneider, C.** (2005) Grapevine and climatic changes: a glance at the situation in Alsace. *Agron. Sustain. Dev.*, **25**, 93–99.

**Edgar, R.C.** (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

**Ehlting, J., Hamberger, B., Million-Rousseau, R. and Werck-Reichhart, D.** (2006) Cytochromes P450 in phenolic metabolism. *Phytochem. Rev.*, **5**, 239–270.

**Falginella, L., Castellarin, S.D., Testolin, R., Gambetta, G. a, Morgante, M. and Gaspero, G. Di** (2010) Expansion and subfunctionalisation of flavonoid 3',5'-hydroxylases in the grapevine lineage. *BMC Genomics*, **11**, 562.

**Falginella, L., Gaspero, G. Di and Castellarin, S.D.** (2012) Expression of flavonoid genes in the red grape berry of "Alicante Bouschet" varies with the histological distribution of anthocyanins and their chemical composition. *Planta*, **236**, 1037–51.

**Farrow, S.C., Hagel, J.M., Beaudoin, G. a W., Burns, D.C. and Facchini, P.J.** (2015) Stereochemical inversion of (S)-reticuline by a cytochrome P450 fusion in opium poppy. *Nat. Chem. Biol.*, **11**, 728–732.

**Feyereisen, R.** (2011) Arthropod CYPomes illustrate the tempo and mode in P450 evolution. *Biochim. Biophys. Acta - Proteins Proteomics*, **1814**, 19–28.

**Galtier, N., Gouy, M. and Gautier, C.** (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*, **12**, 543–548.

**Gouy, M., Guindon, S. and Gascuel, O.** (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol. Biol. Evol.*, **27**, 221–224.

**Gray, D.J., Li, Z.T. and Dhekney, S.A.** (2014) Precision breeding of grapevine (Vitis vinifera L.) for improved traits. *Plant Sci.*, **228**, 3–10.

**Greer, S., Wen, M., Bird, D., Wu, X., Samuels, L., Kunst, L. and Jetter, R.** (2007) The cytochrome P450 enzyme CYP96A15 is the midchain alkane hydroxylase responsible for formation of secondary alcohols and ketones in stem cuticular wax of Arabidopsis. *Plant Physiol.*, **145**, 653–667.

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

**Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O.** (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.

**Guo, J., Zhou, Y.J., Hillwig, M.L., et al.** (2013) CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts. *Proc. Natl. Acad. Sci.*, 1–6.

**Hamberger, B. and Bak, S.** (2013) Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **368**, 20120426.

**Hannah, L., Roehrdanz, P.R., Ikegami, M., Shepard, A. V, Shaw, M.R., Tabor, G., Zhi, L., Marquet, P. a and Hijmans, R.J.** (2013) Climate change, wine, and conservation. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 6907–12.

**Hansen, C.H.** (2001) Cytochrome P450 CYP79F1 from Arabidopsis Catalyzes the Conversion of Dihomomethionine and Trihomomethionine to the Corresponding Aldoximes in the Biosynthesis of Aliphatic Glucosinolates. *J. Biol. Chem.*, **276**, 11078–11085.

**Hatlestad, G.J., Sunnadeniya, R.M., Akhavan, N. a, Gonzalez, A., Goldman, I.L., McGrath, J.M. and Lloyd, A.M.** (2012) The beet R locus encodes a new cytochrome P450 required for red betalain production. *Nat. Genet.*, **44**, 816–20.

**Haudenschild, C., Schalk, M., Karp, F. and Croteau, R.** (2000) Functional expression of regiospecific cytochrome P450 limonene hydroxylases from mint (Mentha spp.) in Escherichia coli and Saccharomyces cerevisiae. *Arch. Biochem. Biophys.*, **379**, 127–36.

**Helliwell, C.A., Sheldon, C.C., Olive, M.R., Walker, A.R., Zeevaart, J.A.D., Peacock, W.J. and Dennis, E.S.** (1998) Cloning of the Arabidopsis ent-kaurene oxidase gene GA3. *Proc. Natl. Acad. Sci.*, **95**, 9019–9024.

**Hofer, R., Boachon, B., Renault, H., et al.** (2014) Dual function of the cytochrome P450 CYP76 family from Arabidopsis thaliana in the metabolism of monoterpenols and phenylurea herbicides. *Plant Physiol.*, **166**, 1149–1161.

**Hughes, R.K., Domenico, S. De and Santino, A.** (2009) Plant cytochrome CYP74 family: Biochemical features, endocellular localisation, activation mechanism in plant defence and improvements for industrial applications. *ChemBioChem*, **10**, 1122–1133.

**Ilc, T., Halter, D., Miesch, L., et al.** (2016) A grapevine cytochrome P450 generates the precursor of wine lactone, a key odorant in wine. *New Phytol.*

**Irmler, S., Schröder, G., St-Pierre, B., Crouch, N.P., Hotze, M., Schmidt, J., Strack, D., Matern, U. and Schröder, J.** (2000) Indole alkaloid biosynthesis in Catharanthus roseus: New enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. *Plant J.*, **24**, 797–804.

**Ito, T. and Meyerowitz, E.M.** (2000) Overexpression of a gene encoding a cytochrome P450, CYP78A9, induces large and seedless fruit in arabidopsis. *Plant Cell*, **12**, 1541–50.

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

**Jones, L., Riaz, S., Morales-Cruz, A., Amrine, K.C.H., McGuire, B., Gubler, W.D., Walker, M.A. and Cantu, D.** (2014) Adaptive genomic structural variation in the grape powdery mildew pathogen, Erysiphe necator. *BMC Genomics*, **15**, 1081.

**Jørgensen, K., Morant, A.V., Morant, M., Jensen, N.B., Olsen, C.E., Kannangara, R., Motawia, M.S., Møller, B.L. and Bak, S.** (2011) Biosynthesis of the cyanogenic glucosides linamarin and lotaustralin in cassava: isolation, biochemical characterization, and expression pattern of CYP71E7, the oxime-metabolizing cytochrome P450 enzyme. *Plant Physiol.*, **155**, 282–92.

**Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L.** (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

**Kraus, P.F. and Kutchan, T.M.** (1995) Molecular cloning and heterologous expression of a cDNA encoding berbamunine synthase, a C--O phenol-coupling cytochrome P450 from the higher plant Berberis stolonifera. *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 2071–2075.

**Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A.** (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–45.

**Langridge, P. and Fleury, D.** (2011) Making the most of "omics" for crop breeding. *Trends Biotechnol.*, **29**, 33–40.

**Larbat, R., Hehn, A., Hans, J., Schneider, S., Jugdé, H., Schneider, B., Matern, U. and Bourgaud, F.** (2009) Isolation and functional characterization of CYP71AJ4 encoding for the first P450 monooxygenase of angular furanocoumarin biosynthesis. *J. Biol. Chem.*, **284**, 4776–85.

**Laudert, D., Pfannschmidt, U., Lottspeich, F., Holländer-Czytko, H. and Weiler, E.W.** (1996) Cloning, molecular and functional characterization of Arabidopsis thaliana allene oxide synthase (CYP 74), the first

enzyme of the octadecanoid pathway to jasmonates. *Plant Mol. Biol.*, **31**, 323–335.

**Lee, S., Badieyan, S., Bevan, D.R., Herde, M., Gatz, C. and Tholl, D.** (2010) Herbivore-induced and floral homoterpene volatiles are biosynthesized by a single P450 enzyme (CYP82G1) in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 21205–10.

**Luo, P., Wang, Y.H., Wang, G.D., Essenberg, M. and Chen, X.Y.** (2001) Molecular cloning and functional identification of (+)-delta-cadinene-8-hydroxylase, a cytochrome P450 mono-oxygenase (CYP706B1) of cotton sesquiterpene biosynthesis. *Plant J.*, **28**, 95–104.

**Magome, H., Nomura, T., Hanada, A., et al.** (2013) CYP714B1 and CYP714B2 encode gibberellin 13-oxidases that reduce gibberellin activity in rice. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 1947–52.

**Martin, D.M., Aubourg, S., Schouwey, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T. and Bohlmann, J.** (2010) Functional annotation, genome organization and phylogeny of the grapevine (Vitis vinifera) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.*, **10**, 226.

**Miettinen, K., Dong, L., Navrot, N., et al.** (2014) The seco-iridoid pathway from Catharanthus roseus. *Nat. Commun.*, **5**, 3606.

**Mikkelsen, M.D., Hansen, C.H., Wittstock, U. and Halkier, B. a** (2000) Cytochrome P450 CYP79B2 from Arabidopsis catalyzes the conversion of tryptophan to indole-3-acetaldoxime, a precursor of indole glucosinolates and indole-3-acetic acid. *J. Biol. Chem.*, **275**, 33712–7.

**Miller, M.A., Pfeiffer, W. and Schwartz, T.** (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *2010 Gateway Computing Environments Workshop (GCE)*. IEEE, pp. 1–8.

**Moses, T., Pollier, J., Faizal, A., Apers, S., Pieters, L., Thevelein, J.M., Geelen, D. and Goossens, A.** (2014) Unravelling the Triterpenoid Saponin Biosynthesis of the African Shrub Maesa lanceolata. *Mol. Plant*.

**Myles, S., Boyko, A.R., Owens, C.L., et al.** (2011) Genetic structure and domestication history of the grape. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 3530–3535.

**Nafisi, M., Goregaoker, S., Botanga, C.J., Glawischnig, E., Olsen, C.E., Halkier, B. a and Glazebrook, J.** (2007) Arabidopsis cytochrome P450 monooxygenase 71A13 catalyzes the conversion of indole-3-acetaldoxime in camalexin synthesis. *Plant Cell*, **19**, 2039–52.

**Nelson, D. and Werck-Reichhart, D.** (2011) A P450-centric view of plant evolution. *Plant J.*, **66**, 194–211.

**Nelson, D.R.** (2009) The cytochrome P450 homepage. *Hum. Genomics*, **4**, 59–65.

**Nelson, D.R., Ming, R., Alam, M. and Schuler, M.A.** (2008) Comparison of Cytochrome P450 Genes from Six Plant Genomes. *Trop. Plant Biol.*, **1**, 216–235.

**Nelson, D.R., Schuler, M. a, Paquette, S.M., Werck-Reichhart, D. and Bak, S.** (2004) Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. *Plant Physiol.*, **135**, 756–772.

**Nützmann, H.-W. and Osbourn, A.** (2014) Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.*, **26**, 91–9.

**Ono, E., Nakai, M., Fukui, Y., et al.** (2006) Formation of two methylenedioxy bridges by a Sesamum CYP81Q protein yielding a furofuran lignan, (+)-sesamin. *Proc. Natl. Acad. Sci.*, **103**, 10116–10121.

**Parage, C., Tavares, R., Rety, S., et al.** (2012) Structural, Functional, and Evolutionary Analysis of the Unusually Large Stilbene Synthase Gene Family in Grapevine. *Plant Physiol.*, **160**, 1407–1419.

**Park, J.H., Halitschke, R., Kim, H.B., Baldwin, I.T., Feldmann, K. a. and Feyereisen, R.** (2002) A knock-out mutation in allene oxide synthase results in male sterility and defective wound signal transduction in Arabidopsis due to a block in jasmonic acid biosynthesis. *Plant J.*, **31**, 1–12.

**Pauli, H.H. and Kutchan, T.M.** (1998) Molecular cloning and functional heterologous expression of two alleles encoding (S)-N-methylcoclaurine 3'-hydroxylase (CYP80B1), a new methyl jasmonate-inducible cytochrome P-450-dependent mono-oxygenase of benzylisoquinoline alkaloid biosynthesis. *Plant J.*, **13**, 793–801.

**Perazzolli, M., Moretto, M., Fontana, P., Ferrarini, A., Velasco, R., Moser, C., Delledonne, M. and Pertot, I.** (2012) Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. *BMC Genomics*, **13**, 660.

**Pfalz, M., Vogel, H. and Kroymann, J.** (2009) The Gene Controlling the Indole Glucosinolate Modifier1 Quantitative Trait Locus Alters Indole Glucosinolate Structures and Aphid Resistance in Arabidopsis. *Plant Cell*, **21**, 985–999.

**Quinlan, A.R. and Hall, I.M.** (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–2.

**R Development Core Team, R.** (2011) R: A Language and Environment for Statistical Computing R. D. C. Team, ed. *R Found. Stat. Comput.*, **1**, 409.

**Ramos, M.J., Coito, J., Silva, H., Cunha, J., Costa, M.M. and Rocheta, M.** (2014) Flower development and sex specification in wild grapevine. *BMC Genomics*, **15**, 1095.

**Richly, E., Kurth, J. and Leister, D.** (2002) Mode of amplification and reorganization of resistance genes during recent Arabidopsis thaliana evolution. *Mol. Biol. Evol.*, **19**, 76–84.

**Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B.** (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

**Seki, H., Sawai, S., Ohyama, K., et al.** (2011) Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. *Plant Cell*, **23**, 4112–23.

**Seo, J., Gordish-Dressman, H. and Hoffman, E.P.** (2006) An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics*, **22**, 808–14.

**Silva, C. Da, Zamperin, G., Ferrarini, A., 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.

**Slater, G.S.C. and Birney, E.** (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

**Sonnhammer, E.L.L. and Durbin, R.** (1995) A Dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, 1–10.

**Stamatakis, A.** (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

**Swaminathan, S., Morrone, D., Wang, Q., Fulton, D.B. and Peters, R.J.** (2009) CYP76M7 is an ent-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell*, **21**, 3315–25.

**Sweetman, C., Wong, D.C., Ford, C.M. and Drew, D.P.** (2012) Transcriptome analysis at four developmental stages of grape berry (Vitis vinifera cv. Shiraz) provides insights into regulated and coordinated gene expression. *BMC Genomics*, **13**, 691.

**Takos, A.M. and Rook, F.** (2012) Why biosynthetic genes for chemical defense compounds cluster. *Trends Plant Sci.*, **17**, 383–388.

**Teoh, K.H., Polichuk, D.R., Reed, D.W., Nowak, G. and Covello, P.S.** (2006) Artemisia annua L. (Asteraceae) trichome-specific cDNAs reveal CYP71AV1, a cytochrome P450 with a key role in the biosynthesis of the antimalarial sesquiterpene lactone artemisinin. *FEBS Lett.*, **580**, 1411–6.

**The angiosperm phylogeny group** (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.*, **161**, 105–121.

**The Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–814.

**Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M.J. van, Salzberg, S.L., Wold, B.J. and Pachter, L.** (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–5.

**Vannozzi, A., Dry, I.B., Fasoli, M., Zenoni, S. and Lucchin, M.** (2012) Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. *BMC Plant Biol.*, **12**, 130.

**Velasco, R., Zharkikh, A., Troggio, M., et al.** (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, **2**.

**Venturini, L., Ferrarini, A., Zenoni, S., et al.** (2013) De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. *BMC Genomics*, **14**, 41.

**Vitulo, N., Forcato, C., Carpinelli, E., 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.

**Wang, Q., Hillwig, M.L., Okada, K., Yamazaki, K., Wu, Y., Swaminathan, S., Yamane, H. and Peters, R.J.** (2012) Characterization of CYP76M5-8 indicates metabolic plasticity within a plant biosynthetic gene cluster. *J. Biol. Chem.*, **287**, 6159–68.

**Wellesen, K., Durst, F., Pinot, F., Benveniste, I., Nettesheim, K., Wisman, E., Steiner-Lange, S., Saedler, H. and Yephremov, A.** (2001) Functional analysis of the LACERATA gene of Arabidopsis provides evidence for different roles of fatty acid ω-hydroxylation in development. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 9694–9699.

**Wittstock, U.** (2000) Cytochrome P450 CYP79A2 from Arabidopsis thaliana L. Catalyzes the Conversion of L-Phenylalanine to Phenylacetaldoxime in the Biosynthesis of Benzylglucosinolate. *J. Biol. Chem.*, **275**, 14659–14666.

**Wu, T.D. and Nacu, S.** (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

**Wu, T.D. and Watanabe, C.K.** (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–75.

**Yang, S., Zhang, X., Yue, J.-X., Tian, D. and Chen, J.-Q.** (2008) Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics*, **280**, 187–198.

**Zohary, D. and Spiegel-Roy, P.** (1972) Beginnings of Fruit Growing in the Old World. *Science (80-. ).*, **187**, 319–327.

# Partie 3 : Étude des variations structurales de type CNV des gènes de résistance à domaine NBS chez la vigne

Bien que la vigne soit sensible à de multiples maladies, elle possède, comme d'autres plantes, de nombreux gènes de résistance dits gènes R. Ces gènes confèrent une résistance à certains virus, bactéries, champignons, oomycètes et nématodes et sont parmi les gènes les plus étudiés chez les plantes. Ils sont divisés en plusieurs familles dont le Nucleotide Binding Site (NBS), nommés d'après un domaine dans les protéines correspondantes à ces gènes, est le plus important en nombre. Les protéines codées par les gènes *NBS* sont connues pour avoir une fonction de récepteur immunitaire intracellulaire. Les gènes *NBS*, possédant pour la plupart une partie Leucine Rich Repeat (LRR), peuvent être divisés en deux sous-familles : les Toll and Interleukin-1 Receptor (TIR) et les non-TIR possédant, pour la plupart, un domaine Coiled-Coil (CC) mais pouvant posséder un domaine Leucine Zipper (LZ) ou aucun domaine N terminal.

Dans le cadre du projet de création variétale de l'INRA de Colmar, les gènes de résistance présentent un intérêt particulier afin d'obtenir des vignes résistantes aux maladies les plus dévastatrices, ce qui a motivé l'étude de cette famille dans le genre *Vitis*. Une collaboration a été réalisée, dans le cadre du projet HealthyGrape, avec d'autres chercheurs de l'INRA de Colmar et les unités ''Étude du Polymorphisme des Génomes végétaux'' et ''Unité de Recherche en Génomique Végétale'' à Evry et ''Unité de Recherche Génomique et Informatique'' à Versailles afin de mettre au point une méthode de détection de variations structurales utilisée à la fois sur les gènes *STS* dans le cadre de mon projet de thèse, mais également sur les gènes de résistance du type *NBS*.

Après analyse des variations structurales de la famille des gènes de résistance du type *NBS*, la majorité des gènes serait présente dans la plupart des génotypes testés. Cependant, certains gènes, dont les spécificités sont encore en cours d'analyse, présentent un plus fort taux de variations que les autres. Certains clusters entiers comptant plusieurs gènes seraient particulièrement soumis à variation. Enfin, il semblerait que, les différentes classes de gènes de type *NBS* ne subirait pas une dynamique évolutive similaire au sein du genre Vitis.

Dans ce projet j'ai participé à la mise au point et au perfectionnement de la méthode de détection de variations structurales, dite méthode de segmentation, décrite dans l'article ci-après.

# Evolutionary dynamics of the NBS resistance gene family from the grapevine reference genome to the *Vitis* genus

Guillaume Barnabé[1], Maharajah Ponnaiah[2], Cécile Guichard[3], Nadia Bentahar[3], Sophie Blanc[1], Gautier Arista[1], Marie-Christine Le Paslier[2], Anne-Françoise Adam-Blondon[4], Sébastien Aubourg[3], Dominique Brunel[2], Camille Rustenholz[1], Didier Merdinoglu[1]

**1** Université de Strasbourg, INRA, SVQV UMR-A 1131, F-68000 Colmar, France, **2** Unité « Etude du Polymorphisme des Génomes Végétaux », US INRA 1279, 2 Rue Gaston Crémieux, 91057 Évry Cedex, France, **3** Unité de Recherche en Génomique Végétale, UMR INRA 1165, Evry Cedex, France, **4** Unité de Recherche Génomique et Informatique, UR INRA 1164, Route de Saint-Cyr, 78026 Versailles Cedex, France

## Abstract

Grapevine is very sensitive to diseases like downy and powdery mildews, which cause important economic loss in viticulture. Alternatives to pesticides exist though the breeding of varieties by introgression of resistance factors identified in wild grapevine species. In this work, we performed an exhaustive annotation of the resistance genes carrying a NBS domain in the homozygous grapevine reference genome PN40024. A total of 829 *NBS-genes* were predicted with 450 potentially complete and functional. Among them, two main sub-families, TIR-NBS and CC-NBS, were identified representing 105 genes and 216 genes, respectively. These genes, found throughout the grapevine genome, are organized into clusters. A total of 747 genes were grouped together forming 122 clusters, with a range of 2 to 25 genes. Most of these clusters are composed of genes from the same sub-family. The evolution analysis of the *Vitis* genus, through CNV detection, revealed a high level of conservation in the whole Vitis genus. Between 4% and 14% of the NBS-genes were potentially absent in the studied genotypes. The *Muscadinia* species showed the highest CNV rate. A lot of genes and particularly CC-NBS clusters were classified as "partially detected", that can be considered as truncated or hemizygous. Altogether, these results suggest that the evolution of the NBS-gene family occurs mainly at the intra-cluster scale through tandem duplications and local rearrangements.

**Keywords**: NBS genes, resistance genes, CNV, evolution, grapevine, *Vitis vinifera*

# Introduction

Grapevine is, economically, the most important fruit trees, and is cultivated worldwide (Vivier and Pretorius, 2002). In France 3% of the agricultural land area constitutes grapevine cultivation. However, this plant is sensitive to a lot of diseases, like the powdery and the downy mildew, with the result of an important economic loss. With the aim to fight these diseases, wine-growers spread each year a lot of pesticides and fungicides. This high use is pointed out in an expert synthesis report (INRA/Cemagref, 2005). It calls for an ideal strategy to improve varieties against pests, and other factors, with a decrease of the pesticide usage.

Grapevines in general are classified into the *Vitis* genus, consisting of two sub-genera, *Euvitis* and *Muscadinia*. These two sub-genera differ in chromosome number (2n=38 and 2n=40, respectively) (Riaz *et al.*, 2008; Blanc *et al.*, 2012), adding limitations to yield a fertile hybrid between them. The *Euvitis* sub-genus comprises more than 60 species (>30 native to China and ~34 in North and Central America), while a single *Vitis* species, *Vitis vinifera*, has originated from Europe (Vivier and Pretorius, 2002). This European grapevine is subdivided into two subspecies, *sativa* (or *vinifera*), the cultivated grapevine, and *sylvestris*, its wild progenitor (Levadoux, 1956; Zohary and Hopf, 1973; Barnaud *et al.*, 2010). The species of the *Muscadinia* sub-genus have a North American origin and are recognized as a major source for resistance to biotic factors.

A large number of disease resistance genes (R-genes) conferring resistance to several pathogens, including viruses, bacteria, fungi, oomycetes, and nematodes, are being isolated from a wide range of plant species (Dangl and Jones, 2001; Meyers *et al.*, 2003; McHale *et al.*, 2006; Liu *et al.*, 2007; Luz *et al.*, 2013). Among the R-genes, the *NBS* genes, which encode for proteins containing Nucleotide Binding Sites (NBS) domain, constitute the largest group (Martin *et al.*, 2003; Belkhadir *et al.*, 2004; Luo *et al.*, 2012; Wan *et al.*, 2012; Liu *et al.*, 2007). These genes are among the most important and larger gene families studied on plants. The encoded proteins have been shown to function as intracellular immune receptors that recognize, directly or indirectly, specific pathogen effectors encoded by avirulence (Avr) genes (Bent and Mackey, 2007). NBS family are commonly classified into two major subfamilies, TIR (Toll and Interleukin-1 Receptor) and non-TIR subfamilies, based on predicted motifs and specific amino acid residues in the proteins (Meyers *et al.*, 1999). Most of the proteins encoded by non-TIR genes have a Coiled-Coil (CC) motif or Leucine Zipper (LZ) (Liu *et al.*, 2007; Luz *et al.*, 2013).

Genome-wide comparison studies made on rice genomes showed that many R gene loci are conserved between the cultivars (Yang *et al.*, 2006), which helps underpinning isolation of R genes using in silico mapping and cloning (Lin and Chen, 2007; Yuan *et al.*, 2011; Liu *et al.*, 2007). This conservation between cultivars is also well documented between genotypes (Grant *et al.*, 1995; Henk *et al.*, 1999; Shen *et al.*, 2006; Yang *et al.*, 2006; Luo *et al.*, 2012). The dynamism (presence/absence polymorphism) of R genes is widely reported as a significant factor of study to understand the R genes mechanisms and evolution (Bergelson *et al.*, 2001; Luo *et al.*, 2012). Evolutionary studies of *Arabidopsis* R genes (Guo *et al.*, 2011) reveal that only about more than half of all R genes share orthologs with other genotypes, supporting evidence for rapid births and deaths of R genes (Michelmore and Meyers, 1998). The exhibit of divergence is not only based on the number of genes, but also on the subclasses of R genes, as *TIR-NBS-LRR* (*TNL*) genes tend to be preferentially lost in grasses and other monocots (Mondragón-Palomino *et al.*, 2002; Akita and Valkonen, 2002; Bai *et al.*, 2002; McDowell and Simon, 2006; Bomblies and Weigel, 2010; Meyers *et al.*, 1999), while predominant in dicot genomes (Yang *et al.*, 2008; Porter *et al.*, 2009; Meyers *et al.*, 2003). In grapevine, Malacarne and co-workers (2012) have predicted 391 resistance gene annotations (RGAs), out of which 346 were anchored on the chromosomes and including CC and TIR types of NBS-LRRs. They studied the transposition around the RGAs and proposed a hypothesis to explain the hexaploid state, that would result from the fusion of two groups of chromosomes (Va and Vc).

In this paper, we present the expertized annotation of the *NBS* gene family in the 12x.2 *Vitis vinifera* (PN40024) genome (National Center for Biotechnology Information Genome ID: 401). Classification of the identified *NBS-LRR* genes based on gene function, structure, phylogenies are described in detail. The first Copy Number Variation (CNV) analysis on the *NBS* gene family and on the NBS-genes clusters were performed and allowed to draw hypotheses on the evolutionary dynamics of this particular gene family in the *Vitis* genus based on 48 *Vitis* and 7 *Muscadinia* genotypes.

## Materials and Methods

### DataSets

The reference genome of genotype PN40024 version 12x.2 ((Jaillon *et al.*, 2007), https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences) was used for annotation and alignments.

The 391 protein and nucleic sequences of *NBS* genes annotated on the sequence of the heterozygous Pinot Noir genome (Malacarne *et al.*, 2012) were downloaded from the IASMA Research and Innovation Centre (http://genomics.research.iasma.it). The 177 *Arabidopsis* resistance genes sequences were downloaded from the protein databases of the Genome Center (http://niblrrs.ucdavis.edu/index.php).

Fifty-five genotypes were resequenced in the framework of various projects: the HealthyGrape project, GrapeReSeq project (PLANT - KBBE2008) and Muscares project (ANR-08-GENM-007). Among these genotypes, 26 are *Vitis vinifera* sp. *vinifera* varieties (PN40024 included), 4 are *Vitis vinifera* sp. *sylvestris*, 18 are other species of the *Vitis* genus (*Vitis aestivalis*, *Vitis amurensis*, *Vitis berlandieri*, *Vitis cinerea*, *Vitis labrusca*, *Vitis lincecumii*, *Vitis rupestris*) and 7 are *Muscadinia rotundofolia* (Supplementary data 1). About 10x sequencing depth of paired-end Illumina GAII (2x100bp) or HiSeq2000 (2x150bp) for each genotype was used

### Annotation of Resistance genes in the reference genome PN40024

The 12x.2 *Vitis vinifera* (PN40024) genome has been explored exhaustively thanks to a similarity search approach based on a reference set made with previously characterized NBS-LRR proteins. This set is composed of 216 reference sequences from different plant species available in Swiss-Prot (Bairoch and Boeckmann, 1991) and from the previous works of Meyers and his co-workers (Meyers *et al.*, 2003). Furthermore, based on the PN40024 annotated proteomes, HMMer (hmmer.org, Eddy, 2009) has also been applied with 4 different HMM profiles related to the NBS-LRR family (PF015282 for TIR, PF00931 for NB-ARC, PF05659 for RPW8 and PF00560 for LRR) which are defined in the PFAM resource (Finn *et al.*, 2014).

The hand-made curation of the annotation was performed using the ARTEMIS platform (Rutherford *et al.*, 2000). This re-annotation task took into account (i) the predictions of the three combiners GAZE (Genoscope; (Jaillon *et al.*, 2007)), EUGENE (URGV; (Schiex *et al.*, 2001)) and JIGSAW (CRIBI; (Allen and Salzberg, 2005)) which are based on predictors specifically trained for the *Vitis* genome annotation, (ii) the results of sequence comparisons (BLASTX and HMMer), and (iii) the spliced alignments of the available cognate transcript sequences (*Vitis* EST and cDNA available in the NCBI database). The goal was to correct and complete the automatic structural annotations and to discriminate between complete genes (*i.e.* perfect full CDS), partial genes (*i.e.* suspended by an unsequenced region) and pseudogenes (*i.e.* disrupted by numerous stop codons, frameshifts and/or small deletions). The curated structural annotation was controlled and completed by multiple alignments and functional annotation based on PFAM motifs and intron-exon gene structure. The prediction of coiled-coil regions was performed using Interproscan "Coils analyses" (version 2.2, (Lupas *et al.*, 1991)).

To define the *NBS-R* genes clusters, three criteria were used based on previous studies (Richly *et al.*, 2002; Yang *et al.*, 2008; Malacarne *et al.*, 2012; Meyers *et al.*, 2003). Two genes were defined as belonging to the same cluster if: (i) the length of the intergenic sequence was less than 200 kb, (ii) both genes were located on the same scaffold, and (iii) there were no more than 8 non *NBS-R* genes between them. The R package "ggplot" (Wickham, 2009) was used to represent clusters on the different chromosomes.

**Similarity analyses**

A multiple alignment of the 829 *R-genes* protein sequences against themselves was performed with the online clustalw2 tool (http://www.ebi.ac.uk/Tools/msa/clustalw2/, default options, (Larkin *et al.*, 2007)). From this alignment, a correlation matrix was created and the R package "heatmap3" (https://cran.r-project.org/web/packages/heatmap3/README.html, (Zhao *et al.*, 2014)) was used to show the similarity levels between each gene. The program Jalview Desktop (http://www.jalview.org/, version 2.9.0b2, (Waterhouse *et al.*, 2009)) was used for the multiple alignment of 5 "TIR" and 5 "CC" protein sequences.

The analysis of TIR and Coils separately was performed with Blast protein alignment (http://blast.ncbi.nlm.nih.gov/Blast.cgi, Camacho *et al.*, 2009) and the use of the Circos tools for the circular representation (http://circos.ca/, (Krzywinski *et al.*, 2009)).

The comparison between the 391 protein sequences of the Malacarne's gene set and our set of 829 *NBS-genes* was performed using reciprocal protein blast with options set to defaults (evalue = 10). Only *R-gene* reciprocal hits with the same chromosome location as the query, an identity score > 80%, and at least one of the query or hit overlap > 80%, were defined to be syntenic genes.

The comparison between the 177 protein sequences of the *Arabidopsis* gene set and our set of 829 *NBS-genes* was performed using a multiple alignment and a distance matrix established with clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/, (Sievers *et al.*, 2011)) with default options. The R package "heatmap3" was used again to represent the matrix.

**CNV analysis in the *Vitis* genus**

The resequencing data of the different accession genotypes were aligned against the PN40024 reference genome using gsnap (version 2013-11-27; Wu and Nacu, 2010) with the following parameters: -B4 -n3 -N0 -m12/18 (according to the read length used for the genome sequencing, 100 bp or 150 bp). The quality protocol Sanger or Illumina was selected depending on the quality encoding for the different datasets. The alignments were curated by (i) selecting only the alignments with the best score for reads with more than one hit and (ii) discarding alignments with equal scores for multiple hits. The depth was calculated for the regions of interest with two different approach (explain below). A normalization step was then performed using the alignment of PN40024 against itself as a reference by calculating the logratio value:

$$logratio = \log_2\left(\frac{(Average\ depth\ /\ Median\ depth\ for\ the\ chromosome)_{genome}}{(Average\ depth\ /\ Median\ depth\ for\ the\ chromosome)_{reference}}\right)$$

From these logratios, a deconvolution method was performed (adapted from Springer et al., 2009) and 5 categories were created. The first category, labelled "not detected", group the logratio values lower than -2. This category groups regions for which low alignment depth or no aligned reads could be found. The sequences are considered as absent in the genome of interest, or too divergent compared to the reference. The logratios close to 0 form the "detected" category, i.e. the normalized sequencing depth is similar for the genome of interest compared to the reference. The logratios greater than 2 are grouped in the "duplicate" category. In this case, the normalized sequencing depth is higher for the genome of interest compared to the reference. The two last categories, "partially detected" and "potentially duplicated", refer respectively to the logratio values lay between -2 and 0, and between 0 and 2. In the first one, the normalized sequencing depth is lower for the genome of interest compared to the reference.

These could be caused by partial or heterozygous absence in the regions. The "potentially duplicated" category refer to the normalized sequencing depth slightly greater for the genome of interest compared to the reference. These could be caused by partial or heterozygous duplications in the regions.

The first approach was performed at the gene scale. From the curated alignments, the depth was calculated for the exons of all the genes annotated in the PN40024 genome, *NBS-genes* included. After the logratio value was determined for each exon, a last step was involved to define the potential level of detection of the *NBS-genes* by merging exons results according to the following rules:

(1) Duplicated gene: ≥75% of the detected exons were duplicated.
(2) Potentially duplicated gene: at least one exon was half-duplicated or duplicated.
(3) Detected gene: ≥75% of the exons were detected.
(4) Partially detected gene: at least one exon was detected.
(5) Not detected gene: none of the exons were detected.

The second approach was performed at the scale of the NBS clusters. With the R package DNAcopy version 1.36.0 (Seshan and Olshen, 2014), a segmentation approach was undertaken to identify genomic regions harbouring significantly homogeneous logratios. The algorithm generates 10 kb windows throughout the 20 chromosomes, separated by a 1 kb step, and computes the average sequencing depth for the windows. The segments were classified into the five categories previously described using the same deconvolution method. The clusters were then classified as "not detected", "partially detected", "detected", "potentially duplicated" and "duplicated" regarding the major class of segments describing the region.

## Results

### Characterization of the *NBS-gene* family in the grapevine reference genome

Using our annotation pipeline, we defined 829 potential *NBS* resistance genes, out of which 54% were complete genes (450), 41% were pseudogenes (336) and 5% were partial genes (43). The identification of the resistance genes with a NBS domain in the grapevine reference genome PN40024 was managed by the PFAM and the Coils detection. We chose to classify into sub-families only the 450 complete genes, *i.e.* with a complete CDS. The incomplete protein sequences of the pseudogenes and the partial genes prevented from identifying clear structural domains. The classification showed the presence of the two main groups of *NBS* genes in plants, the CC-NBS with 211 *CC-NBS-LRR* genes and 5 *CC-NBS* genes, and the TIR-NBS with 94 *TIR-NBS-LRR* genes, 9 *TIR* genes and 2 *TIR-NBS-LIM* genes. Among the other *NBS* genes, we had also detected a huge number of *NBS-LRR* without clearly identified N-terminal domain (116 genes), and 13 with a RPW8 N-terminal domain (Table 1).

**Table 1: Sub-families of *NBS-genes* identified in the PN40024 genome and their schematic structure.** Only the 450 complete genes were annotated. The pseudogenes and the partial genes are grouped in the "pseudogenes" row.

| | Sub-families | Structure | No. of seq. |
|---|---|---|---|
| Completes genes | CC-NBS | CC NBS | 5 |
| | CC-NBS-LRR | CC NBS LRR | 211 |
| | NBS | NBS | 14 |
| | NBS-LRR | NBS LRR | 102 |
| | RPW8-NBS-LRR | RPW8 NBS LRR | 13 |
| | TIR | TIR | 9 |
| | TIR-NBS-LRR | TIR NBS LRR | 94 |
| | TIR-NBS-LIM | TIR NBS LIM | 2 |
| Pseudogenes | - | NA | 379 |

The identification and classification of *NBS* resistance genes was performed in the heterozygous Pinot Noir genome (Velasco *et al.*, 2007) and published by Malacarne and co-workers in 2012. A total of 391 *NBS-genes* were predicted, hereafter called *IASMA* genes. Their dataset was composed of 143 CC (111 CC-NBS-LRR and 32 CC-NBS), 33 TIR (27 TIR-NBS-LRR and 6 TIR-NBS) and 215 NBS (145 NBS-LRR and 70 NBS). Based on blast protein alignments, we compared this prediction with our results. About 60% (232) of the heterozygous Pinot Noir *NBS-genes* had a strong similarity with one of our homozygous PN40024 *NBS-genes* (Supplementary data 2). Three quarters of the matching genes were complete and between them, mostly were classified as *CC-NBS-LRR* (110 genes, data not shown). Some *TIR-NBS-LRR*, *NBS-LRR* and *RPW8-NBS-LRR* genes were also present in both datasets (22, 51 and 4 genes, respectively). A multiple alignment highlighted that among the 232 PN40024 genes with a good alignment, 100 had two strong Italian matches that shared close percentage identity and overlap. So, among the 159 *IASMA-specific* genes, only 59 are potentially specific. However, only six out of these 100 co-aligned genes of the heterozygous Pinot Noir, were localized on the same chromosome than the *IASMA* gene with the best match and were suspected of being an allelic form of the best match. For the other 94 genes, they were localized on a different chromosome.

**NBS-genes organization in the grapevine reference genome**

Despite their presence on every chromosome, the *NBS-genes* were unevenly distributed along the 20 chromosomes (19 chromosomes and an unknown chromosome that bring unanchored scaffolds together) of the PN40024 reference genome (Figure 1 and Supplementary data 3). They are mostly located on chromosomes 9, 13 and 18 with 103, 128, and 123 *NBS-genes*, respectively. Chromosomes 2, 4 and 8 are the poorest with 3, 1, and 2 *NBS-genes*, respectively. According to the literature, three parameters were chosen to define the NBS-gene clusters (see Material and Methods section). A total of 747 *NBS-genes* (90%) were grouped in 122 clusters (Figure 1). A range of 0 to 22 clusters are found per chromosome (6 on average) and they are composed of 2 to 26 genes (6 on average) for a length of 3,9kb to 1Mb (179 kb on average). As expected, the chromosomes with the greatest number of *NBS-genes* (chromosomes 9, 13 and 18) were also the chromosomes with the greatest number of clusters (respectively 12, 22 and 16). Notably, chromosome 12 carries one more cluster than chromosome 9 whereas it carries less *NBS-genes* (Figure 1 & Supplementary data 3). The only chromosome for which no cluster could be identified is chromosome 4, which carries only one *NBS-gene*.

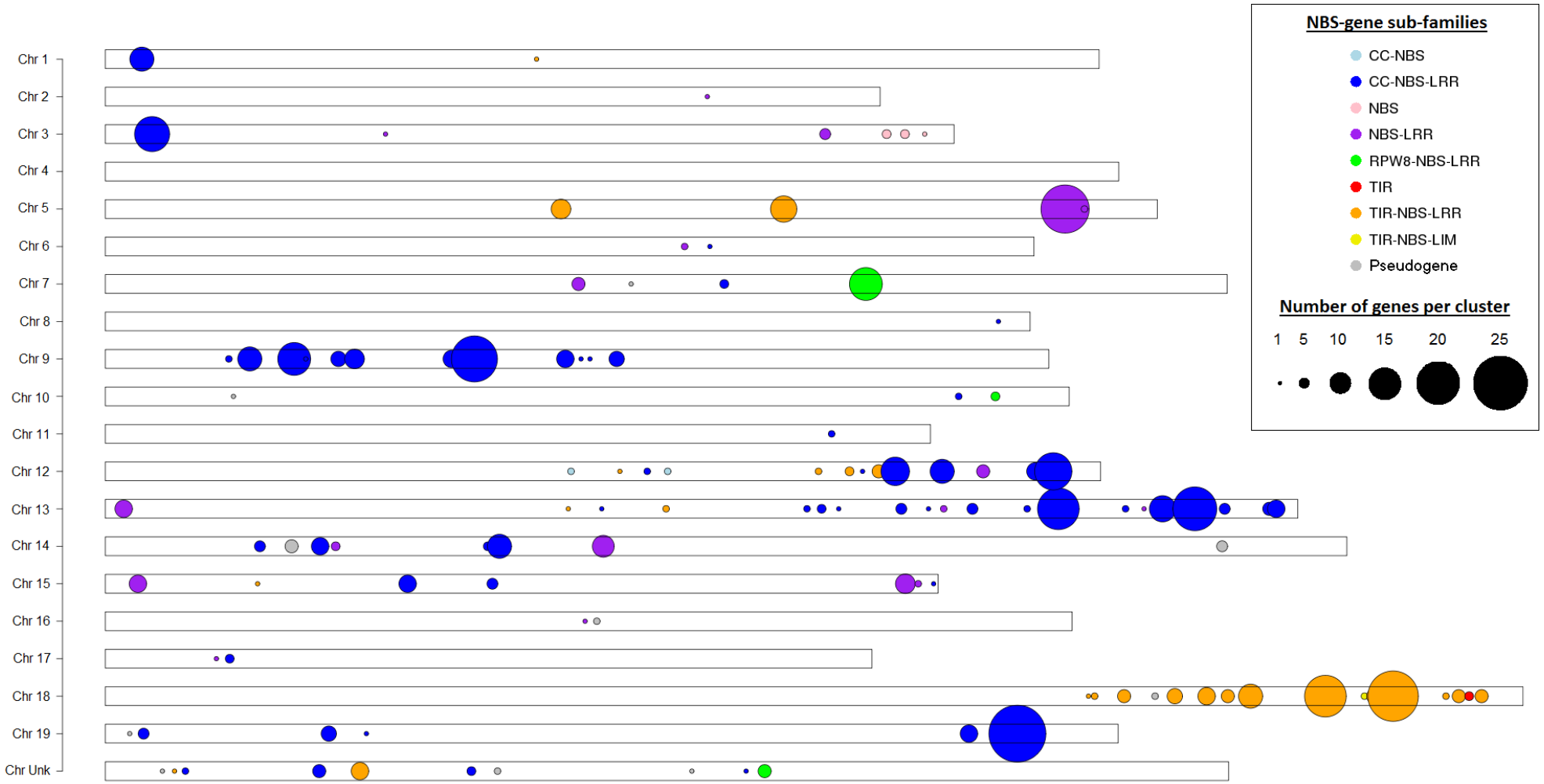



**Figure 1: Localization of NBS-gene clusters across the grapevine reference genome.** Representation of the 19 chromosomes of the grapevine genome. The unknown chromosome is the concatenation of the undefined scaffolds. Each cluster was represented by a solid circle, its size defining the total number of genes. The solid circle color showed the sub-family in majority in each cluster.

Most clusters are composed of a mix of complete, pseudo- and partial genes. Also, by considering only the complete genes, most clusters are composed of genes belonging to the same sub-family. Except for one cluster on chromosome 18 containing one *CC* and five *TIR NBS-genes*, 25 clusters with two different sub-families are composed of *CC-NBS-LRR* and *NBS/NBS-LRR* genes. Five other clusters are formed by *TIR* and *TIR-NBS-LRR* genes. More than 60% of the *NBS-genes* encoding a TIR domain (TIR, TIR-NBS-LRR and TIR-NBS-LIM) are localized at one extremity of chromosome 18. All these genes are part of 16 clusters and notably, the two *TIR-NBS-LIM* genes are clustered together with a *TIR-NBS-LRR* gene. On the contrary, the organization of the *CC-NBS-LRR* and *CC-NBS* genes do not harbour such a chromosome-specific pattern. They are spread in all the genome, chromosome 4 included, and form 59 of the 122 clusters (48%).

A bootstrap statistical analysis with 100,000 iterations was performed in order to check whether this level of clustering could be observed by chance. For each iteration, a set of 829 random genes were sampled and the clusters were identified using the same three parameters than the one we chose to define the NBS-gene clusters. Finally, the average number of clusters generated was greater (165), the total number of clustered genes (423), the mean length of clusters (148,503 bp) and the amount of genes per cluster (2.6 genes per cluster) were lower. The probability to obtain 747 clustered genes or more was found to be $< 0.00001$. Thus, this test confirms that the *NBS-gene* family is highly clustered in the grapevine genome.

**Evolution of the NBS-genes family**

The multiple alignment of the 829 *NBS-genes* with themselves highlighted the sequence diversity between the different sub-families. The genes with different protein structure shared less than 40% identity (Figure 2) whereas, for the genes belonging to the same family, a better percentage of identity was observed. By looking at the two main classes of *NBS-genes*, the ones encoding a TIR domain share a stronger sequence conservation (more than 50%), no matter the chromosome, than the ones encoding a CC domain. To find the cause of this greatest conservation of the *TIR* genes, the consensus alignment of 10 randomly selected *NBS-genes* (5 *TIR-NBS-LRR* and 5 *CC-NBS-LRR*; Supplementary data 4) was performed. The consensus sequence for the 5 *TIR NBS-genes* shows an important conservation of the protein sequence all along the different domains. However, the consensus sequence for the 5 *CC NBS-genes* shows a good conservation of its functional domains except for the Coil-coiled domain, which explains the general least conservation level of the *CC* compared to the *TIR NBS-genes*.

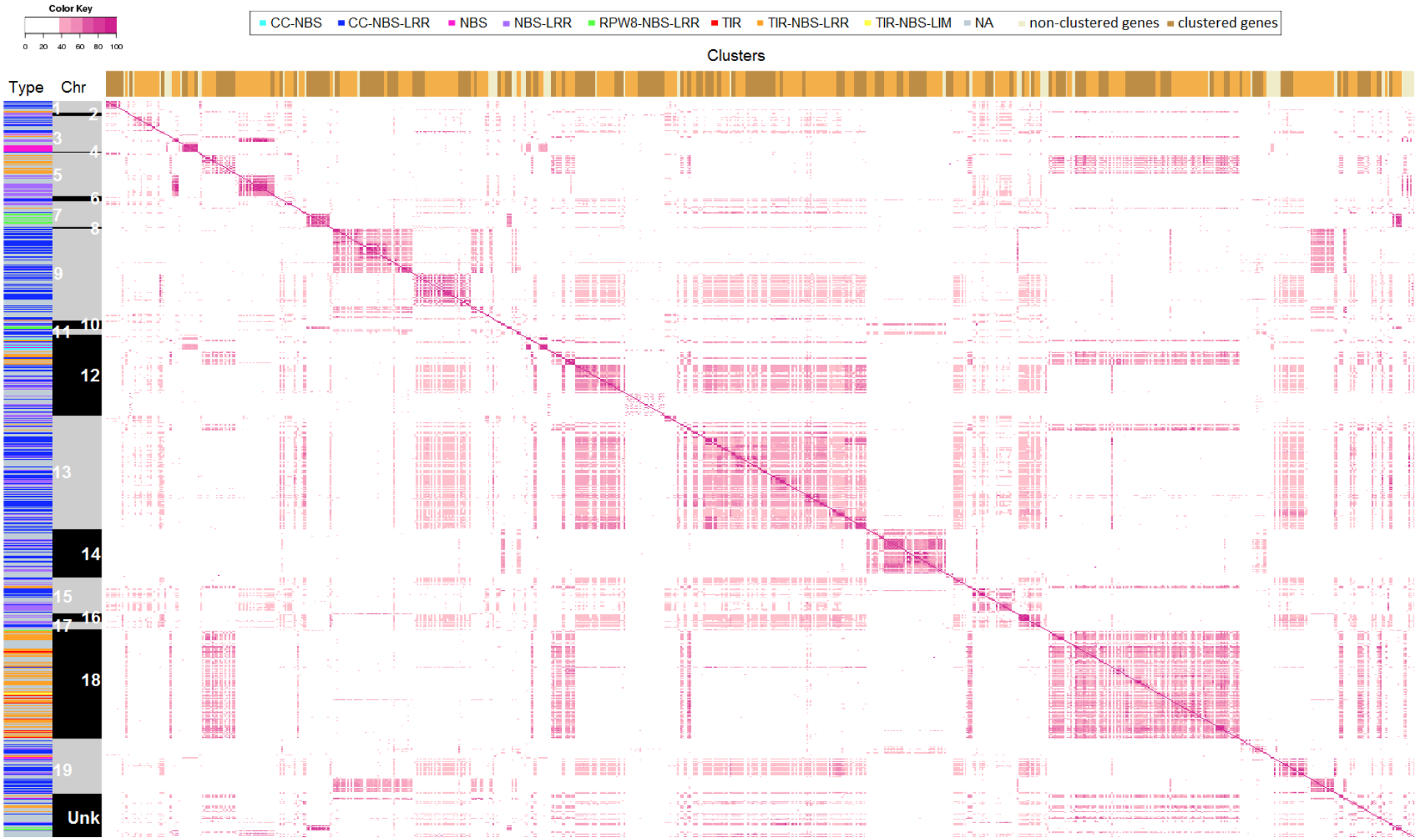



**Figure 2: Identity matrix of the 829 grapevine *NBS-genes* in the PN40024 reference genome.** The color is correlated with the similarity between genes, i.e. pink mean 40% of similarity between the two genes, purple 100%. The diagonal is the alignment of one gene with himself. The chromosomes and gene sub-families are indicated on the vertical axis, the clusters and lonely genes in the horizontal axis.

The similarity between genes with the same N-terminal domain seems correlated with their physical distance on the chromosome. For example, on Figure 2, the darkest spots of chromosome 18 can be observed along the diagonal, corresponding to the highly similar *TIR NBS-genes* that are close to each other. Generally, the more the physical distance between two genes increases, the more the percentage of identity decreases. Within clusters, the similarity is mostly very high and drops by considering genes in other clusters. However, there are few exceptions, for example on chromosome 16, which genes share high percentage identity even across clusters. Thus, we can hypothesize that the organization of the NBS clusters was mostly shaped by tandem duplications. However, it was shown in the *Arabidopsis* genome that the expansion of the *NBS* genes is the result of both tandem and large-scale segmental duplications (Leister, 2004; Yang *et al.*, 2008). To check whether the same events controlled the formation of the NBS-gene clusters in the grapevine genome, two nucleic blast analysis were performed between the 105 *TIR NBS-genes* and between the 216 *CC NBS-genes*, separately. For each gene, the best hit localized in a different cluster was represented on a circle representation (Figure 3). The *CC* genes that are spread on most of the chromosomes share mostly similarities with genes on the same chromosome (green lines) and fewer across different chromosomes (blue lines). Moreover, the observed similarity relationships between the *CC* genes involved mostly single genes and no clear segmental duplications could be identified. In contrast, the *TIR* genes, detected mainly at the end of chromosome 18, shares similarities with genes on the same chromosome (green lines) as well as with genes on a different chromosome (blue lines). Like the *CC* genes, the *TIR* genes were likely involved in many single gene duplications but three potential segmental duplications could be identified. Several genes of cluster 8 on chromosome 18 are similar to several genes of cluster 8 on chromosome 12, just like cluster 3 on chromosome 18 with cluster 1 on chromosome 5 and also clusters 9 and 12, both on chromosome 18. So the organization of the NBS clusters was mostly shaped by tandem and single gene duplications, except for the *TIR* genes, for which larger segmental duplications could be identified.

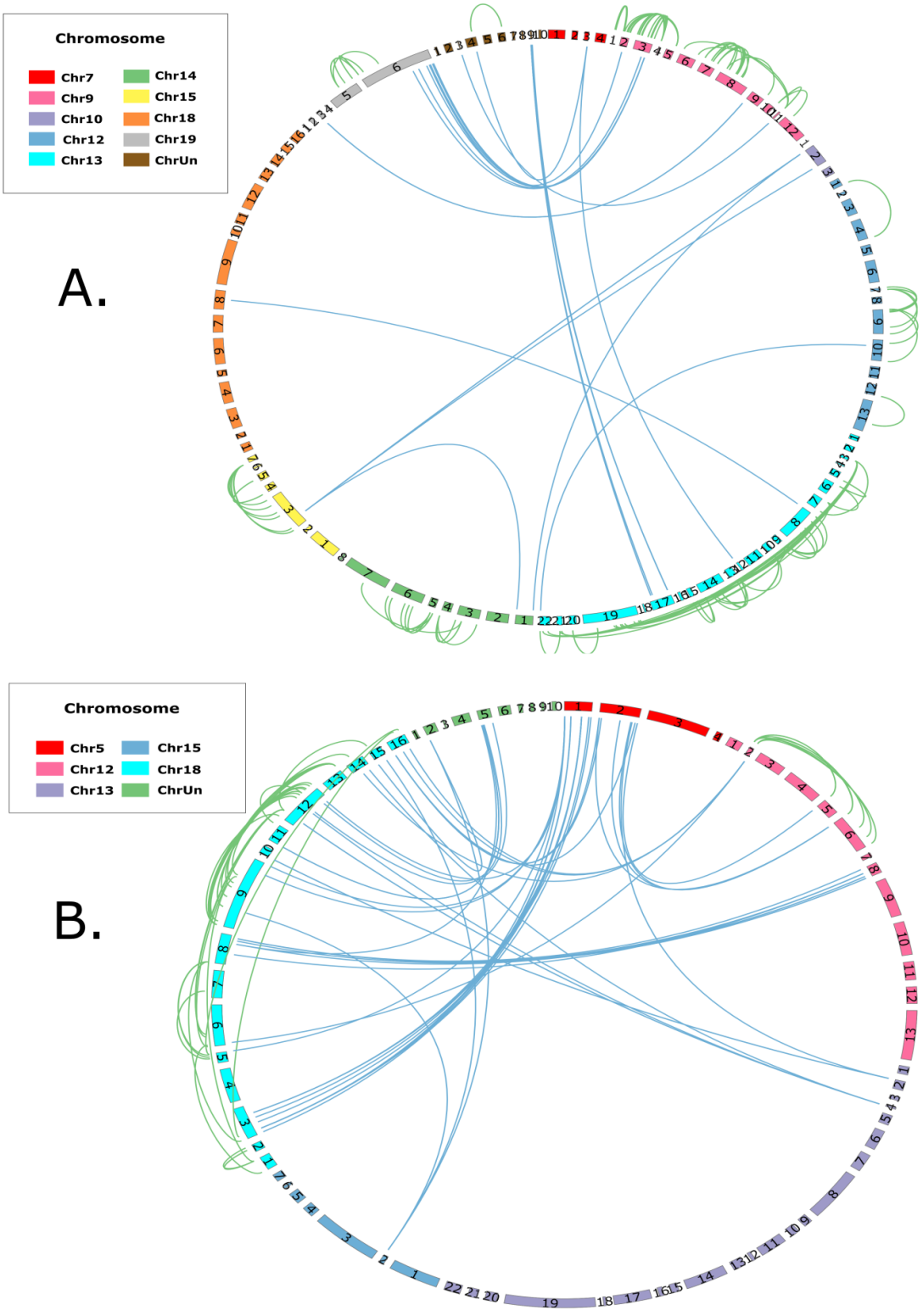



**Figure 3: Illustration of the duplication events that contribute to the evolution of the two main *NBS-gene* sub-families in the grapevine genome.** The link between genes localized in different NBS-genes cluster, represented by colored box, was performed by nucleic blast and was represented with the Circos tool for the *CC NBS-genes* (A), and the *TIR NBS-genes* (B). The similarities between genes of the same chromosome are represented by green link, those between genes of different chromosomes by blue link.

The first resistance gene inventory was performed by Meyers and coworkers in 1999 in the *Arabidopsis thaliana* genome. To perform a comparison with our *NBS-gene* annotations, we considered only the 177 *NBS-genes*, composed of 54 *CC-NBS-LRR*, 93 *TIR-NBS-LRR* and 30 *TIR genes*. Two major differences could be observed. First, the proportion of *NBS-genes* with a TIR N-terminal domain is greater than the ones with a CC domain (123 *TIR* genes for 54 *Coils* genes) in *Arabidopsis thaliana*. Moreover, whereas the *TIR NBS-genes* are mostly grouped on chromosome 18 in the grapevine genome, they are spread across the five chromosomes in the *Arabidopsis thaliana* genome. On the contrary, the *CC NBS-genes* are spread along multiple chromosomes, for both grapevine and *Arabidopsis thaliana* genome. The multiple alignment of the 829 grapevine *NBS-genes* with 177 *Arabidopsis NBS-genes* was preformed (Supplementary data 5). According to the matrix, the *Arabidopsis TIR NBS-genes* shared a good sequence homology with the grapevine *TIR NBS-genes* (between 40% and 60%). In spite of the difference of localization between the *TIR NBS-genes* of the two plants, they seemed to be well conserved. However, the *CC* genes of the two genomes shown two different behaviours. Some *CC* genes identified in the grapevine genome shared a good homology with numerous *Arabidopsis CC* genes, such as those at the beginning of the chromosome 9 of PN40024, for which homologies were found with several *Arabidopsis* clusters on chromosome 1, 4 and 5. The fact that grapevine genes have multiple homologs in *Arabidopsis* genome is consistent with the results of Jaillon and coworkers (2007). An opposite situation is observed on grapevine chromosomes 9, 12, 13, and 19 carrying highly similar genes, for which homology could be found only with a single *Arabidopsis* cluster on chromosome 3.

**CNV analysis of the NBS-genes in the Vitis genus**

To investigate the presence / absence and copy number variations of the *NBS-genes* in 54 genotypes of the *Vitis* genus, a CNV analysis was performed at the gene scale (Figure 4 and Supplementary data 6 and 7). On the 44766 total calls (829 *NBS-genes* x 54 genotypes), 10% were grouped in the "not detected" category, *i.e.* these *NBS-genes* were potentially absent in some genotypes. The "detected" category, *i.e.* the genes that are present, represented 56% of the total calls. The "partially detected" category, *i.e.* genes that are truncated or hemizygous, represented 28%. The "potentially duplicated" and the "duplicated" categories are more anecdotal with 6% and 1%, respectively. Our analysis revealed that the *NBS-genes* are globally well conserved among the *Vitis* genus and that entire gene deletions or duplications not be the major evolutionary mechanisms shaping this gene family.

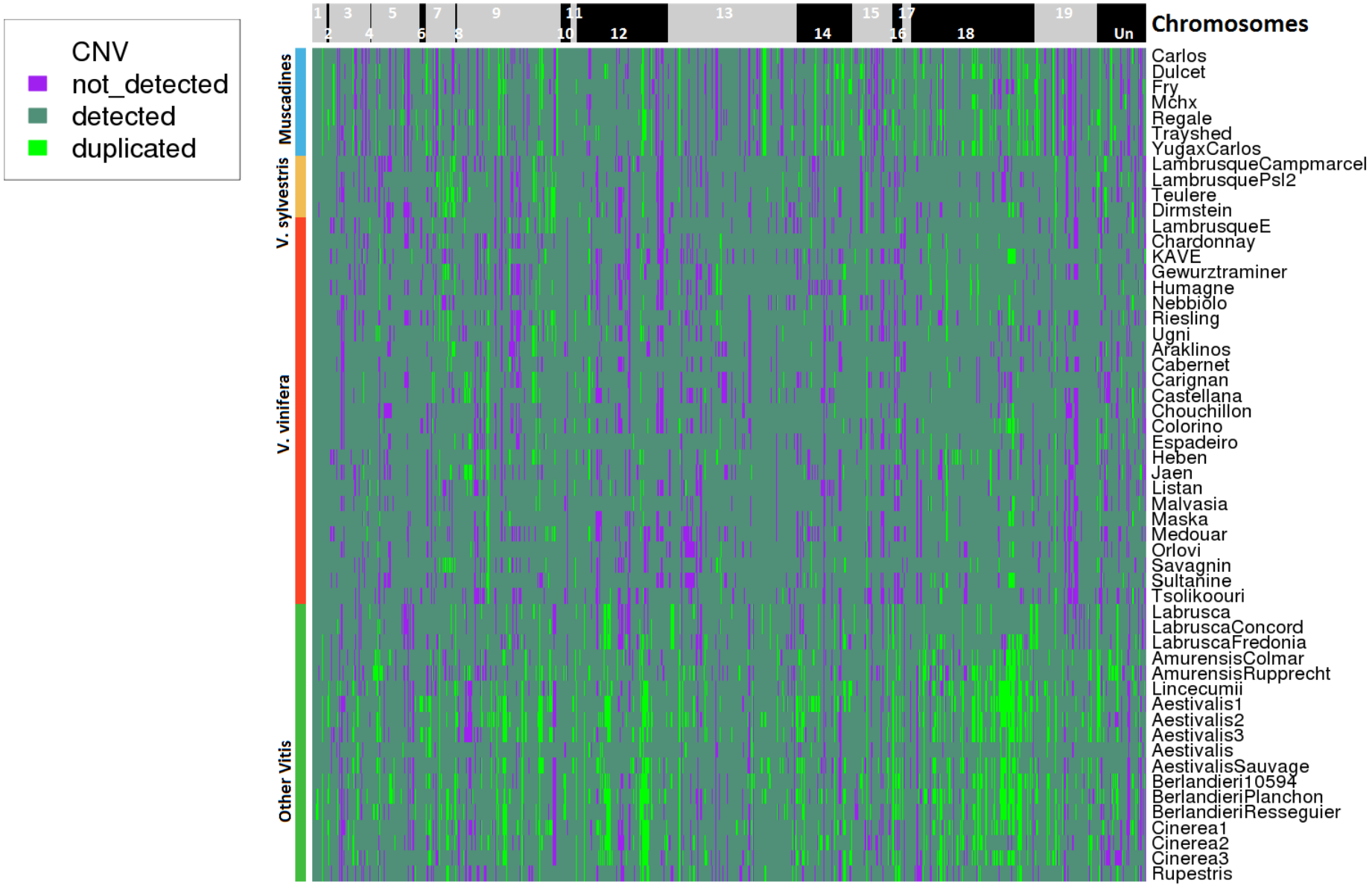



**Figure 4: *NBS-gene* conservation among 54 Vitis genomes.** The CNV category for each of the 829 *NBS-genes* (y-axis, the first *NBS-gene* in chromosome one on the left, the last gene in chromosome Unknown on the right) was color-coded for each of the 54 genotypes (x-axis). For this heatmap, the "partially_detected" and "potentially_duplicated" genes are colored like the "detected" and "duplicated" genes, respectively.

By considering the "not detected" genes as absent and all the others as present with at least one copy, we can assess the presence / absence variations into more details. Surprisingly, the American *Vitis aestivalis* genotype harboured the least number of putatively absent genes (37) and *Vitis vinifera* cv Tsolikoouri the highest (120). On average, the percentage of present genes increased from the *Muscadinia* group (87.6%), the *Vitis sylvestris* group (88.4%), the *Vitis vinifera* group (90.5%) to the *Vitis* species group (92.2%). By considering the *NBS-genes* individually, 353 were found to be present in all the studied genotypes and up to 712 genes in more than 40 genotypes. Twenty-four genes (2.9%; 13 pseudo- or partial genes) were absent in more than half of the genotypes and represented 18% of the "not detected" calls. Thus, strong disparities of presence / absence rate were identified at the gene scale.

Among the present genes, some were classified as "potentially duplicated" or "duplicated", *i.e.* at least partially duplicated. On average, the percentage of duplicated genes increased from the *Vitis vinifera* group (2.6%), the *Vitis sylvestris* group (3.2%), the *Muscadinia* group (7.8%) to the *Vitis species* group (11.7%). By considering the *NBS-genes* individually, 32 genes (3.9%; 19 pseudo- or partial genes) were duplicated in more than a third of the genotypes and represented 27% of the "potentially duplicated" or "duplicated" calls.

Among the present genes, most were classified as "partially detected". On average, the percentage of "partially detected" genes increased from the *Vitis sylvestris* group (25.7%), the *Vitis* species group (26.9%), the *Vitis vinifera* group (28.0%) to the *Muscadinia* group (35.6%). By considering the *NBS-genes* individually, 165 genes (19.9%; 113 pseudo- or partial genes) were "partially detected" in more than half of the genotypes and represented 45% of the "partially detected" calls. Thus, most of the variations were observed in this category and represented either truncated or hemizygous genes, i.e. no complete loss nor complete duplication. Moreover, blocks of neighbour genes and clusters seemed to be enriched in "partially detected" and "not detected" calls, suggesting that rearrangements regarding NBS-genes are not occurring randomly in the grapevine genome.

To investigate whether whole clusters (genes and intergenic regions) were putatively absent or duplicated in the Vitis genus, a second CNV analysis was performed at the cluster scale using a segmentation approach (Supplementary data 8 and 9). The majority of the clusters belonged to the "partially detected" or the "detected" categories and only 5, 5 and 2 calls were classified as "not detected", "potentially duplicated" and "duplicated", respectively. The three potentially absent clusters are all composed of *CC NBS-genes* on chromosome 13 of *Vitis vinifera* cv Sultanine and Orlovi. Regarding the potentially duplicated clusters, no clear pattern is noticeable. The "partially detected" clusters represented 29% of the calls and especially 23 clusters were classified as "partially detected" in more than half of the genotypes. Fourteen and six of these clusters are composed of *CC* and *NBS* genes, respectively, which is about 25% of each cluster type harbouring rearrangements in most of the studied genotypes. Thus, we can hypothesize that the rearrangements of *NBS-genes* occur mostly at the gene scale within the clusters and that they affect mostly clusters of *CC* and *NBS-genes*.

Based on the logratio values obtained for each exon, the dendrogram of the 54 genotypes was build (Figure 5). This tree shown a similar organization to the phylogenetic network produced by Wan and coworkers (2013). Among the four groups of species, the *Muscadinia* group, like the *Vitis sylvestris* and the American/Asian *Vitis* groups appeared well defined. Only the *Vitis vinifera* genotypes do not group together and are split by the *Vitis sylvestris* genotypes. One of these groups is composed of five genotypes (Carignan, Castellana, Medouar, Listan, Malvasia), whereas the twenty others form the second group. A similar dendrogram was obtained based on the logratio calculated for each cluster (supplementary data 10). Finally, the evolution of the *NBS-gene* family seemed to follow the same pattern than the evolution acting for speciation in the *Vitis* genus.

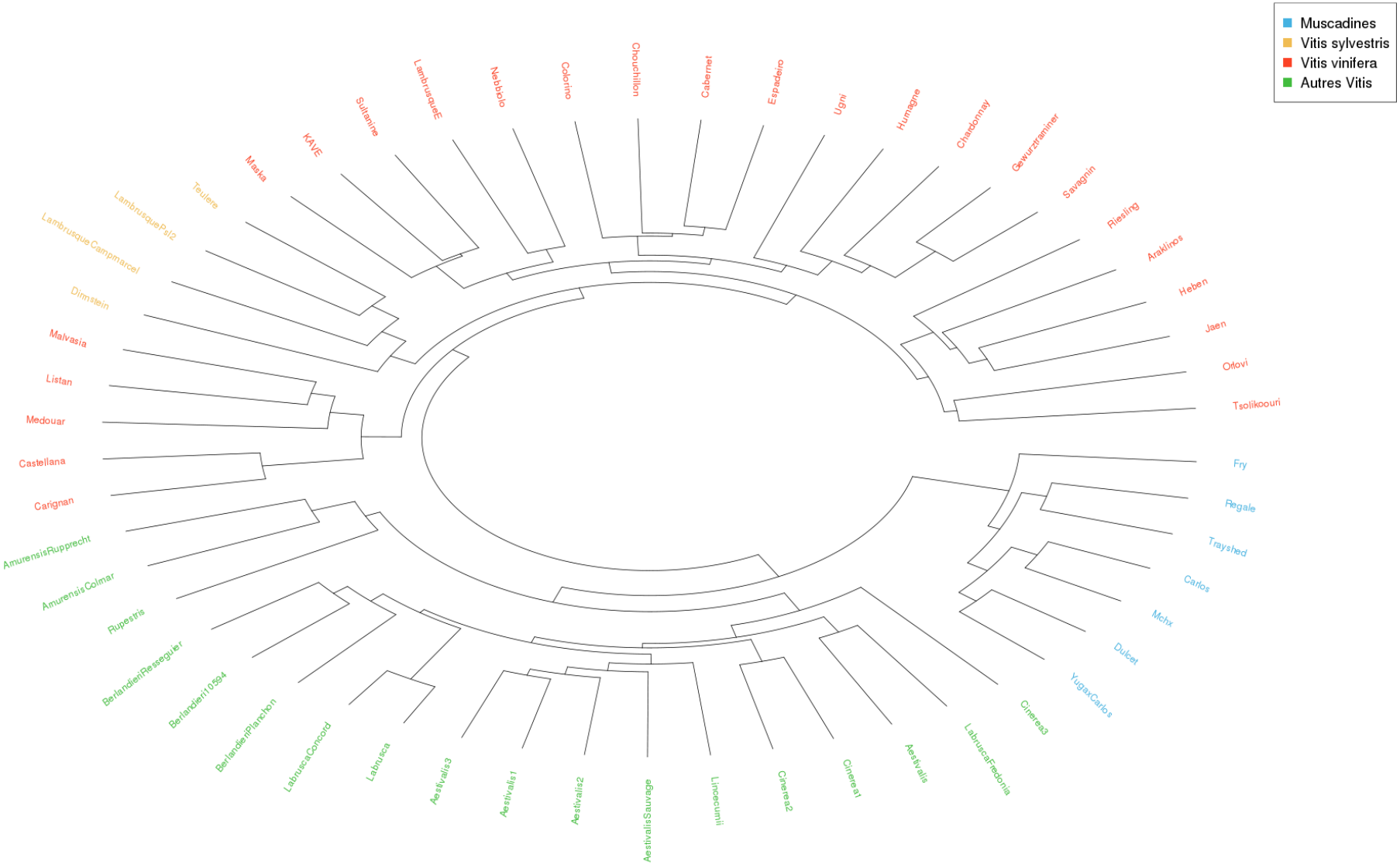



**Figure 5: Dendrogram of the 54 genomes based on the CNV analysis of the *NBS-genes*.** The dendrogram was produced from the logratio values of exons of all the *NBS-genes.*

# Discussion

## Organization of the *NBS-genes* in the grapevine reference genome

This study reports the first annotation of the *NBS-genes* in the grapevine reference genome PN40024 12x.2, and their conservation among the *Vitis* genus. Based on the identification of the NBS domain, we annotated 829 *NBS-genes*, which is twice the number detected by Yang and coworkers (2008) (314) and by Malacarne and coworkers (2012) (391). However, among our 829 genes, only 450 were complete genes (54%), which is more consistent with their results. It seems that our study is more exhaustive because of our detection of pseudogenes and partial genes. Their detection is helpful for studying the organization of this gene family in clusters and their evolution.

Based on the identification of Pfam domains, 216 *NBS-genes* with a CC domain, 105 with a TIR domain, 116 without N-terminal domain and 13 with the RPW8 domain were detected. As the CC domain is not highly conserved at the sequence level but rather at the structural level, some *NBS-genes* without identified N-terminal domain may also be composed of a CC domain. The ratio in favour of the *CC NBS-genes* is consistent with other *NBS* annotation managed for the first assembly of the PN40024 reference genome (Yang *et al.*, 2008) or other grapevine genome (Malacarne *et al.*, 2012). The enrichment in CC-NBS rather than TIR-NBS is also consistent with the trend observed in most plants (Shao *et al.*, 2016; Yue *et al.*, 2012; Jacob *et al.*, 2013).

In order to define the NBS-gene clusters, we used criteria reported in the literature based on physical distance and the number of non *NBS-genes* between two *NBS-genes* (Yang *et al.*, 2008; Malacarne *et al.*, 2012; Richly *et al.*, 2002; Meyers *et al.*, 2003). Most of the clusters defined this way harboured sequence homology of successive genes within the borders of the clusters. This observation was found to be consistent with our hypothesis regarding *NBS-genes* undergoing intra-cluster rearrangements. However, few exceptions of similarity extended over neighbour clusters were identified, which pointed out that these criteria may not be optimal in the case of grapevine reference genome. The use of a maximal physical distance of 320 kb without a cutoff on the number of *non-NBS* genes between two *NBS-genes*, for example, would result in the clustering of the genes at the end of chromosome 3, while they were separated in three clusters using the previous criteria. However, with this new parameter, five clusters composed of either *CC* or *TIR NBS-genes* would form two larger clusters mixing the two classes.

The organization of *NBS-genes* into clusters of highly similar genes was demonstrated in many plant genomes (Andolfo *et al.*, 2013; Christie *et al.*, 2016; Zheng *et al.*, 2016; Yang and Wang, 2015; Andersen *et al.*, 2016; Yang *et al.*, 2008). In the tomato genome, the clusters do not only include NBS but also receptor-like protein and receptor-like kinase genes, which would favour coordinated transcription (Andolfo *et al.*, 2013). In the eucalyptus genome as well, expression hotspots could be identified within the NBS clusters (Christie *et al.*, 2016), suggesting an adaptive advantage for the plant to keep the *NBS-genes* close to each other in order to optimize transcription.

**Evolution patterns for *TIR* and *CC NBS-genes***

In the grapevine reference genome, we identified a clear tendency for genes to be highly conserved within clusters. Some large inter-cluster rearrangements could be identified involving *TIR NBS-genes* specifically. This tendency to local rather than large segmental duplications was already described in sorghum, barley, tomato, poplar and grapevine (Yang and Wang, 2015; Andersen *et al.*, 2016; Yang *et al.*, 2008; Andolfo *et al.*, 2013). But, for the first time to our knowledge in grapevine, we assessed the diversity of the *NBS-gene* family among the *Vitis* genus through a CNV analysis. Most of the identified variations were classified as "partially detected", meaning they probably imply gene truncation or hemizygosity. Large duplicated or deleted regions could not be found in any group of *Vitis* species. Friedman and Baker (2007) reviewed the main processes of evolution of this gene family and reported that clusters of tandemly duplicated genes permit sequence exchanges via recombinatorial mispairing and generate high haplotypic diversity. They also pointed out the putative role of epigenetic modifications in the instability of these regions. Together, these results depict an intra-cluster evolution of the *NBS-gene* family not only in the reference genome but among the *Vitis* genus that implies tandem duplications and shuffling of the different functional domains to generate diversity.

The *TIR* and *CC NBS-genes* are organized differently in the grapevine reference genome. On the one hand, the *TIR NBS-genes* are mostly located on chromosome 18, show high intra- and inter-cluster similarities and some segmental duplications between clusters. On the other hand, the *CC NBS-genes* are distributed on all the 19 chromosomes, show high similarities within clusters but lower between clusters and no large segmental duplication was identified. These characteristics suggest, indeed, a different evolution history between these two classes of *NBS-genes* as previously mentioned in other plants (Chen *et al.*, 2010; Yang *et al.*, 2008). Two hypotheses are stated. Yue and coworkers (2012) suggested a higher selective constraint for *TIR* or a higher diversifying selection for *CC NBS-genes* that could explain their different levels of conservation. Another hypothesis was stated by Shao and coworkers (2016) who explained that the *CC NBS-gene* initial expansion is older in the angiosperm lineage than the *TIR*, which number remains stable during the same period. However, after the Cretaceous-Palaeogene boundary, both classes of *NBS-genes* underwent an expansion due to an explosion of the pathogen pressure.

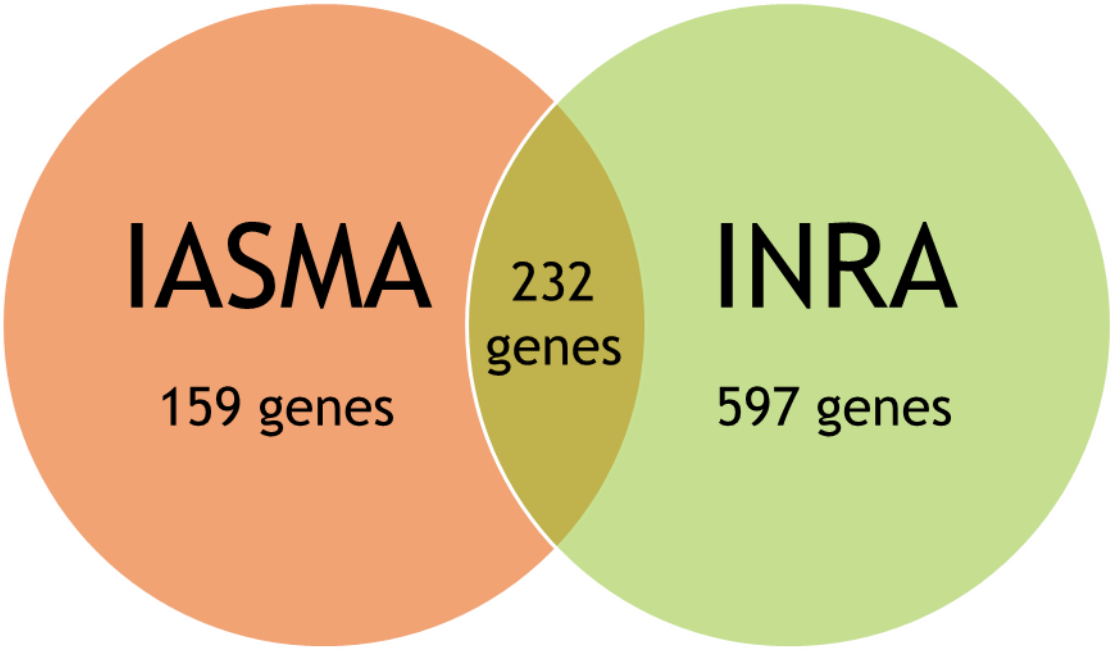
Finally, we report the first evaluation of the diversity of the *NBS-gene* family in the *Vitis* genus through a CNV analysis. This approach added a new dimension to the hypotheses stated on the grapevine reference genome regarding the mechanisms driving the evolution dynamics of this large resistance gene family. Thus, by comparing the copy number in several *Vitis* genotypes compared to the reference genome, we confirmed a clear tendency for an intra-cluster evolution of the *NBS-gene* family that implies tandem duplications and shuffling of the different functional domains. These local rearrangements seemed prevalent for CC-NBS clusters, suggesting a higher diversifying selection for this class. However, this relative estimation of the copy number is biased in favour of genotypes that are close to the reference genome and may not be as accurate for the *Muscadinia* genotypes, for example. To circumvent this limitation and to discover new *NBS-genes*, *de novo* assembly of the non-mapping reads, RenSeq or even whole genome assembly using long could be valuable alternatives (Zmieńko *et al.*, 2014; Young *et al.*, 2016; Andolfo *et al.*, 2014).

# Supplementary data

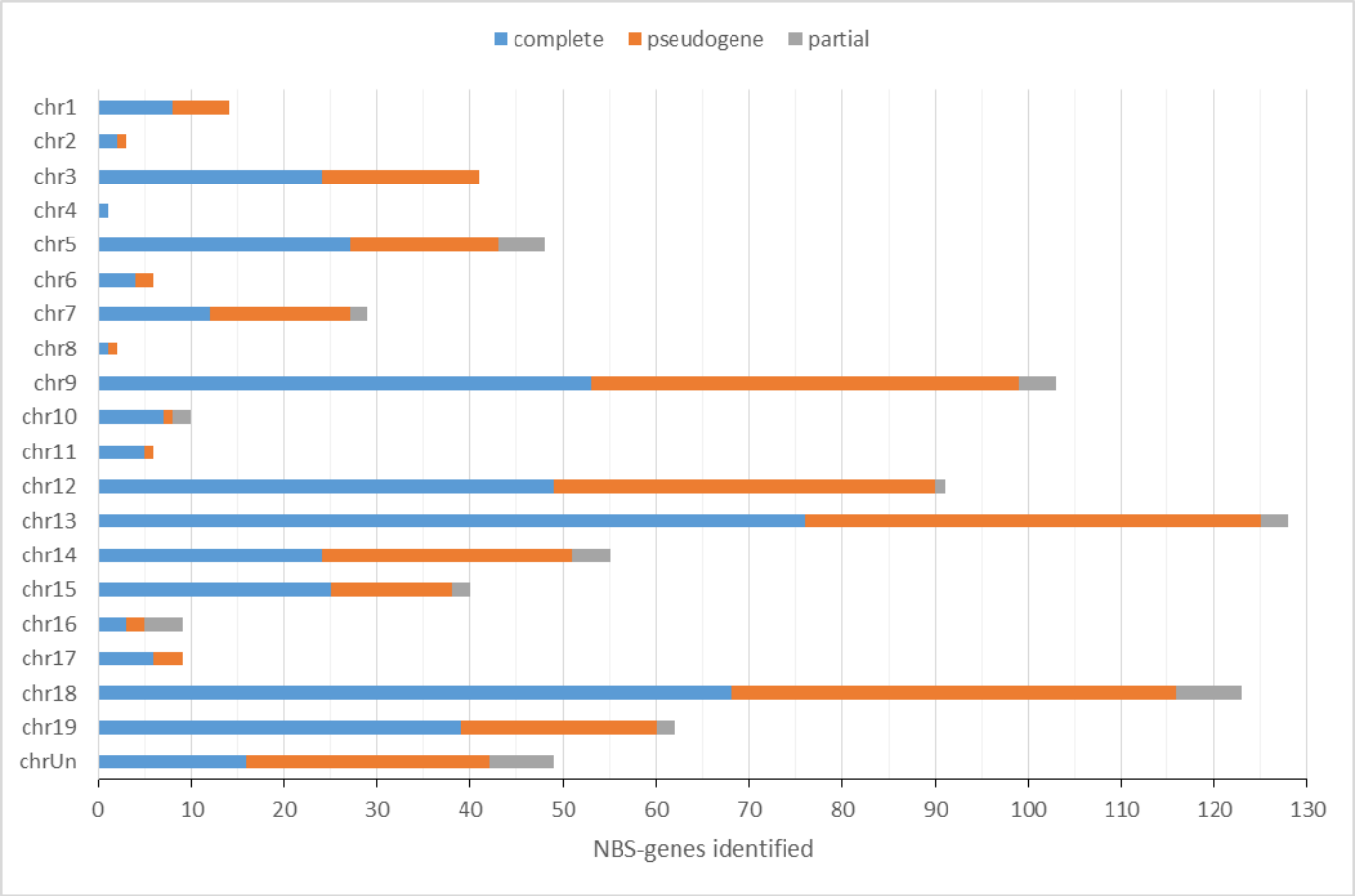
**Supplementary data 1: Studied genotypes.**

| Accession | Species | Project |
|---|---|---|
| Aestivalis-1 | *V. aestivalis* | GrapeReSeq |
| Aestivalis-2 | *V. aestivalis* | GrapeReSeq |
| Aestivalis-3 | *V. aestivalis* | GrapeReSeq |
| Vitis_aestivalis | *V. aestivalis* | GrapeReSeq |
| Vitis_aestivalis_sauvage | *V. aestivalis* | GrapeReSeq |
| Berlandieri-10594 | *V. berlandieri* | GrapeReSeq |
| Planchon | *V. berlandieri* | GrapeReSeq |
| Resseguier | *V. berlandieri* | GrapeReSeq |
| Cinerea-1 | *V. cinerea* | GrapeReSeq |
| Cinerea-2 | *V. cinerea* | GrapeReSeq |
| Cinerea-3 | *V. cinerea* | GrapeReSeq |
| Concord | *V. labrusca* | GrapeReSeq |
| Fredonia | *V. labrusca* | GrapeReSeq |
| Labrusca | *V. labrusca* | GrapeReSeq |
| Vitis_lincecumii | *V. lincecumii* | GrapeReSeq |
| Lambrusque_Campmarcel | *V. sylvestris* | GrapeReSeq |
| Lambrusque_psl2 | *V. sylvestris* | GrapeReSeq |
| Sylvestris-Dirmstein-male | *V. sylvestris* | GrapeReSeq |
| Teulere_sauvage | *V. sylvestris* | GrapeReSeq |
| Araklinos | *V. vinifera* | GrapeReSeq |
| Cabernet_franc | *V. vinifera* | GrapeReSeq |
| Carignan_noir | *V. vinifera* | GrapeReSeq |
| Castellana-blanca | *V. vinifera* | GrapeReSeq |
| Chouchillon | *V. vinifera* | GrapeReSeq |
| Colorino | *V. vinifera* | GrapeReSeq |
| Espadeiro-tinto | *V. vinifera* | GrapeReSeq |
| Heben | *V. vinifera* | GrapeReSeq |
| Jaen | *V. vinifera* | GrapeReSeq |
| Lambrusque_e | *V. vinifera* | GrapeReSeq |
| Listan_Prieto | *V. vinifera* | GrapeReSeq |
| Malvasia-di-sardegna | *V. vinifera* | GrapeReSeq |
| Maska | *V. vinifera* | GrapeReSeq |
| Medouar | *V. vinifera* | GrapeReSeq |
| Orlovi-nogti | *V. vinifera* | GrapeReSeq |
| PN40024 | *V. vinifera* | GrapeReSeq |
| Savagnin | *V. vinifera* | GrapeReSeq |
| Sultanine | *V. vinifera* | GrapeReSeq |
| Tsolikoouri | *V. vinifera* | GrapeReSeq |
| Riesling_49 | *V. vinifera* | HealthyGrape |
| Gewurztraminer_643 | *V. vinifera* | HealthyGrape |
| Nebbiolo | *V. vinifera* | HealthyGrape |

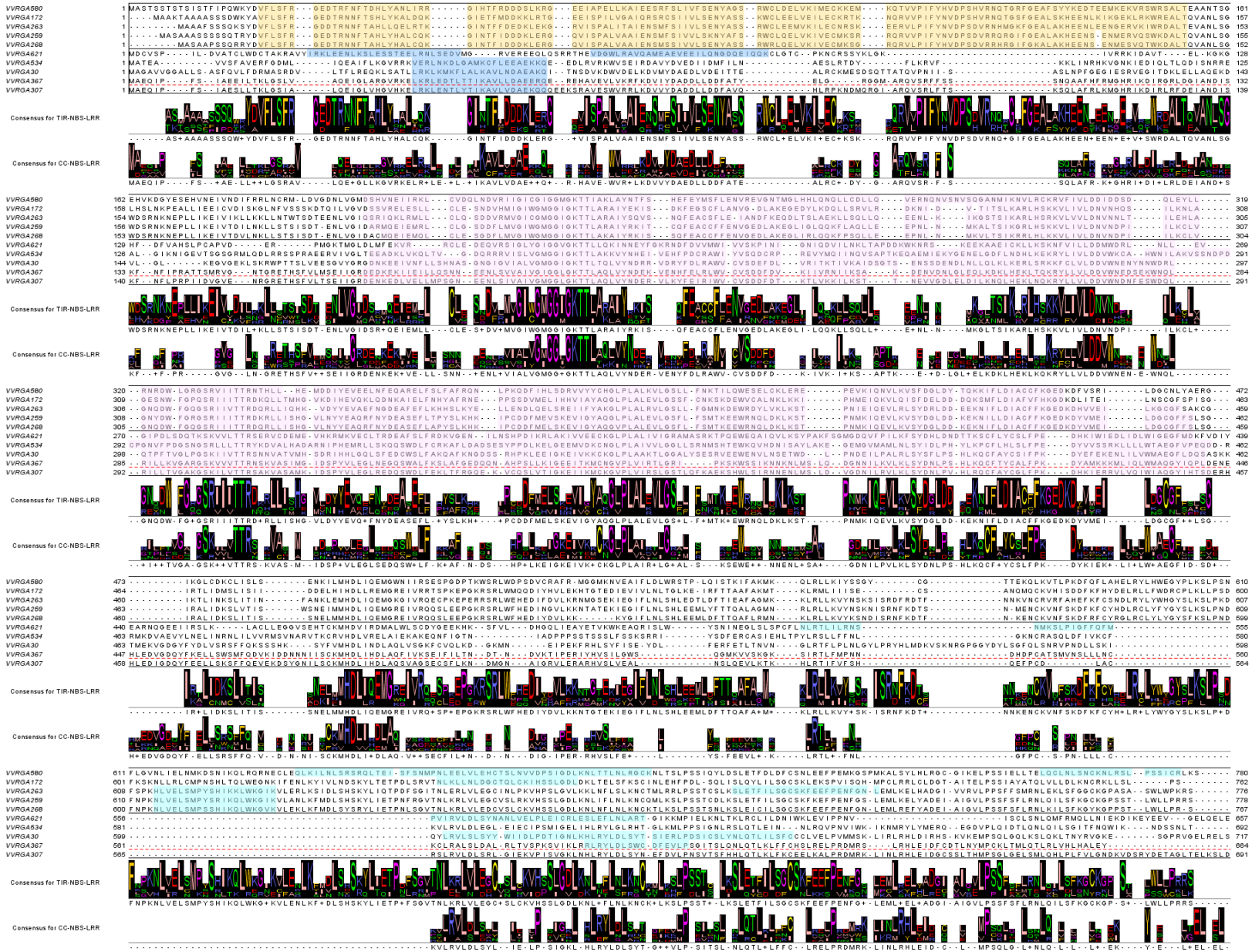
| | | |
|---|---|---|
| Chardonnay_95 | *V. vinifera* | HealthyGrape |
| Humagne | *V. vinifera* | HealthyGrape |
| KAVE | *V. vinifera* | HealthyGrape |
| Ugni_blanc_479 | *V. vinifera* | HealthyGrape |
| Vitis_amurensis_Colmar | *V. amurensis* | HealthyGrape |
| Vitis_amurensis_Rupprecht | *V. amurensis* | HealthyGrape |
| Carlos | *M. rotundifolia* | Muscares |
| Dulcet | *M. rotundifolia* | Muscares |
| Fry | *M. rotundifolia* | Muscares |
| Mchx | *M. rotundifolia* | Muscares |
| Regale | *M. rotundifolia* | Muscares |
| Trayshed | *M. rotundifolia* | Muscares |
| YugaxCarlos | *M. rotundifolia* | Muscares |



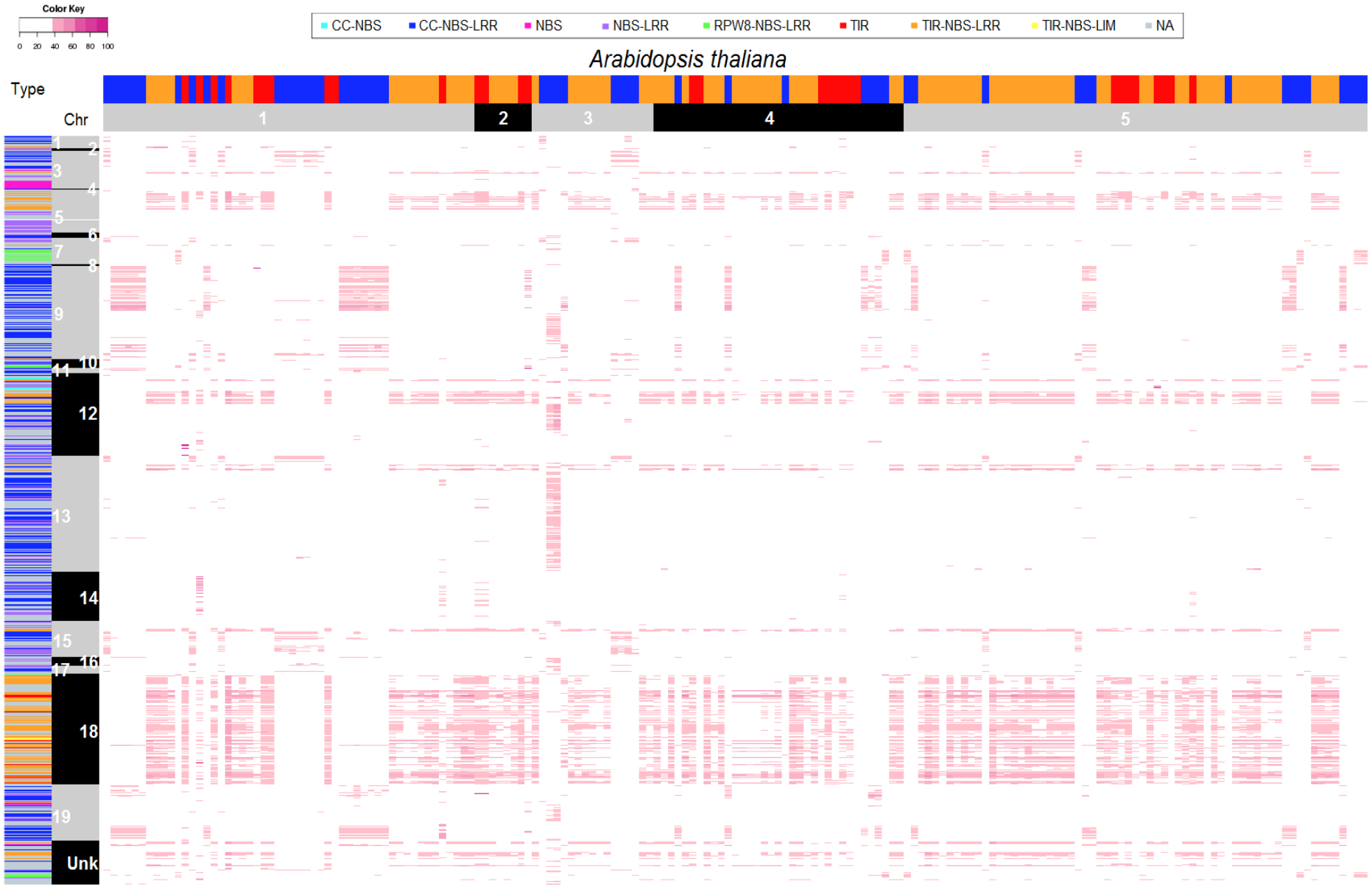

**Supplementary data 2: *NBS-gene* comparison with Malacarne et al. (2012) study.** With both identity and overlapping threshold are set at 80%, 332 *IASMA* genes had strong alignments with 232 *INRA* genes. A great part of the *INRA* genes that matched *IASMA* genes were complete genes and were classified as "CC-NBS-LRR".




**Supplementary data 3: *NBS-gene* repartition across the grapevine reference genome.** Even if the *NBS-genes* are spread along the 19 chromosomes of PN40024, the distribution varies depending on the chromosome.




**Supplementary data 4: Sequence conservation of the different domains that composed the *NBS-genes*.** The coding sequence of the TIR domain was underline in yellow, the Coil domain in dark blue, the NBS domain in pink and the LRR domains in light blue. The consensus sequences represent the conservation rate of each amino-acid respectively for the 5 *TIR-NBS-LRR* and for the 5 *CC-NBS-LRR genes*.




**Supplementary data 5: Identity matrix of the *NBS-gene* in the *Arabidopsis* and the grapevine genomes.** The 829 PN40024 *NBS-genes* are represented on the X axis and the 177 *Arabidopsis NBS-genes* on the Y axis. The color is correlated with the similarity between genes, i.e. pink mean 40% of similarity between the two genes, purple 100%. The chromosomes and gene sub-families are indicated for each genome.

**Supplementary data 6: *NBS-gene* detection through CNV analysis.** The average percentage for each CNV category were calculated for each species group.

| | not detected | partially detected | detected | potentially duplicated | duplicated |
|---|---|---|---|---|---|
| ***Muscadinia*** | 12,34 | 35,57 | 44,29 | 6,77 | 1,03 |
| ***V. sylvestris*** | 11,58 | 25,69 | 59,50 | 2,96 | 0,27 |
| ***V. vinifera*** | 9,47 | 26,91 | 61,09 | 2,32 | 0,20 |
| **Other *Vitis*** | 7,84 | 26,91 | 53,53 | 10,15 | 1,57 |

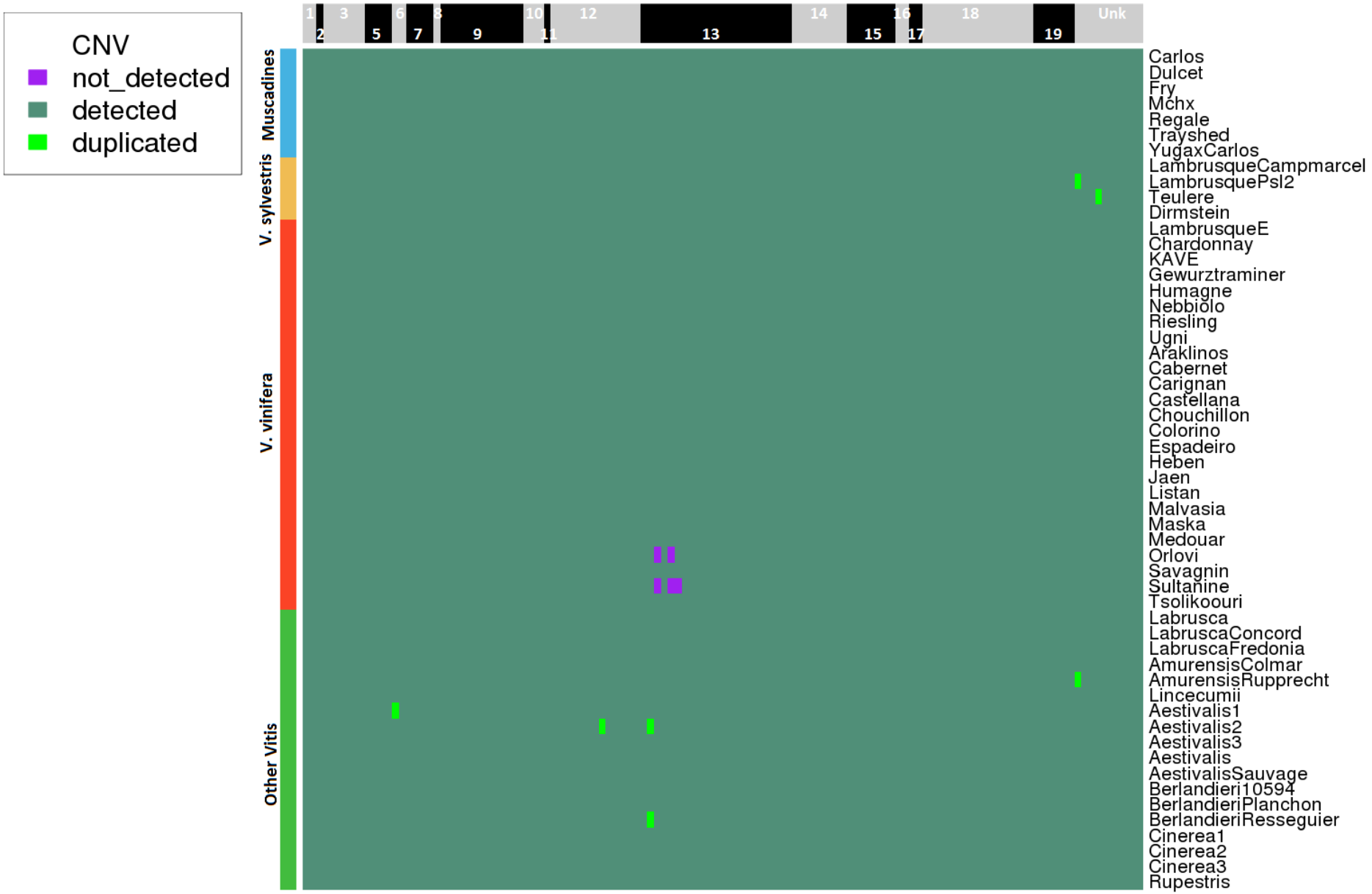
**Supplementary data 7: Proportion of each CNV category in the 54 *Vitis* genomes.** The "present" percentage is the sum of the other percentages, "not_detected" excluded. Blue genotypes are the *Muscadinia* species, yellow the *Vitis sylvestris*, red the *Vitis vinifera* and green the American/Asian Vitis species.

| | not detected (in %) | partially detected (in %) | detected (in %) | potentially duplicated (in %) | duplicated (in %) | present (in %) |
|---|---|---|---|---|---|---|
| Carlos | 13.75 | 36.91 | 43.55 | 5.07 | 0.72 | 86.25 |
| Dulcet | 12.55 | 32.93 | 45.11 | 8.32 | 1.09 | 87.45 |
| Fry | 12.55 | 38.36 | 41.86 | 5.91 | 1.33 | 87.45 |
| Mchx | 12.91 | 37.27 | 44.27 | 4.58 | 0.97 | 87.09 |
| Regale | 10.98 | 34.38 | 45.11 | 8.56 | 0.97 | 89.02 |
| Trayshed | 10.98 | 35.46 | 44.75 | 7.84 | 0.97 | 89.02 |
| YugaxCarlos | 12.67 | 33.66 | 45.36 | 7.12 | 1.21 | 87.33 |
| Lambrusque_Campmarcel | 12.30 | 29.67 | 56.09 | 1.69 | 0.24 | 87.70 |
| Lambrusque_psl2 | 9.77 | 20.87 | 66.47 | 2.65 | 0.24 | 90.23 |
| Teulere | 11.82 | 25.45 | 58.75 | 3.74 | 0.24 | 88.18 |
| Dirmstein | 12.42 | 26.78 | 56.69 | 3.74 | 0.36 | 87.58 |
| Lambrusque_e | 10.37 | 30.88 | 56.33 | 2.17 | 0.24 | 89.63 |
| Chardonnay | 8.93 | 28.83 | 60.19 | 1.81 | 0.24 | 91.07 |
| KAVE | 11.82 | 35.22 | 49.58 | 3.14 | 0.24 | 88.18 |
| Gewurztraminer | 9.41 | 26.18 | 61.76 | 2.41 | 0.24 | 90.59 |
| Humagne | 10.74 | 30.28 | 56.09 | 2.77 | 0.12 | 89.26 |
| Nebbiolo | 8.93 | 33.05 | 55.85 | 1.93 | 0.24 | 91.07 |
| Riesling | 11.70 | 33.05 | 51.39 | 3.62 | 0.24 | 88.30 |
| Ugni | 9.53 | 30.40 | 56.69 | 3.14 | 0.24 | 90.47 |
| Araklinos | 8.69 | 25.21 | 64.54 | 1.09 | 0.48 | 91.31 |
| Cabernet | 8.44 | 25.81 | 64.17 | 1.33 | 0.24 | 91.56 |
| Carignan | 917 | 32.21 | 55.73 | 2.77 | 0.12 | 90.83 |
| Castellana | 10.86 | 22.68 | 63.45 | 2.65 | 0.36 | 89.14 |
| Chouchillon | 9.29 | 27.38 | 62.12 | 1.09 | 0.12 | 90.71 |
| Colorino | 8.93 | 25.69 | 61.40 | 3.74 | 0.24 | 91.07 |
| Espadeiro | 7.60 | 30.04 | 60.07 | 2.17 | 0.12 | 92.40 |
| Heben | 8.69 | 28.59 | 57.90 | 4.58 | 0.24 | 91.31 |
| Jaen | 11.94 | 28.71 | 55.49 | 3.74 | 0.12 | 88.06 |
| Listan | 10.25 | 24.61 | 62.36 | 2.65 | 0.12 | 89.75 |
| Malvasia | 7.60 | 26.78 | 63.45 | 1.93 | 0.24 | 92.40 |
| Maska | 7.48 | 25.57 | 64.66 | 2.17 | 0.12 | 92.52 |
| Medouar | 14.35 | 25.57 | 58.02 | 1.69 | 0.36 | 85.65 |
| Orlovi | 9.89 | 27.26 | 61.64 | 0.97 | 0.24 | 90.11 |
| Savagnin | 6.63 | 22.80 | 67.55 | 2.90 | 0.12 | 93.37 |
| Sultanine | 10.49 | 25.33 | 61.40 | 2.65 | 0.12 | 89.51 |
| Tsolikoouri | 14.48 | 27.50 | 56.69 | 1.21 | 0.12 | 85.52 |
| Labrusca | 8.32 | 26.66 | 59.11 | 5.67 | 0.24 | 91.68 |
| Labrusca Concord | 7.84 | 22.44 | 64.29 | 5.07 | 0.36 | 92.16 |
| Labrusca Fredonia | 10.01 | 28.71 | 53.32 | 6.88 | 1.09 | 89.99 |
| Amurensis_Colmar | 5.31 | 27.02 | 58.14 | 7.60 | 1.93 | 94.69 |

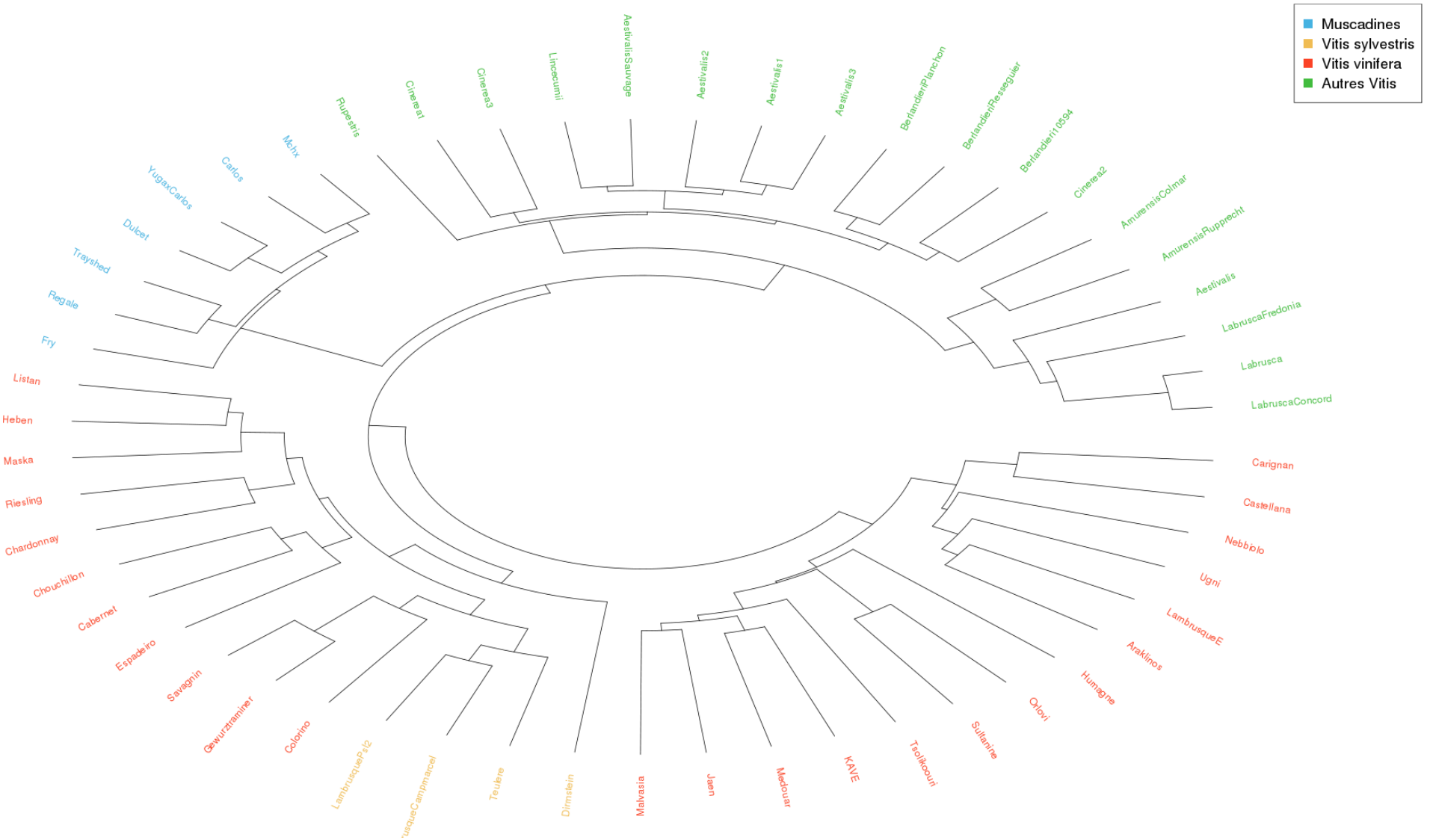
| | | | | | | |
|---|---|---|---|---|---|---|
| Amurensis_Rupprecht | 10.49 | 28.23 | 48.85 | 9.89 | 2.53 | 89.51 |
| Lincecumii | 7.84 | 29.92 | 48.85 | 11.94 | 1.45 | 92.16 |
| Aestivalis1 | 6.27 | 25.09 | 51.03 | 14.60 | 3.02 | 93.73 |
| Aestivalis2 | 6.03 | 23.88 | 52.71 | 14.11 | 3.26 | 93.97 |
| Aestivalis3 | 6.51 | 25.57 | 52.71 | 13.03 | 2.17 | 93.49 |
| Aestivalis | 4.46 | 26.54 | 64.90 | 4.10 | 0.00 | 95.54 |
| Aestivalis_sauvage | 7.60 | 26.42 | 53.20 | 11.58 | 1.21 | 92.40 |
| Berlandieri 10594 | 6.88 | 29.07 | 50.18 | 12.30 | 1.57 | 93.12 |
| Berlandieri Planchon | 7.12 | 22.20 | 53.44 | 15.32 | 1.93 | 92.88 |
| Berlandieri Resseguier | 7.12 | 27.62 | 51.63 | 11.70 | 1.93 | 92.88 |
| Cinerea1 | 9.41 | 27.99 | 49.10 | 11.70 | 1.81 | 90.59 |
| Cinerea2 | 9.65 | 30.52 | 48.01 | 11.10 | 072 | 90.35 |
| Cinerea3 | 9.17 | 26.66 | 51.51 | 11.10 | 1.57 | 90.83 |
| Rupestris | 11.10 | 29.92 | 52.59 | 4.95 | 1.45 | 88.90 |

**Supplementary data 8: NBS-gene clusters detection with the segmentation analysis.** The average of the different level of cluster detection were calculated for each genome and report in the following table for each group of *Vitis*.

| | not detected | partially detected | detected | potentially duplicated | duplicated |
|---|---|---|---|---|---|
| ***Muscadinia*** | 0,00 | 47,66 | 52,34 | 0,00 | 0,00 |
| ***V. sylvestris*** | 0,00 | 30,53 | 69,06 | 0,41 | 0,00 |
| ***V. vinifera*** | 0,16 | 28,95 | 70,89 | 0,00 | 0,00 |
| **Other *Vitis*** | 0,00 | 21,95 | 77,82 | 0,14 | 0,09 |




**Supplementary data 9: NBS-gene cluster conservation among the 54 Vitis genomes.** The CNV result of the 122 *NBS-genes* clusters is represented with a color-scaled. For this heatmap, the "partially_detected" and "potentially_duplicated" genes are labelled like "detected" and "duplicated" respectively.




**Supplementary data 10: Dendrogram of the 54 genomes based on the segmentation analysis of the clusters of *NBS-genes*.** The dendrogram was produced from the logratio values of the segments corresponding to the clusters defined previously (the segment start corresponding to the first gene start and the segment end corresponding to the last gene end).

# References

**Akita, M. and Valkonen, J.P.T.** (2002) A novel gene family in moss (Physcomitrella patens) shows sequence homology and a phylogenetic relationship with the TIR-NBS class of plant disease resistance genes. *J. Mol. Evol.*, **55**, 595–605.

**Allen, J.E. and Salzberg, S.L.** (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–603.

**Andersen, E.J., Ali, S., Reese, R.N., Yen, Y. and Neupane, S.** (2016) Diversity and Evolution of Disease Resistance Genes in Barley ( Hordeum vulgare L .). , 99–108.

**Andolfo, G., Jupe, F., Witek, K., et al.** (2014) Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC Plant Biol.*, **14**, 120.

**Andolfo, G., Sanseverino, W., Rombauts, S., Peer, Y. Van de, Bradeen, J.M., Carputo, D., Frusciante, L. and Ercolano, M.R.** (2013) Overview of tomato (Solanum lycopersicum) candidate pathogen recognition genes reveals important Solanum R locus dynamics. *New Phytol.*, **197**, 223–237.

**Bai, J., Pennill, L.A., Ning, J., et al.** (2002) Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. *Genome Res.*, **12**, 1871–1884.

**Bairoch, A. and Boeckmann, B.** (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **19 Suppl**, 2247–2249.

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

**Belkhadir, Y., Subramaniam, R. and Dangl, J.L.** (2004) Plant disease resistance protein signaling: NBS-LRR proteins and their partners. *Curr. Opin. Plant Biol.*, **7**, 391–399.

**Bent, A.F. and Mackey, D.** (2007) Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions. *Annu. Rev. Phytopathol.*, **45**, 399–436.

**Bergelson, J., Kreitman, M., Stahl, E. a and Tian, D.** (2001) Evolutionary dynamics of plant R-genes. *Science*, **292**, 2281–2285.

**Blanc, S., Wiedemann-Merdinoglu, S., Dumas, V., Mestre, P. and Merdinoglu, D.** (2012) A reference genetic map of Muscadinia rotundifolia and identification of Ren5, a new major locus for resistance to grapevine powdery mildew. *Theor. Appl. Genet.*, **125**, 1663–1675.

**Bomblies, K. and Weigel, D.** (2010) Arabidopsis and relatives as models for the study of genetic and genomic incompatibilities. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **365**, 1815–1823.

**Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L.** (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 1–9.

**Chen, Q., Han, Z., Jiang, H., Tian, D. and Yang, S.** (2010) Strong positive selection drives rapid diversification of R-Genes in arabidopsis relatives. *J. Mol. Evol.*, **70**, 137–148.

**Christie, N., Tobias, P.A., Naidoo, S. and Külheim, C.** (2016) The Eucalyptus grandis NBS-LRR Gene Family: Physical Clustering and Expression Hotspots. *Front. Plant Sci.*, **6**, 1238.

**Dangl, J.L. and Jones, J.D.** (2001) Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.

**Eddy, S.R.** (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.

**Finn, R.D., Bateman, A., Clements, J., et al.** (2014) Pfam: The protein families database. *Nucleic Acids Res.*, **42**.

**Friedman, A.R. and Baker, B.J.** (2007) The evolution of resistance genes in multi-protein plant resistance systems. *Curr. Opin. Genet. Dev.*, **17**, 493–499.

**Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W. and Dangl, J.L.**

(1995) Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. *Science*, **269**, 843–6.

**Guo, Y.-L., Fitz, J., Schneeberger, K., Ossowski, S., Cao, J. and Weigel, D.** (2011) Genome-Wide Comparison of Nucleotide-Binding Site-Leucine-Rich Repeat-Encoding Genes in Arabidopsis. *Plant Physiol.*, **157**, 757–769.

**Henk, A.D., Warren, R.F. and Innes, R.W.** (1999) A new Ac-like transposon of arabidopsis is associated with a deletion of the RPS5 disease resistance gene. *Genetics*, **151**, 1581–1589.

**Jacob, F., Vernaldi, S. and Maekawa, T.** (2013) Evolution and conservation of plant NLR functions. *Front. Immunol.*, **4**.

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

**Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A.** (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–45.

**Larkin, M., Blackshields, G., Brown, N., et al.** (2007) ClustalW and ClustalX version 2. *Bioinformatics*, **23**, 2947–2948.

**Leister, D.** (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet.*, **20**, 116–122.

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

**Lin, F. and Chen, X.M.** (2007) Genetics and molecular mapping of genes for race-specific all-stage resistance and non-race-specific high-temperature adult-plant resistance to stripe rust in spring wheat cultivar Alpowa. *Theor. Appl. Genet.*, **114**, 1277–1287.

**Liu, J., Liu, X., Dai, L. and Wang, G.** (2007) Recent Progress in Elucidating the Structure , Function and Evolution of Disease Resistance Genes in Plants. *J. Genet.*, **34**, 765–776.

**Luo, S., Zhang, Y., Hu, Q., Chen, J., Li, K., Lu, C., Liu, H., Wang, W. and Kuang, H.** (2012) Dynamic Nucleotide-Binding Site and Leucine-Rich Repeat-Encoding Genes in the Grass Family. *Plant Physiol.*, **159**, 197–210.

**Lupas, A., Dyke, M. Van and Stock, J.** (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–4.

**Luz, L. de A., Silva, M.C.C., Ferreira, R. da S., Santana, L.A., Silva-Lucca, R.A., Mentele, R., Oliva, M.L.V., Paiva, P.M.G. and Coelho, L.C.B.B.** (2013) Structural characterization of coagulant Moringa oleifera Lectin and its effect on hemostatic parameters. *Int. J. Biol. Macromol.*, **58**, 31–38.

**Malacarne, G., Perazzolli, M., Cestaro, A., Sterck, L., Fontana, P., Peer, Y. van de, Viola, R., Velasco, R. and Salamini, F.** (2012) Deconstruction of the (paleo)polyploid grapevine genome based on the analysis of transposition events involving NBS resistance genes. *PLoS One*, **7**.

**Martin, G.B., Bogdanove, A.J. and Sessa, G.** (2003) Understanding the functions of plant disease resistance proteins. *Annu. Rev. Plant Biol.*, **54**, 23–61.

**McDowell, J.M. and Simon, S.A.** (2006) Recent insights into R gene evolution. *Mol. Plant Pathol.*, **7**, 437–448.

**McHale, L., Tan, X., Koehl, P. and Michelmore, R.W.** (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biol.*, **7**, 212.

**Meyers, B.C., Dickerman, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sobral, B.W. and Young, N.D.** (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.*, **20**, 317–332.

**Meyers, B.C., Kozik, A., Griego, A., Kuang, H. and Michelmore, R.W.** (2003) Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis. *Plant Cell Online*, **15**, 809–834.

**Michelmore, R.W. and Meyers, B.C.** (1998) Clusters of resistance genes in plants evolve by divergent

selection and a birth-and-death process. *Genome Res.*, **8**, 1113–1130.

**Mondragón-Palomino, M., Meyers, B.C., Michelmore, R.W. and Gaut, B.S.** (2002) Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. *Genome Res.*, **12**, 1305–1315.

**Porter, B.W., Paidi, M., Ming, R., Alam, M., Nishijima, W.T. and Zhu, Y.J.** (2009) Genome-wide analysis of Carica papaya reveals a small NBS resistance gene family. *Mol. Genet. Genomics*, **281**, 609–626.

**Riaz, S., Tenscher, a C., Smith, B.P., Ng, D. a and Walker, M. a** (2008) Use of SSR markers to assess identity, pedigree, and diversity of cultivated muscadine grapes. *J. Am. Soc. Hortic. Sci.*, **133**, 559–568.

**Richly, E., Kurth, J. and Leister, D.** (2002) Mode of amplification and reorganization of resistance genes during recent Arabidopsis thaliana evolution. *Mol. Biol. Evol.*, **19**, 76–84.

**Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B.** (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

**Schiex, T., Moisan, A. and Rouzé, P.** (2001) EugÉne: An eukaryotic gene finder that combines several sources of evidence. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 111–125.

**Seshan, V.E. and Olshen, A.B.** (2014) DNAcopy : A Package for Analyzing DNA Copy Data. *Bioconductor Vignette*, 1–7.

**Shao, Z.-Q., Xue, J.-Y., Wu, P., Zhang, Y.-M., Wu, Y., Hang, Y.-Y., Wang, B. and Chen, J.-Q.** (2016) Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiol.*, pp.01487.2015.

**Shen, J., Araki, H., Chen, L., Chen, J.Q. and Tian, D.** (2006) Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in Arabidopsis thaliana. *Genetics*, **172**, 1243–1250.

**Sievers, F., Wilm, A., Dineen, D., et al.** (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

**Springer, N.M., Ying, K., Fu, Y., et al.** (2009) Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genet.*, **5**, e1000734.

**Velasco, R., Zharkikh, A., Troggio, M., et al.** (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, **2**.

**Vivier, M.A. and Pretorius, I.S.** (2002) Genetically tailored grapevines for the wine industry. *Trends Biotechnol.*, **20**, 472–478.

**Wan, H., Yuan, W., Ye, Q., et al.** (2012) Analysis of TIR- and non-TIR-NBS-LRR disease resistance gene analogous in pepper: characterization, genetic variation, functional divergence and expression patterns. *BMC Genomics*, **13**, 502.

**Wan, Y., Schwaninger, H.R., Baldo, A.M., Labate, J.A., Zhong, G.-Y. and Simon, C.J.** (2013) A phylogenetic analysis of the grape genus (Vitis L.) reveals broad reticulation and concurrent diversification during neogene and quaternary climate change. *BMC Evol. Biol.*, **13**, 141.

**Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J.** (2009) Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

**Wickham, H.** (2009) *ggplot2*,.

**Wu, T.D. and Nacu, S.** (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

**Yang, S., Feng, Z., Zhang, X., Jiang, K., Jin, X., Hang, Y., Chen, J.Q. and Tian, D.** (2006) Genome-wide investigation on the genetic variations of rice disease resistance genes. *Plant Mol. Biol.*, **62**, 181–193.

**Yang, S., Zhang, X., Yue, J.-X., Tian, D. and Chen, J.-Q.** (2008) Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics*, **280**, 187–198.

**Yang, X. and Wang, J.** (2015) Genome-wide analysis of NBS-LRR genes in sorghum genome revealed several events contributing to NBS-LRR gene evolution in grass species. *Evol. Bioinforma.*, **12**, 9–21.

**Young, N.D., Zhou, P. and Silverstein, K.A.T.** (2016) Exploring structural variants in environmentally sensitive gene families. *Curr. Opin. Plant Biol.*, **30**, 19–24.

**Yuan, B., Zhai, C., Wang, W., Zeng, X., Xu, X., Hu, H., Lin, F., Wang, L. and Pan, Q.** (2011) The Pik-p resistance to Magnaporthe oryzae in rice is mediated by a pair of closely linked CC-NBS-LRR genes. *Theor. Appl. Genet.*, **122**, 1017–1028.

**Yue, J.X., Meyers, B.C., Chen, J.Q., Tian, D. and Yang, S.** (2012) Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes. *New Phytol.*, **193**, 1049–1063.

**Zhao, S., Guo, Y., Sheng, Q. and Shyr, Y.** (2014) Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinformatics*, **15**, P16.

**Zheng, F., Wu, H., Zhang, R., Li, S., He, W., Wong, F.-L., Li, G., Zhao, S. and Lam, H.-M.** (2016) Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family. *BMC Genomics*, **17**, 402.

**Zmieńko, A., Samelak, A., Kozłowski, P. and Figlerowicz, M.** (2014) Copy number polymorphism in plant genomes. *Theor. Appl. Genet.*, **127**, 1–18.

**Zohary, D. and Hopf, M.** (1973) Domestication of Pulses in the Old World. *Science (80-. ).*, **182**, 887–894.

# Partie 4 : Génomique comparative et transcriptomique de la famille des gènes *STS* chez la vigne (*Vitis vinifera*) et ses proches parents

Dans le cadre de cette thèse, une étude spécifique des gènes codant pour les stilbènes synthases (STS) est détaillée dans le manuscrit présenté ci-dessous. L'expression des gènes codant pour les gènes *STS*, ainsi que leur présence / absence dans différents génomes de vignes ont été étudiées en utilisant, entre autres, les outils et méthodes décrites précédemment.

Nous avons montré que, dans les conditions expérimentales étudiées, certains gènes ne sont pas ou peu exprimés tandis que d'autres ont une expression plus importante. La plupart des gènes étant faiblement ou pas exprimés sont les gènes partiels et les pseudo-gènes. Les gènes *STS* sont principalement exprimés dans les baies mûres et les feuilles subissant un stress biotique et leur amplitude d'expression est variable, bien que les profils d'expression soient similaires. Comme les stilbènes sont des molécules de défense et de réponse au stress, sachant qu'un stress oxydatif apparaît lors de la maturation des baies, ces résultats correspondent au fait que la production de stilbènes soit favorisée dans ces conditions.

L'étude des variations structurales des gènes *STS* dans les différents génotypes choisis montre que l'amplification de la famille de gènes précédemment observée dans le génome de référence n'est pas un artéfact et que de manière générale il y a une tendance à la conservation de la majorité des gènes *STS* dans les autres génomes étudiés. En observant le comportement de certains gènes, il est aussi possible de remarquer que certains ont tendance à être plus souvent dupliqués ou délétés que d'autres mais que les gènes les plus conservés sont généralement ceux les plus exprimés.

# Comparative genomics and transcriptomics of the stilbene synthase gene family in grapevine (*Vitis vinifera*) and its wild relatives

Gautier Arista, Guillaume Barnabé, Lauriane Renault, Sophie Blanc, Philippe Hugueney, Camille Rustenholz

Université de Strasbourg, INRA, SVQV UMR-A 1131, F-68000 Colmar, France.

## Abstract

Stilbenes such as resveratrol are considered as the main phytoalexins in *Vitis* species and are synthesized through the activity of the enzyme stilbene synthase (STS). Analysis of the grapevine reference genome has revealed an unusually large family of 48 *STS* genes with redundant function. This raises the question, as to whether the characteristics of this gene family are conserved in other cultivated grapevine varieties and wild *Vitis* species. In this work, we have used large-scale transcriptomics and Copy Number Variation (CNV) analyses in order to get a better picture of the *STS* gene family throughout the *Vitis* genus. We propose that *STS* genes clusters were formed through several rounds of segmental duplications in tandem in the grapevine genome and we show that this family is subjected to a fine-tuned regulation relying on a combination of both highly and low-expressed genes. Finally, we propose that a small number of highly expressed *STS* genes, conserved throughout the *Vitis* genus, may constitute the core set of genes necessary for stilbene biosynthesis.

**Key words:** grapevine, stilbene synthases, transcriptomics, RNA-Seq, CNV

# Introduction

Stilbenes are phenylpropanoids characterized by a 1,2-diphenylethylene backbone, which are produced by a number of unrelated plants, including peanut, pine trees, Japanese knotweed and grapevine (reviewed in Chong *et al.*, 2009). The enzyme stilbene synthase (STS) is characteristic of stilbene-producing plants and catalyses the biosynthesis of the stilbene backbone. In most stilbene-producing plants, *STS* genes form small families of closely related paralogs. However, grapevine is a noteworthy exception, as its genome has been shown to contain a large family of *STS* genes. A first analysis suggested that more than 20 copies of *STS* genes could be found in grapevine (Sparvoli *et al.*, 1994). In previous studies, manual annotation of the *STS* gene family in the grapevine reference genome (Jaillon *et al.*, 2007) identified 48 *STS* genes organised in two clusters on chromosomes 10 and 16 (Parage *et al.*, 2012, Vannozzi *et al.*, 2012). Furthermore, functional analysis of a selection of 10 *STS* genes showed that they indeed encoded enzymes with stilbene synthase activity, suggesting that the 32 full-length *STS* genes present in the grapevine genome may actually have the same function. Finally, Parage *et al.* (2012) showed that this family was subjected to negative selective pressures, resulting in a tendency to keep all genes functional. These results raised the question, as to why such a large family of redundant genes is conserved in grapevine. One hypothesis is that this high number of *STS* genes may allow fine-tuning of their expression and specific responses to a large array of biotic or abiotic stresses, conferring grapevine a better adaptability to its environment. A first insight was provided by Vannozzi *et al.* (2012) who used a combination of microarray and RNA-Seq approaches to analyse the expression of *STS* genes. Most *STS* genes showed no or low constitutive expression, except in roots, where they were constitutively expressed. In addition, *STS* genes could be separated in different groups regarding their expression patterns in response to various stresses. Three different stresses were tested: infection with downy mildew (biotic stress), wounding and UV treatment (abiotic stresses). Three groups of *STS* genes were defined based on their response intensity following UV stress: low response, intermediate response and high response. The regulation of the highly responsive group was shown to be different from the other groups, suggesting a transcriptional subfunctionalization among *STS* genes in grapevine. Regulation of *STS* gene expression has been shown to rely, at least partly, on the two R2R3-MYB–type transcription factors MYB14 and MYB15, which regulate the stilbene biosynthetic pathway in response to stresses and during berry development (Höll *et al.*, 2013).

However, most of our current knowledge of the *STS* gene family relies on studies performed on the reference PN40024 genome (Parage *et al.*, 2012, Vannozzi *et al.*, 2012). This raises the question, as to whether the characteristics of this gene family are conserved in other cultivated grapevine varieties and wild *Vitis* species. In order to broaden our knowledge of the *STS* family and its regulation, we decided to extend the analysis of the *STS* gene family both at the structural and the transcriptomic levels. Firstly, we analysed further the structure of the *STS* cluster on chromosome 16 in the PN40024 genome, which allowed us to propose hypotheses regarding the evolution of this large cluster of highly similar genes. Secondly, we extended the analysis of the regulation of the *STS* family by using a large-scale transcriptomics analysis involving all RNA-Seq data available in public databases. Finally, we performed a Copy Number Variation (CNV) analysis of *STS* genes in genomes of a selection of cultivated varieties and wild grapevine species, in order to get a better picture of this unusually large family of genes with identical function throughout the *Vitis* genus.

# Material & Methods

## Phylogeny

The phylogeny analysis was performed with the cDNA sequences of *STS* genes from PN40024, using the annotation published previously (Parage *et al.*, 2012; Vannozzi *et al.*, 2012). The multiple alignment was performed using MUSCLE version 3.8.31 (Edgar, 2004) with default parameters. The phylogenetic tree, available in supplementary data, was constructed with Mega version 7.0.18 (Kumar *et al.*, 2016) using the maximum likelihood method, a bootstrap of 100, a Jukes-Cantor model, the rates among sites being gamma distributed in five categories and using all sites for the missing data treatment. The tree was then visualised and coloured using iTOL version 3 (itol.embl.de; Letunic & Bork, 2007). A matrix of identity percentage between STS gene pairs was computed from a multiple alignment of the cDNA sequences performed with Clustal Omega (Sievers *et al.*, 2011). This matrix was used to identify the groups of similar STS genes.

## Expression profile analysis

Data from 73 RNA-Seq experiments (Da Silva *et al.*, 2013; Jones *et al.*, 2014; Perazzolli *et al.*, 2012; Ramos *et al.*, 2014; Sweetman, Wong, Ford, & Drew, 2012; Venturini *et al.*, 2013) were retrieved from NCBI SRA database (http://www.ncbi.nlm.nih.gov/sra). From the SRA files, fastq-dump, available in the SRA Toolkit Package version 2.3.4 (http://www.ncbi.nlm.nih.gov/books/NBK158900) was used to generate the fastq files. Alignments were performed on the 12x.0 version of the grapevine reference genome PN40024 (Jaillon et al., 2007; https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences) using GSNAP version 2013-11-27 (Wu and Nacu, 2010) with the parameters: -B 4, -N 1, -n 3, --nofails and the right quality protocol, Illumina or Sanger, depending on the experiment. The output alignments were parsed to keep the best, unique and paired, if paired-end reads, using a homemade Perl script. Using the manual annotation performed in Parage *et al.* (2012) and Vannozzi *et al.* (2012), the number of fragments aligned on each gene was calculated using htseq-count, from the HTSeq suite version 0.6.0 (Anders *et al.*, 2015), with the parameters: -m intersection-nonempty and -s no. The FPKM (Fragments Per Kilo base of exon per Million fragments mapped) were calculated using a homemade R script (R version 3.0.2; R Development Core Team, 2016).

The experiments were grouped into five different categories regarding the sample and experimental conditions: flowers, young berries, ripe berries, leaves and leaves under biotic stress. To simplify the data, the mean FPKM value was calculated per category resulting in 5 expression values per gene. The expression profiles of the *STS* genes on these five categories were shown on a heatmap build with R.

**Differential expression analysis**

The RNA-Seq datasets published by Perazzolli *et al.* (2012) and Jones *et al.* (2014) were selected for a differential expression analysis as they used triplicates in their experimental design to study the impact of both downy and powdery mildew infections on gene expression in grapevine. The number of fragments aligned on the V1 gene annotation available in the Grape Genome Database hosted at CRIBI (http://genomes.cribi.unipd.it/grape/; Vitulo et al., 2014) and on the manual annotation of *STS* genes performed in Parage *et al.* (2012) and Vannozzi *et al.* (2012) was counted using htseq-count as previously mentioned. The package edgeR (version 3.4.2; Robinson, McCarthy, & Smyth, 2009) was used to perform the differential expression analysis. Genes with more than 0.2 reads per million mapped reads were considered as expressed and used for further analysis. The method "TMM" provided in the package was used to normalise the data and the glmQLFTest model was run to access the differentially genes. For the downy mildew analysis (Perazzolli *et al.*, 2012), the datasets from the non-inoculated leaves were used as reference. For the powdery mildew analysis (Jones *et al.*, 2014), the datasets from leaves 12 hours after inoculation were used as reference. Only genes with False Discovery Rate (FDR) lower than 5% were considered differentially expressed.

**Copy Number Variation analysis**

A total of 56 genotypes were analysed and compared to the reference genome: genotypes from Europe (25 *Vitis vinifera* spp *vinifera* and 4 *Vitis vinifera* spp *sylvestris*), Asia (2 *Vitis amurensis*), North America (16 genotypes of *Vitis cinerea*, *aestivalis*, *labrusca*, *rupestris* and 7 *Muscadinia rontundifolia*) and 2 hybrids between *Vitis* and *Muscadinia*. About 10x sequencing depth of paired-end Illumina GAII (2x100bp) or HiSeq2000 (2x150bp) for each genotype was used. Alignments were performed on the 12x.0 version of the grapevine reference genome PN40024 using GSNAP as previously described except with –N 1 parameter. A parsing step using a homemade perl script used the edit distance to keep the best and unique alignments.

The threshold of edit distance compared to the reference genome was adjusted according to the genotype species: more than 98% identity for PN40024, 95% for the *Vitis vinifera* and *Vitis sylvestris* genotypes, 90% for the other *Vitis* species and 85% for the *Muscadinia* genotypes. The sequencing depth of every base and the median depth for the whole chromosome were computed. Sliding windows of 1 kb separated by 200bp were defined and a normalized logratio was calculated for each of them as followed:

$$LogRatio = \log_2\left(\frac{(Average\ depth\ /\ Median\ depth\ for\ the\ chromosome)_{genome}}{(Average\ depth\ /\ Median\ depth\ for\ the\ chromosome)_{reference}}\right)$$

With the R package DNAcopy version 1.36.0 (Seshan & Olshen, 2014), a segmentation approach was undertaken to identify genomic regions harbouring significantly homogeneous logratios. A histogram of the logratios distribution of the segments was drawn and five normal distributions were determined to cover this distribution based on the method adapted from Springer *et al.* (2009). The highly negative, slightly negative, null, slightly positive and highly positive logratios were represented by a normal distribution each. This approach allowed the classification of the segments in five categories: not detected, partially detected, detected, potentially duplicated and duplicated. Adjacent segments of the same category were grouped together. The focus was made on the clusters of *STS* genes (chr10:14210000..14310000 and chr16:16230000..16720000). The *STS* genes were classified depending on the segments to which they belonged. If at least 75% of the gene length was in the duplicated category, the gene was considered as duplicated. If at least 75% of the gene length was potentially duplicated (with adding the duplication category too), the gene was considered as potentially duplicated. Then, a gene was considered as detected if the addition of the segment lengths from the detected category and the duplicated categories was covering at least 75% of the gene. Finally, if the addition of the segment lengths of the previous categories and the partially detected was at least 25%, the gene was considered as partially detected. If not, it was considered as not detected.

# Results

## Organisation of the *STS* gene family in the PN40024 reference genome

The *STS* gene family is greatly amplified in grapevine compared to other stilbene-producing plants and is organised in two clusters on chromosome 10 and 16. In order to better describe and understand this family in the reference genome, a phylogeny was first constructed. This new phylogeny differs from that presented in Parage *et al.* (2012) and Vannozzi *et al.* (2012) by the fact that partial and pseudogenes were included and that nucleotide instead of protein sequences were used. However, the same nomenclature (from *VvSTS1* to *VvSTS48*) was used. In the phylogenetic tree shown in Supplementary Figure 1, three groups can be identified, in agreement with the groups identified previously by Vannozzi *et al.* (2012). Based on this phylogeny, colours were assigned to the *STS* genes according to their similarity (Figure 1), in order to better visualize relationships among the clusters, which may reflect the history of the *STS* gene family. Genes with the same colour share a minimum of 95% identity, except for *VvSTS14* and *VvSTS26*, these two genes being partial and are not overlapping. The overall minimum identity among *STS* genes is 66% (excluding *VvSTS14* and *VvSTS26*). Based on colour similarity, it is possible to identify multiple redundant patterns in the large *STS* genes cluster, represented by blocks of genes with high similarity, probably resulting from recent duplication events. Four groups of genes numbered 1 to 4 were identified, each of them made of 2 to 4 blocks of genes (Figure 1). For example, the block 4a spanning *VvSTS39* to *VvSTS42* and the block 4b (*VvSTS43*-*VvSTS46*) are made of similar genes and are organized the same way, suggesting that these 2 blocks may result from a duplication event. However, the blocks that are worth highlighting are 1a, 1b and 1c, corresponding to *STS* genes *VvSTS10*-*VvSTS14*, *VvSTS15*-*VvSTS19* and *VvSTS21*-*VvSTS24*. *VvSTS11* is a pseudogene with at least 80% identity with *VvSTS12*, *16* and *22*, even if it has not been assigned to any group.

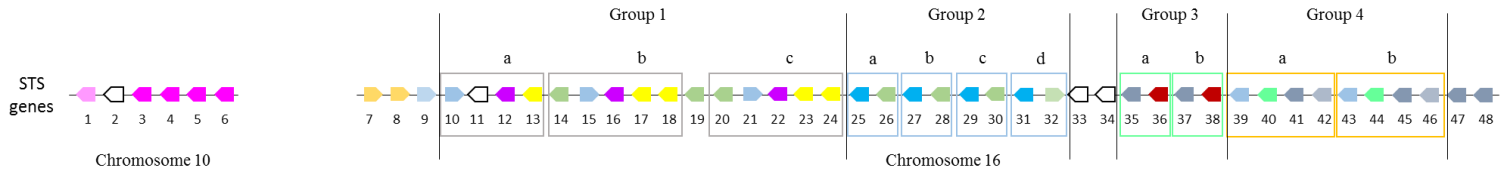



**Figure 1: Gene organization and similarity patterns in the PN40024 of *STS* gene clusters**
The percentage of identity was calculated based on nucleotide sequences and colours were assigned to *STS* genes according to their similarity. Four groups of genes numbered 1 to 4 were identified, each of them made of 2 to 4 blocks of genes numbered a to d. Genes with the same colour share at least 95% identity.

Figure 2 shows that the conservation level between these three subgroups is very high even in the intergenic regions, which suggests that these duplications may be recent. Notably, groups 1b and 1c seem more similar between each other than with group 1a. This suggests that the duplication between groups 1b and 1c is more recent. Similarly, Figure 2 shows that the conservation level between these four subgroups is very high even in the intergenic regions, which suggests that these duplications may be recent. Notably, groups 2a, 2b and 2c seem more similar between each other than with group 2d. For groups 3 and 4, despite the duplications that may have happened, the intergenic regions do not seem to share high levels of conservation. So we hypothesized that the duplications involving groups 3 and 4 may be older than the ones involving groups 1 and 2. Even though *VvSTS33* and *VvSTS34* have not been assigned to any group, they are very similar to *VvSTS36* and *VvSTS38* and to *VvSTS10*, *15*, *21*, *39* and *43*, respectively. For cluster on chromosome 10, no clear duplication pattern could be identified (data not shown).

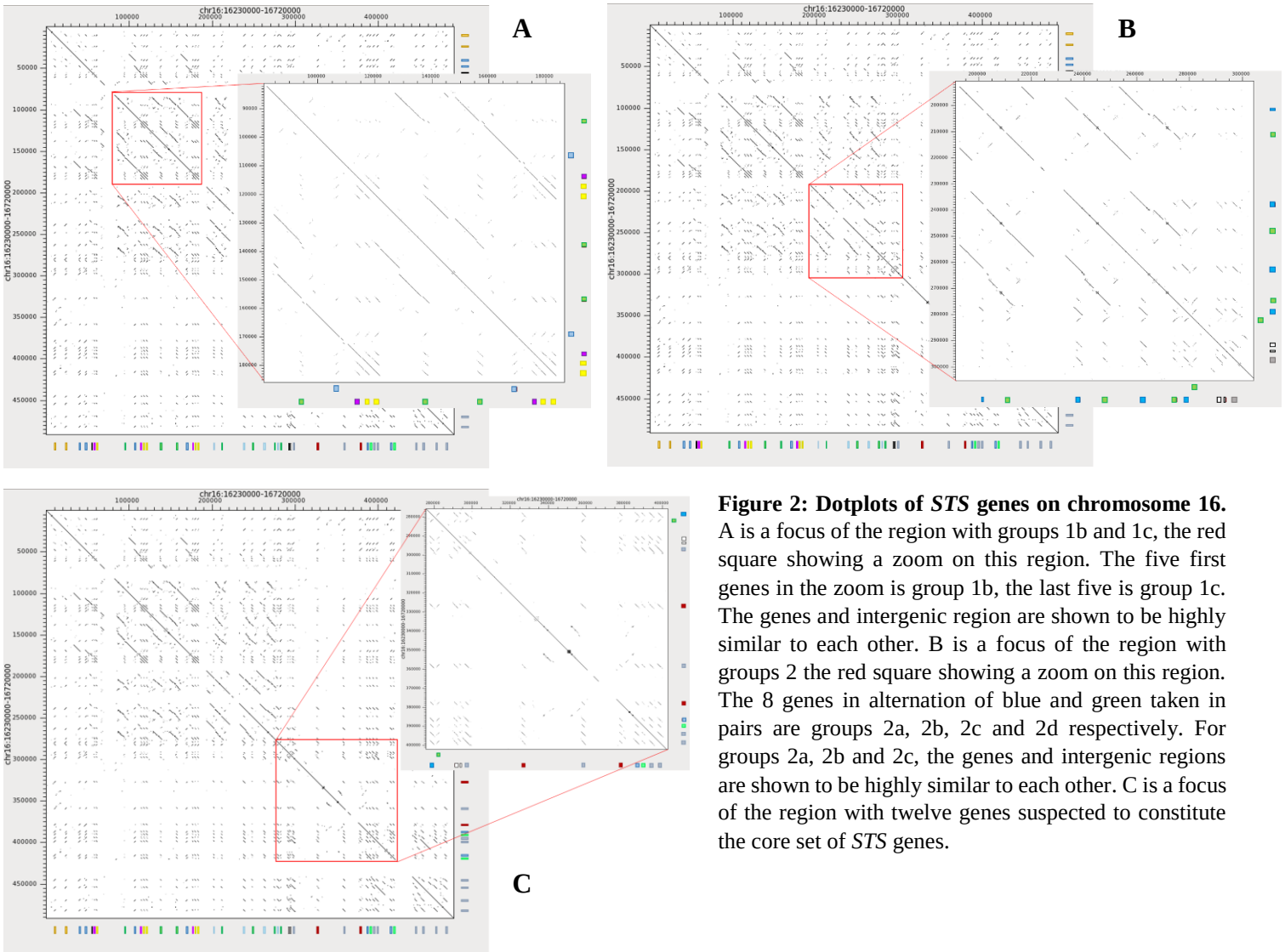



**Figure 2: Dotplots of *STS* genes on chromosome 16.** A is a focus of the region with groups 1b and 1c, the red square showing a zoom on this region. The five first genes in the zoom is group 1b, the last five is group 1c. The genes and intergenic region are shown to be highly similar to each other. B is a focus of the region with groups 2 the red square showing a zoom on this region. The 8 genes in alternation of blue and green taken in pairs are groups 2a, 2b, 2c and 2d respectively. For groups 2a, 2b and 2c, the genes and intergenic regions are shown to be highly similar to each other. C is a focus of the region with twelve genes suspected to constitute the core set of *STS* genes.

## Large-scale expression analysis of the *STS* gene family

Vannozzi *et al.* (2012) have provided a first analysis of *STS* gene expression focused on biotic and abiotic stress conditions. In order to get a broader insight into the regulation of this family, expression of *STS* genes was analysed in 81 publicly-available RNA-Seq experiments (Supplementary table 1), including, flowers, berries and leaves subjected to biotic stress. Global expression analysis showed that some *STS* genes had low or no expression in all conditions tested, whereas others were highly expressed (Figure 3). In particular, 69% of the partial and pseudogenes showed low expression or were considered as not expressed. Among non-expressed genes, only *VvSTS6* is complete and *VvSTS26* is partial. For expressed genes, highest expression rates were observed in ripe berries and leaves subjected to biotic stress. Even though expression profiles were globally similar among the *STS* gene family, the amplitude of expression varies from one gene to another. *VvSTS7*, *9*, *10*, *18*, *31*, *36* and *48* had the highest expression. These results suggest that even if *STS* genes expression profiles are similar among the tested conditions, with increased expression in ripe berries and under biotic stress, a small number of *STS* genes are highly expressed and account for most of the expression of the family.

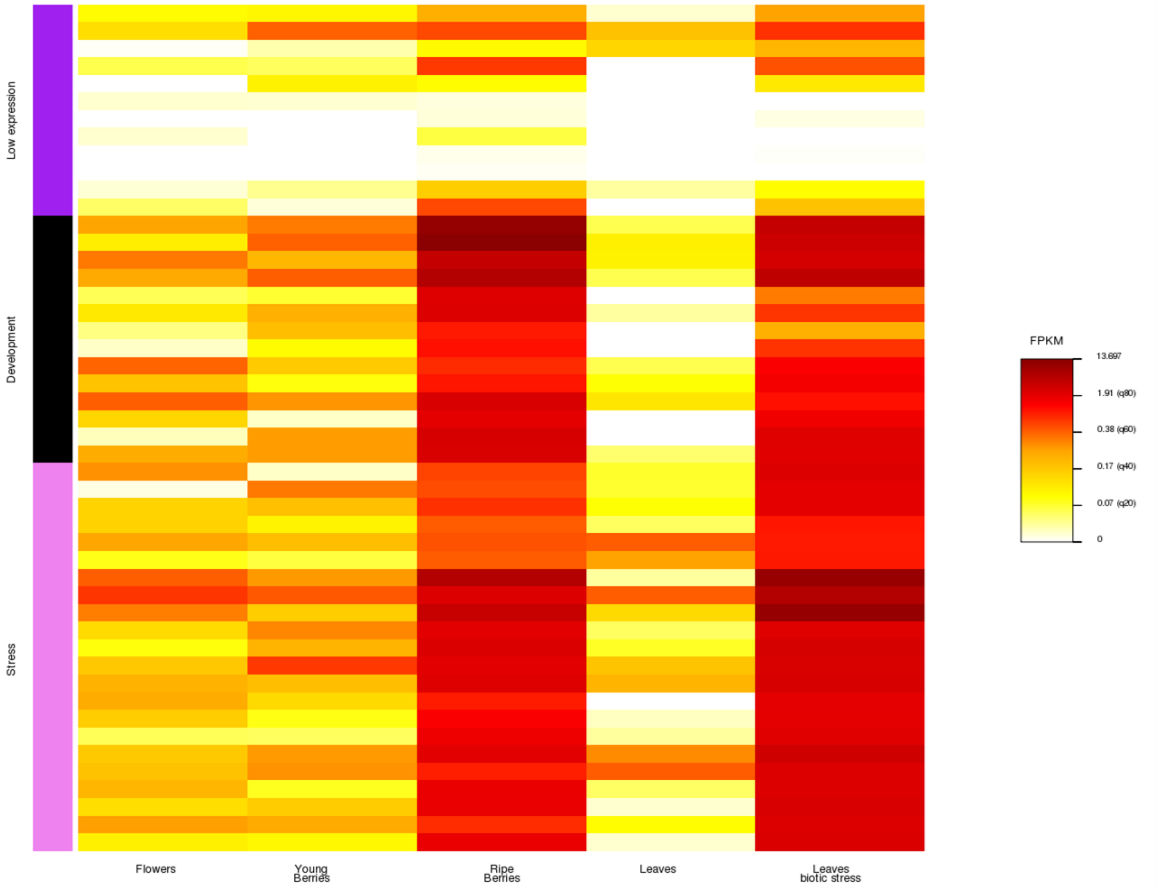



**Figure 3: heatmap of the expression for *STS* genes in the different categories.** The genes are horizontally ordered regarding their profile as shown by the clustering on the left. The expression value is in FPKM and genes in white are not expressed while red colour represents genes with high expression.

Powdery and downy mildew are major grapevine diseases by *Erysiphe necator* and *Plasmopara viticola*, respectively. As stilbenes are part of the defence metabolism, RNA-Seq data of grapevine leaves infected by these pathogens were studied in detail using a differential gene expression analysis (Supplementary figure 3). As a general result, 26 *STS* genes were significantly differentially expressed in at least one experiment, 22 of them being complete genes (*VvSTS13* and *25* are partial genes and *VvSTS1* and *18* are pseudogenes). The log fold changes varied between -1.33 and 6.23 for the differentially expressed genes. Even though expression levels of *STS* genes were higher in response to powdery mildew infection (data not shown), the fold changes were lower compared to downy mildew infection (Figure 4). Eighteen *STS* genes were found to be differentially expressed following downy mildew infection. For powdery mildew, the experiment was originally designed to analyse the transcriptome of *E. necator* (Jones *et al.*, 2014). As the earliest time point available was 12 hours post-infection (hpi), it was used as a reference. Four *STS* genes were differentially expressed at 24hpi and all of them were downregulated compared to their expression level at 12hpi. At 72hpi and 144hpi, 3 and 19 *STS* genes were differentially expressed and upregulated compared to 12hpi, respectively. Out of the 26 *STS* genes differentially expressed in at least one experiment, half were affected by both pathogens, 8 only by *E. necator* and 5 only by *P. viticola*. This confirms that in spite of their similar global expression profile, a fine regulation is applied on the *STS* gene family. In addition, all the 26 *STS* genes differentially expressed under biotic stress but *VvSTS1* (downregulated in powdery experiment at 24hpi compared to 12hpi) are part of the cluster on chromosome 16.

Comparing young and ripe berry experiments, 27 *VvSTS* genes were considered to be induced (FPKM > 1 in ripe berries and log fold change > 2). Five *VvSTS* were specifically induced in ripe berries.

Investigating expression patterns along the *STS* gene clusters, we could not identify blocks of neighbouring genes showing similar expression levels, and thus, we did not see evidence of co-expression and co-regulation of neighbour *STS* genes under the studied conditions. However, by comparing the expression patterns of *STS* genes across similarity groups, the highly similar *VvSTS10*, *15* and *21* within group 1 exhibited similar response to downy mildew infection. The same was observed for *VvSTS25*, *27*, *29* and *31* within group 2. Thus, gene regulation was probably conserved following the duplication events, which is supported by the high level of conservation of intergenic regions within group 1 and group 2, the putatively most recent duplicated groups.

**CNV analysis**

To investigate the presence / absence and copy number variations of the *STS* genes in other grapevine genotypes, a CNV analysis was performed using a segmentation approach. Based on their logratio values, segments were classified into five categories. The segments classified as "detected" harboured logratios close to 0, i.e. the normalized sequencing depth was similar for the genome of interest compared to the PN40024 reference. The "detected" *STS* genes were considered as present in the genome of interest with the same copy number than the reference. They represented the large majority of the 2688 total calls (48 *STS* genes x 56 genotypes) with 67% (1792) "detected" calls. The segments classified as "duplicated" harboured logratios greater than 1, i.e. the normalized sequencing depth was higher for the genome of interest compared to the reference. The "duplicated" *STS* genes were considered as present in the genome of interest with a greater number of copies than the reference. They were the least represented category with 2% of the total calls. The segments classified as "potentially duplicated" harboured intermediate logratios between "detected" and "duplicated" segments. These could be caused by partial or heterozygous duplications in the regions. The "potentially duplicated" *STS* genes were considered as present in the genome of interest with, at least, the same copy number than the reference. They represented a small proportion of the total calls with 3%. The segments classified as "not detected" harboured logratios lower than -2, i.e. the normalized sequencing depth was lower for the genome of interest compared to the reference. The "not detected" *STS* genes were considered as absent in the genome of interest or too divergent compared the reference. They represented a small proportion of the total calls with 3%. The segments classified as "partially detected" harboured intermediate logratios between "detected" and "not detected" segments. These could be caused by partial or heterozygous absence in the regions. The "partially detected" *STS* genes were not considered as strictly absent but, at least, partly present in the genome of interest compared to the reference.

Based on the number of "not detected" *STS* genes, we found that the selected *Vitis* and *Muscadinia* genotypes harboured between 42 and 48 members of the gene family (Figure 4). A third of the genotypes (18) did not have any "not detected" *STS* genes, i.e. harboured the full set of 48 *STS* genes like the reference genome. Only 4 genotypes were potentially missing 5 or 6 members of this gene family. All of these 4 genotypes were cultivars of *Vitis vinifera* spp *vinifera* (Carignan Noir, Chardonnay, Colorino and Jaen).

Among the studied genotypes, 38% did not harbour any "duplicated" *STS* genes and we detected 4 or 5 duplicated members for only 2 genotypes (*Muscadinia rotundifolia* Carlos and *Vitis vinifera* spp *vinifera* Cabernet Franc) (Figure 4). So *STS* genes copy number seems to be very constant among the *Vitis* and *Muscadinia* genotypes with only few variations. Thus, our analysis revealed that the amplification of the *VvSTS* gene family is not specific to the reference genome and is observed in all the studied genotypes. No clear pattern of conservation or variability could be identified among the species of grapevine. Unexpectedly, *Vitis sylvestris* subspecies are showing the least differences with the reference genome, which is a *V. vinifera* derived from Pinot Noir. So we can hypothesized that the expansion of the *STS* gene family is probably older than speciation, but that a strong negative selection pressure acted on maintaining a high identify percentage among them.

However, several *STS* genes were biased, as they were more prone to be not detected or duplicated than others. Only 15 *STS* genes were not detected in at least one genotype. *VvSTS26* (partial), *VvSTS27* (complete), *VvSTS43* (complete) and *VvSTS4* (pseudogene) were the most frequently ''not detected'' genes, with 19, 13, 13 and 8 genotypes, respectively. These four genes, which represented 60% of the "not detected" even, seem to be part of three blocks (*VvSTS2* to *6*, *26* to *28*; *43* to *46*) that accounted for 54% of the "partially detected" and 80% of the "not detected" events. Similarly, 13 *STS* genes were potentially duplicated in one genotype at least. *VvSTS11* (pseudogene), *VvSTS12* (partial), *VvSTS13* (partial) and *VvSTS14* (partial) were the most frequently duplicated genes, with 26, 5, 4 and 9 genotypes, respectively, which represented 73% of the "duplicated" events. These four genes defined one block that also accounted for 53% of the "potentially duplicated" events. Thus, these four blocks of *STS* genes seem to concentrate the variations and the dynamic of the whole gene family. On the opposite, several *STS* genes were more conserved than others. *VvSTS10* (complete), *VvSTS36* (complete) and *VvSTS41* (complete) were classified as "detected" for 54 out of 56 genotypes, i.e. they were present with the same number of copies than the reference genome. The highly conserved genes seem to be more scattered in the clusters but two blocks (9 to 10 and 36 to 39) could be identified. Thus, different levels of selective pressure may act on the *STS* genes and, most of the time, they may have an effect on contiguous blocks of genes.

These blocks of *STS* genes harbouring consistent patterns of CNV across the 56 studied genotypes could also be analysed regarding the similarity blocks that we identified in the reference genome. The group 3 (*VvSTS35* to *VvSTS38*) is the least subject to variations and is the most frequently detected group across the studied genotypes. Thus a specific selective constraint may be applied to this group of neighbour genes and probably spread to the neighbour subgroups 2d and 4a. On the contrary, the *STS* genes on the chromosome 10 and in the groups 2a, 2b and 4b seem to be enriched in "not detected" and "partially detected" events. So, these whole groups of *STS* may not be present in some genotypes. *V. vinifera* seems to be the most affected species regarding the putative absence of group 4b, whereas for the groups 2a and 2b, both *V. vinifera* and the other *Vitis* species seem to be affected. For the group of *STS* on chromosome 10, the pattern of putative absences is more scattered across species. The group 1a is notably enriched in "duplicated" and "potentially duplicated" events across the all species under study, which means that one more segmental duplication involving the group 1 might have happened in several *Vitis* genomes.

We then analysed CNV in the light of the expression patterns. Among the 12 *STS* genes classified as "detected" in more than 50 genotypes, 8 belongs to the 10 most expressed *STS* under in the studied conditions. Respectively, 6 and 6 out of these 12 *STS* genes also belongs to the 10 most induced *VvSTS* during berry maturation and downy mildew infection. These *STS* genes are mostly part of the groups 2d, 3a, 3b and 4a, which are probably the oldest ones, as proposed previously. On the contrary, among the 15 *STS* genes classified at least once as "not detected", 6 are in the 10 least expressed *STS* in the studied conditions. They mostly belong to the groups 1b, 1c, 2a, 2b, and 4b, which are probably the most recent ones as we proposed above. So, there seems to be a correlation between the age of the duplications that shaped the *STS* gene family within the reference genome, the conservation level of the *STS* among the studied genotypes and their expression level, especially during berry maturation and downy mildew infection.

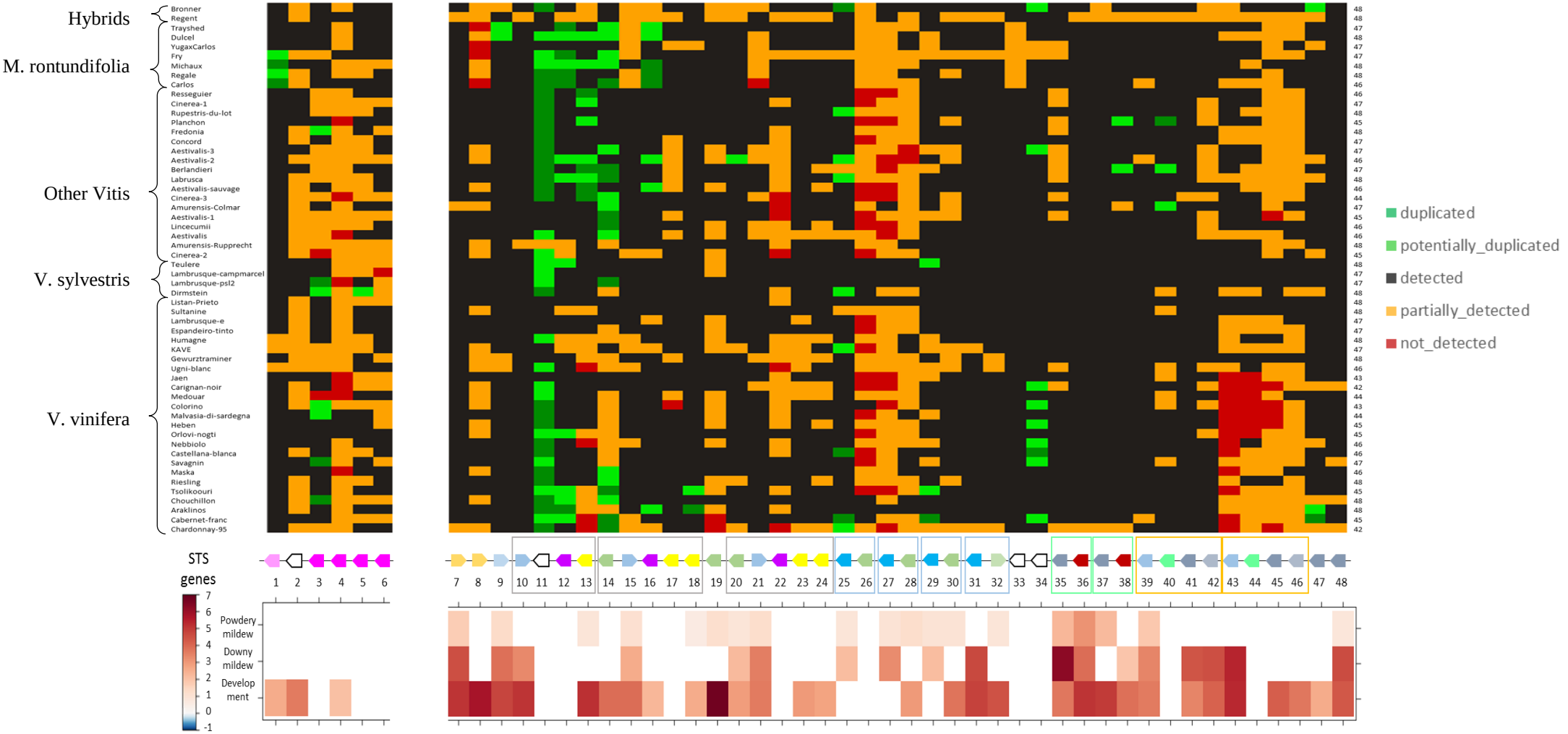



**Figure 4: conservation of STS genes linked to their expression.** The top part of the figure is a heatmap showing *STS* genes in the different genotypes. The red and orange squares are showing not detected and partially detected genes, respectively, black portions are genes detected and therefore conserved compared to the reference genome and light and dark green squares correspond to potentially duplicated and duplicated genes, respectively. On the left of the heatmap are shown the name of the different genotypes with their species: *Vitis vinifera*, *Vitis sylvestris*, Other *Vitis*, *Muscadinia rontundifolia* and hybrid of *Muscadinia rontundifolia* with *Vitis vinifera*. The numbers on the right represent the number of all but not detected *STS* genes in the genotypes. The bottom part of the figure is a heatmap showing the expression of *STS* genes. "Development" represents the ratio between the mean FPKM of ripe berries and the mean FPKM of young berries. Similarly, "Downy mildew" represents the FPKM ratio between leaves infected by *P. viticola* 24 hpi and control leaves. "Powdery mildew" represents the FPKM ratio between leaves infected by *E. necator* 72 hpi and 12 hpi. The ratios are presented in a logarithm scale. Down-regulated and up-regulated genes are indicated in blue and red, respectively.

# Discussion

### Blocks of duplicated genes can be identified in *STS* gene clusters

Using the phylogeny of the family and analysing the similarity levels in the intergenic regions, we have revealed similarity patterns of *STS* gene clusters, suggesting that they were formed through several rounds of segmental duplications in tandem involving two to five genes. In a previous study, a genome-wide detection of segmental duplications in grapevine allowed the identification of seven duplication events in *STS* clusters (five on chromosome 16 and two on chromosome 10) (Giannuzzi *et al.*, 2011). Despite the lower resolution of this analysis at the whole genome scale, the borders of the duplications corresponded well (more or less one gene) with the groups identified in this work.

### Toward a fine subfunctionalization of *STS* genes in grapevine

Even though the publicly available RNA-Seq dataset did not cover all plant organs and developmental stages, general patterns of *STS* gene expression could be identified. In general, *STS* genes were more expressed in ripe berries and leaves under biotic stress. This latter pattern of expression is consistent with the fact that stilbene biosynthesis is part of the response to biotic stress shown in Jiao and coworkers (2016) who showed that an allele from *Vitis pseudoreticulata* can confer resistance to powdery mildew due to its promoter region. Berry maturation resembles abiotic stress to some extent, due to oxidative reactions occurring during ripening (Pilati *et al.*, 2007).

Differential expression of *STS* genes during development or in response to stress conditions has led to the hypothesis of transcriptional subfunctionalization within the *STS* family (Vannozzi *et al.*, 2012). Indeed, the three group of *STS* genes, identified based on their sequence similarity, showed specific expression profiles. Group A, located on chromosome 10, was more expressed during development whereas group B and C were both specifically expressed under stress with a greater amplitude for group B. These results are in agreement with those of Vannozzi *et al.* (2012) and confirm that *STS* genes are mostly expressed in response to stress conditions, with similar expression profiles but with variable amplitudes. Some *STS* genes are indeed showing a high expression under specific conditions (*VvSTS10*, *18* and *31*) or in general (*VvSTS7*, *9*, *36* and *48*).

The fact that *STS* genes are not equally induced in response to stress suggests a fine-tuned regulation of the family. The two transcription factors MYB14 and MYB15 have been identified as regulators of the stilbene biosynthetic pathway (Höll *et al.*, 2013). Further investigation of *STS* gene regulation showed that MYB14 was a direct regulator of some of these genes, thus influencing resveratrol biosynthesis and plant defence (Fang *et al.*, 2014; Duan *et al.*, 2016). Very recently, a study focused on the MYB transcription factors family in grapevine suggested that *STS* genes could be regulated not only by MYB14 and MYB15, but also by MYB13 (Wong *et al.*, 2016). Thus, all these results suggest that grapevine *STS* family is subjected to fine transcriptional subfunctionalization and controlled by several transcription factor belonging to the MYB family.

**A subset of highly expressed *STS* genes is conserved throughout the *Vitis* genus**

To assess the expansion level of the *STS* gene family in the *Vitis* genus, a CNV approach based detection of variations in resequencing depth along genomic regions, was chosen. Similar approaches have been used to efficiently detect deletions and duplications in various organisms like human, *Drosophila*, cattle and *Arabidopsis thaliana* (Daines *et al.*, 2009; Yoon *et al.*, 2009; Cao *et al.*, 2011; Bickhart *et al.*, 2012; Luo *et al.*, 2016). The CNV analysis revealed that the expansion of the *STS* genes family was not restricted to the PN40024 reference genome, but was indeed observed in all 56 studied *Vitis* genomes. However, for particular genes and whole groups of genes, a higher variability among the different genotypes was highlighted. Thus, group 1a was shown to be potentially more duplicated than others, which suggests that an extra duplication event might have occurred in some genotypes compared to the reference genome. Conversely, some groups were potentially absent in several genotypes. As these particular groups were suggested to be tandemly duplicated recently several times in the reference genome, we hypothesized that one less duplication event may have occurred in some genotypes. Examples of CNV analyses performed on large gene families organized in clusters in plants are scarce in the literature. However, a CNV analysis was recently performed on the C-Repeat binding factor (CBF) gene family, organized in clusters and playing a major role in response to cold temperature in barley (Francia *et al.*, 2016). Variations in copy number for two members of this family among winter and spring genotypes of barley were identified and linked to frost resistance. Altogether, these studies suggest that gene families are highly subject to CNV, which may have strong impact on plant phenotype.

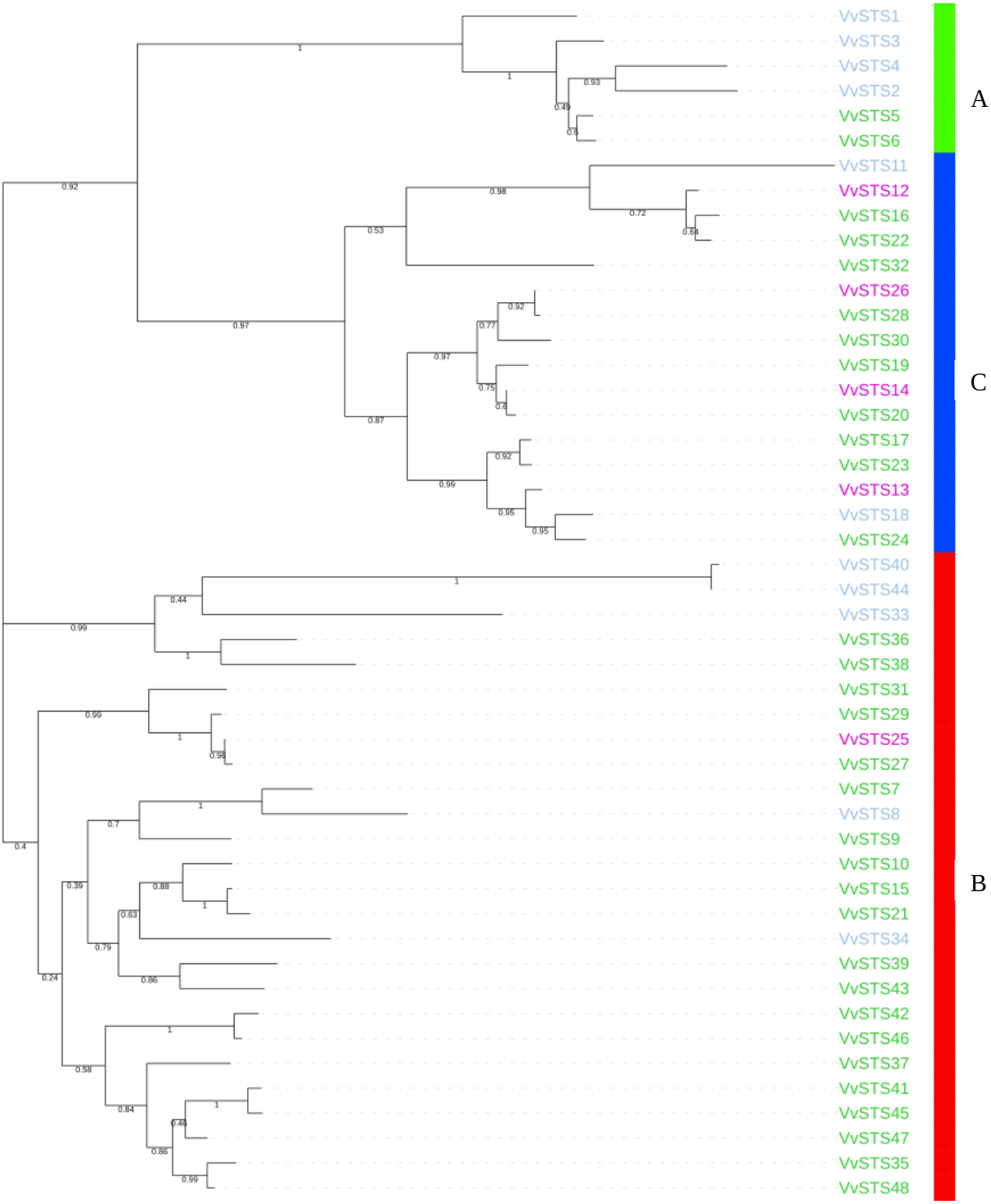
Chen and coworkers (2010) performed a CNV analysis of the resistance genes with a nucleotide-binding site (NBS) domain between *A. thaliana* and *A. lyrata*. They found that 20% of these genes were present in only one of the genomes and absent in the other one, which is a higher rate than what we found for the *STS* family. However, the NBS gene family has been shown to be subjected to strong positive selection whereas the *STS* gene family is highly constrained. In addition, we found that the *STS* genes that are more subjected to CNV tend to be less expressed than the more conserved ones. A similar observation was made by Pinosio *et al* (2016), who characterized the CNV in poplar genome. They found that genes affected by structural variations showed lower expression levels than average. However, these genes tended to be pseudogenes or harboured higher levels of dN/dS suggesting relaxed selective pressures, in contrast to *STS* genes.

On chromosome 16, *VvSTS31* to *VvSTS42* seems to be the most conserved set of genes among the studied genotypes, these genes being interestingly highly induced and expressed in response to stress. Although duplications occurred during the formation of this block of twelve genes, the whole block may be older than the others, as the intergenic regions are highly divergent. In addition, transposable elements annotated by Parage et al. (2012) in this region were not found to be duplicated elsewhere in the *STS* clusters. Thus, we propose that this region carrying twelve *STS* genes might constitute the core set of *STS* genes and that later duplication events may have resulted in a dosage effect on transcription, as suggested in other studies (Iovene *et al.*, 2013; Wang *et al.*, 2015), thus boosting the number of transcripts under stress conditions.

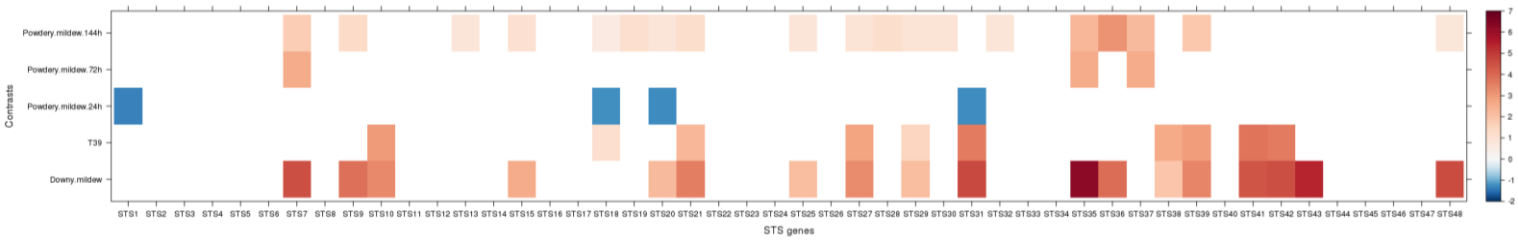
## Conclusion

Similarity patterns suggest that *STS* genes clusters were formed through several rounds of segmental duplications in tandem in grapevine genome. *STS* genes exhibited similar expression profiles during berry maturation and in response to stresses. However, in the tested conditions, a small number of highly expressed genes accounted for most of gene expression in the family, which appears to be subjected to a fine-tuned regulation relying on a combination of both highly and low-expressed genes. A CNV analysis confirmed the expansion of the *STS* gene family in the *Vitis* genus with a global conservation of gene number. However, some blocks of genes were probably subjected to one more or one less duplication in certain genotypes compared to the reference genome. Interestingly, older duplicated blocks carrying *VvSTS31* to *VvSTS42* not only showed higher expression and induction under stress, but also higher conservation in the *Vitis* genus, suggesting that they may constitute the core set of *STS* genes necessary for stilbene biosynthesis. Later duplications of these genes may have induced a dosage effect on transcription and introduced more flexibility in gene regulation. This probably conferred an adaptive advantage to grapevine under stress conditions, resulting in the current negative selective pressure acting on the whole family.

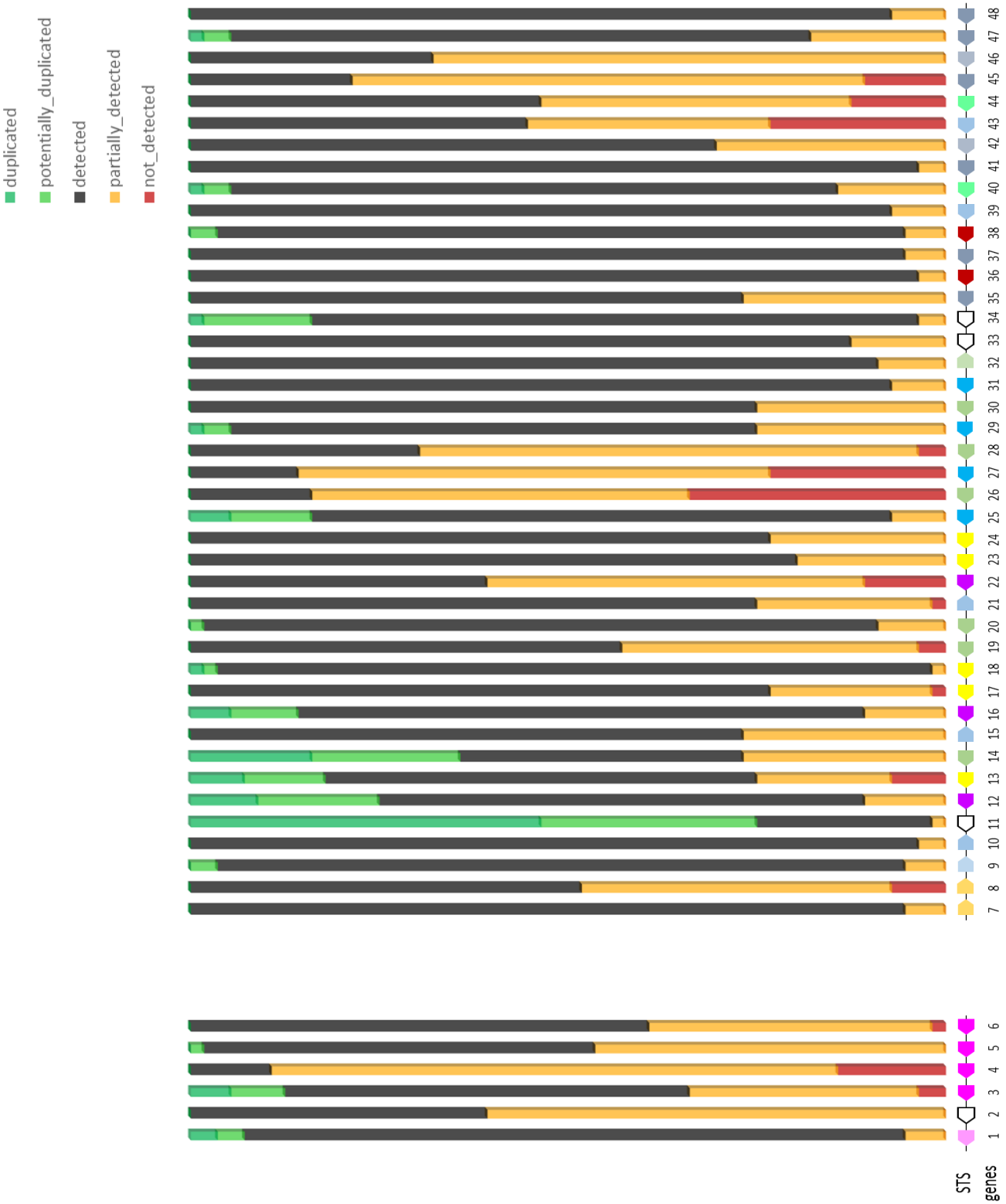
## Supplementary data




**Supplementary figure 1: phylogenic tree of the *STS* gene family.** Sequences used are nucleotides. The green, purple and blue colours correspond respectively to complete, partial and pseudogenes. The green, blue and red bands are representing the three different phylogenic groups.




**Supplementary figure 3: significant fold change of *STS* genes in different contrasts.** Downy mildew represents the contrast between infection with downy mildew and no infection, T39 represents the contrast between T39 and no infection, Powdery mildew 24h, 72h, 144h represent the contrasts between infection with powdery mildew at respectively 24 hpi vs. 12 hpi, 72 hpi vs. 12 hpi and 144 hpi vs. 12 hpi. Each square represents a gene differentially expressed with blue being down regulation and red being up regulation




**Supplementary Figure 4: histogram representing the *STS* genes among different genotypes.** In the histogram, the red and orange bars are showing not detected and partially detected genes, respectively, black portions are genes detected and therefore conserved compared to the reference genome and light and dark green bars correspond to potentially duplicated and duplicated genes, respectively.

**Supplementary table 1: Recapitulative table of RNA-Seq experiments analysed.** bp means base pair. hpi = hours post inoculation.

| Group / Project | Specie / Variety | Reads | Studied conditions |
|---|---|---|---|
| Da Silva | Tannat | Illumina<br>paired 2x100 bp | Berries 1 week after flowering<br>Skins 5 and 7 weeks after flowering<br>Seeds 7 weeks after flowering |
| Jones | Carignan | Illumina<br>paired 2x100 bp | Leaves infected with powdery mildew 12, 24, 72 and 144 hpi |
| Perazzolli | Pinot Noir | Illumina<br>paired 2x100 bp | Leaves control<br>Leaves treated with T39<br>Leaves infected with downy mildew 24 hpi<br>Leaves treated with T39 and infected with downy mildew 24 hpi |
| Ramos | Vitis sylvestris | Illumina 51 bp | Flowers at B, D, G and H developmental stages |
| Sweetman | Shiraz | Illumina 100 bp | Young berries<br>Berries before véraison<br>Berries after véraison<br>Ripe berries |
| Venturini | Corvina | Illumina<br>paired 2x51 bp | Young berries<br>Mid ripening berries<br>Mid withering berries (2 months after harvest) |
| Vitaroma | Carménère<br>Gewurztraminer<br>Petit Verdot<br>Riesling | Illumina<br>paired 2x38 bp | Young berries<br>Mid ripening berries |

# References


**Anders, S., Pyl, P.T. and Huber, W.** (2015) HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

**Bickhart, D.M., Hou, Y., Schroeder, S.G., et al.** (2012) Copy number variation of individual cattle genomes using next-generation sequencing. , 778–790.

**Cao, J., Schneeberger, K., Ossowski, S., et al.** (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.*, **43**, 956–963.

**Chen, Q., Han, Z., Jiang, H., Tian, D. and Yang, S.** (2010) Strong positive selection drives rapid diversification of R-Genes in arabidopsis relatives. *J. Mol. Evol.*, **70**, 137–148.

**Chong, J., Poutaraud, A. and Hugueney, P.** (2009) Metabolism and roles of stilbenes in plants. *Plant Sci.*, **177**, 143–155.

**Daines, B., Wang, H., Li, Y., Han, Y., Gibbs, R. and Chen, R.** (2009) High-throughput multiplex sequencing to discover copy number variants in Drosophila. *Genetics*, **182**, 935–941.

**Duan, D., Fischer, S., Merz, P., Bogs, J., Riemann, M. and Nick, P.** (2016) An ancestral allele of grapevine transcription factor MYB14 promotes plant defence. *J. Exp. Bot.*, **67**, 1795–1804.

**Edgar, R.C.** (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

**Fang, L., Hou, Y., Wang, L., Xin, H., Wang, N. and Li, S.** (2014) Myb14, a direct activator of STS, is associated with resveratrol content variation in berry skin in two grape cultivars. *Plant Cell Rep.*, **33**, 1629–1640.

**Francia, E., Morcia, C., Pasquariello, M., Mazzamurro, V., Milc, J.A., Rizza, F., Terzi, V. and Pecchioni, N.** (2016) Copy number variation at the HvCBF4–HvCBF2 genomic segment is a major component of frost resistance in barley. *Plant Mol. Biol.*, **92**, 161–175.

**Giannuzzi, G., D'Addabbo, P., Gasparro, M., Martinelli, M., Carelli, F.N., Antonacci, D. and Ventura, M.** (2011) Analysis of high-identity segmental duplications in the grapevine genome. *BMC Genomics*, **12**, 436.

**Höll, J., Vannozzi, A., Czemmel, S., et al.** (2013) The R2R3-MYB transcription factors MYB14 and MYB15 regulate stilbene biosynthesis in Vitis vinifera. *Plant Cell*, **25**, 4135–49.

**Iovene, M., Zhang, T., Lou, Q., Buell, C.R. and Jiang, J.** (2013) Copy number variation in potato - an asexually propagated autotetraploid species. *Plant J.*, **75**, 80–89.

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

**Jiao, Y., Xu, W., Duan, D., Wang, Y. and Nick, P.** (2016) A stilbene synthase allele from a Chinese wild grapevine confers resistance to powdery mildew by recruiting salicylic acid signalling for efficient defence. *J. Exp. Bot.*, **67**, erw351.

**Jones, L., Riaz, S., Morales-Cruz, A., Amrine, K.C., McGuire, B., Gubler, W.D., Walker, M.A. and Cantu, D.** (2014) Adaptive genomic structural variation in the grape powdery mildew pathogen, *Erysiphe necator*. *BMC Genomics*, **15**, 1081.

**Kumar, S., Stecher, G. and Tamura, K.** (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.*, **33**, msw054.

**Letunic, I. and Bork, P.** (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

**Luo, S., Yu, J.A. and Song, Y.S.** (2016) Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads. , 1–21.

**Parage, C., Tavares, R., Rety, S., et al.** (2012) Structural, Functional, and Evolutionary Analysis of the

Unusually Large Stilbene Synthase Gene Family in Grapevine. *Plant Physiol.*, **160**, 1407–1419.

**Perazzolli, M., Moretto, M., Fontana, P., Ferrarini, A., Velasco, R., Moser, C., Delledonne, M. and Pertot, I.** (2012) Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. *BMC Genomics*, **13**, 660.

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

**Pinosio, S., Giacomello, S., Faivre-Rampant, P., et al.** (2016) Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Mol. Biol. Evol.*, **33**, 2706–19.

**R Development Core Team** (2016) R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput. Vienna Austria*, **0**, {ISBN} 3-900051-07-0.

**Ramos, M.J.N., Coito, J.L., Silva, H.G., Cunha, J., Costa, M.M.R. and Rocheta, M.** (2014) Flower development and sex specification in wild grapevine. *BMC Genomics*, **15**, 1095.

**Robinson, M.D., McCarthy, D.J. and Smyth, G.K.** (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

**Sievers, F., Wilm, A., Dineen, D., et al.** (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

**Silva, C. Da, Zamperin, G., Ferrarini, A., 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.

**Sparvoli, F., Martin, C., Scienza, A., Gavazzi, G., Tonelli, C. and Celoria, V.** (1994) Cloning and molecular analysis of structural genes involved in flavonoid and stilbene biosynthesis in grape (Vitis vinifera L.). *Plant Mol Biol*, **75969**, 743–755.

**Springer, N.M., Ying, K., Fu, Y., et al.** (2009) Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genet.*, **5**, e1000734.

**Sweetman, C., Wong, D.C., Ford, C.M. and Drew, D.P.** (2012) Transcriptome analysis at four developmental stages of grape berry (Vitis vinifera cv. Shiraz) provides insights into regulated and coordinated gene expression. *BMC Genomics*, **13**, 691.

**Vannozzi, A., Dry, I.B., Fasoli, M., Zenoni, S. and Lucchin, M.** (2012) Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. *BMC Plant Biol.*, **12**, 130.

**Venturini, L., Ferrarini, A., Zenoni, S., et al.** (2013) De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. *BMC Genomics*, **14**, 41.

**Vitulo, N., Forcato, C., Carpinelli, E., 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.

**Wang, Y., Xiong, G., Hu, J., et al.** (2015) Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet.*, **47**, 944–948.

**Wong, D.C.J., Schlechter, R., Vannozzi, A., Höll, J., Hmmam, I., Bogs, J., Tornielli, G.B., Castellarin, S.D. and Matus, J.T.** (2016) A systems-oriented analysis of the grapevine R2R3-MYB transcription factor family uncovers new insights into the regulation of stilbene accumulation. *DNA Res.*, **0**, dsw028.

**Wu, T.D. and Nacu, S.** (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

**Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J.** (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

# Conclusion générale

Cette thèse ayant été réalisée dans l'unité ''Santé de la Vigne et Qualité du Vin'', le choix des familles de gènes étudiées s'est naturellement porté sur des gènes impliqués dans la qualité des raisins (*cytochromes P450*) et dans les mécanismes de résistance et de défense de la vigne contre les pathogènes (*gènes de résistance à domaine NBS*, gènes *STS* et gènes codant pour les endo-β-1,3-glucanases). Ces familles, impliquées dans des voies métaboliques très différentes, ont également des caractéristiques structurales très différentes, résumées dans le tableau 1.

Au cours de cette thèse, l'annotation manuelle de trois familles de gènes a pu être exploitée et a permis de mener des études de génomique structurale approfondies. En effet, l'organisation des familles de gènes de cytochromes P450, à domaine NBS et *STS* a été analysée dans le génome de référence de la vigne. Les familles de gènes de cytochrome P450 et à domaine NBS sont structurées en sous-familles sur la base de la conservation de leur séquence protéique et sur la base de la présence de certains domaines fonctionnels, respectivement. Les trois familles étudiées sont organisées en clusters dans le génome de référence. Malgré les caractéristiques différentes de ces familles, des hypothèses similaires ont pu être émises quant à la formation des clusters. Les clusters sont formés, dans la majorité des cas, par des gènes appartenant à la même sous-famille et donc partageant des similarités de séquences très fortes. Les analyses menées dans les différents articles proposent que les clusters soient la résultante de duplications en tandem (c'est-à-dire proximale) d'un gène ou de bloc de plusieurs gènes. Les gènes à domaines TIR-NBS semblent faire exception puisque quelques exemples de duplications segmentales lointaines impliquant plusieurs gènes ont pu être identifiés. Cette organisation en clusters formés à partir de duplications en tandem de gènes est un mécanisme commun chez les plantes et en particulier pour les familles de gènes impliqués dans le métabolisme secondaire (Panchy *et al.*, 2016; Cannon *et al.*, 2004).

Une analyse transcriptomique a été menée sur les familles de gènes des cytochromes P450, des endo-β-1,3-glucanases et des *STS*. Bien que des variations dans le niveau d'expression aient été observées, globalement, ces familles de gènes présentent une expression significative dans les organes ou les conditions dans lesquels leur fonction s'exerce. Ainsi, les profils d'expression des gènes *STS* sont similaires avec une forte expression dans les feuilles soumises à un stress biotique et dans les baies mûres, deux conditions de stress pour la plante induisant la production de composés de défense tels que les stilbènes. De même, les niveaux d'expression les plus élevés ont été détectés dans les conditions de stress pour les gènes codant pour les endo-β-1,3-glucanases. Les profils d'expression pour les gènes de cytochromes P450 se sont avérés beaucoup plus divers, probablement du fait de l'implication des différentes sous-familles de P450 dans des voies métaboliques extrêmement diverses. Ainsi des gènes présentant une expression constitutive et d'autres présentant des profils d'induction dans certaines conditions ont pu être identifiés. Cependant, certaines sous-familles comptant un grand nombre de membres organisés en cluster ne semblaient pas présenter de profils similaires ou particuliers, ce qui permettrait d'émettre l'hypothèse d'une potentielle sub-fonctionnalisation transcriptionnelle de ces gènes (Panchy *et al.*, 2016) comme cela a été suggéré pour la sous-familles CYP75 (Falginella *et al.*, 2010).

L'approche de génomique comparative basée sur la détection de CNV pour les membres des familles de gènes à domaine NBS et *STS* a permis de mettre en évidence une conservation globale du nombre de copies au sein des génomes du genre *Vitis*. La majeure partie des variations observées était catégorisée comme ‘‘partiellement détecté’’ que l'on peut interpréter comme les gènes ayant le plus divergé au niveau de leur séquence, qui ne pourraient être présents que de façon tronquée ou sous forme hémizygote. Les deux familles de gènes présentaient des disparités entre gènes puisque certains semblent plus soumis à variations que les autres jusqu'à former des blocs de gènes consécutifs dans le cas des gènes *STS*. Pour les gènes à domaines NBS également, les clusters composés de gènes à domaines CC-NBS semblent davantage soumis à variations. Cependant, alors que les variations observées au sein de la famille des gènes à domaine NBS structurent les génotypes selon la phylogénie des espèces, ce n'est pas le cas pour les gènes *STS*. Une hypothèse serait que la pression de sélection qui s'exerce sur la famille des gènes *STS* soit variable selon les génotypes alors qu'elle serait plus constante pour les gènes à domaine NBS et suivrait la dynamique de spéciation.

Enfin les relations entre l'organisation, l'expression et l'évolution des membres d'une famille de gènes n'ont été approfondies que pour les gènes *STS* sur lesquels des travaux avaient été préalablement menés dans l'équipe. Cette étude a permis de mettre en évidence 12 gènes *STS* dont l'expression est parmi les plus fortes en conditions de stress et qui présentent les plus fort taux de conservation au sein du genre *Vitis*. Cependant, aucune organisation nette n'a pu être mise en évidence quant à ce groupe de gènes, à savoir qu'il n'est pas constitué de gènes fortement similaires entre eux ou de grands blocs de gènes. Afin de valider l'importance de ces gènes dans la synthèse de stilbènes, il serait intéressant de compléter l'analyse avec des données de RNA-Seq issus d'organes riches en stilbènes de façon constitutive comme les racines et les parties ligneuses. Cette caractérisation approfondie de la famille des gènes *STS* pourrait également être menée sur les autres familles étudiées durant ma thèse voire sur de nouvelles, telles que la famille des terpènes synthases impliquées dans la synthèse d'arômes dans les baies.

Ma thèse permet donc d'apporter des connaissances fondamentales nouvelles et d'appréhender la dynamique évolutive du génome de la vigne à l'échelle de plusieurs familles de gènes ainsi que d'étudier le rôle de la régulation sur la conservation de la structure de ces familles dans différents génomes. De plus, mes travaux ont permis de mettre en place un outil d'analyse du transcriptome facilitant l'accès au profil d'expression de gènes d'intérêt dans les différentes expériences analysées. Cela a d'ailleurs permis de valider le profil d'expression des gènes candidats codant pour les endo-β-1,3-glucanases qu'avaient identifié expérimentalement d'autres collègues. Grâce aux outils développés, aux analyses réalisées et aux collaborations, mon travail a déjà été valorisé dans une publication et diverses communications et sera valorisé par des publications actuellement en cours de préparation.

Ces travaux contribuent à la caractérisation de facteurs de défense efficaces et durables ainsi que des gènes impliqués dans la synthèse d'arômes dans la vigne. Ceux-ci pourront contribuer à orienter les choix de gènes à intégrer dans les programmes de création variétale mis en œuvre à l'INRA de Colmar dont le but est de développer des variétés résistantes aux principales maladies de la vigne tout en maintenant un fort potentiel qualitatif. Ainsi, par mon approche, certes fondamentale, j'ai pu contribuer à diminuer l'impact de la viticulture sur l'environnement sans transiger sur la qualité de la boisson alcoolisée la plus consommée en France.

**Tableau 1 :** Résumé des caractéristiques structurales, transcriptionnelles et de génomique comparative des quatre familles de gènes étudiées.

| Familles de gènes codant pour… | Impliquées dans… | Nombre de gènes | Sous-familles | % identité nucléotidique | Clusters | Formation des clusters | Expression | CNV au sein du genre *Vitis* |
|---|---|---|---|---|---|---|---|---|
| **endo-β-1,3-glucanases** | Défense | 23 | 1 | 72 à 76% (entre EGase1, 2 et 3) | 2 | × | Essentiellement stress et baies mûres | × |
| **cytochromes P450** | Métabolisme secondaire, Arômes et défense | 579 | 48 (certaines en expansion chez la vigne et d'autres non) | 51 à 86% en moyenne au sein de chaque sous-famille; <40% entre sous-familles | 85 | Duplications en tandem de 1 à plusieurs gènes | Par sous-famille, pas de claires tendances à l'expression spécifique dans une condition | × |
| **protéines à domaine NBS** | Résistance | 829 | 8 (2 principales : avec domaine TIR et avec domaine CC) | CC : 24 à 99% (moy. 39%) TIR : 28 à 99% (moy. 56%) | 122 | CC : duplications en tandem TIR : duplications en tandem et segmentales lointaines | × | Globalement conservés, variabilité essentiellement dans la catégorie des ‘‘partiellement détectés’’ |
| **STS** | Défense | 48 | Non | 70 à 99,9% | 2 | Duplications en tandem de 1 à plusieurs gènes | Essentiellement stress et baies mûres | Globalement conservés, variabilité essentiellement dans la catégorie des ‘‘partiellement détectés’’ |

# Perspectives

## 1. Approfondir les connaissances sur l'expression et la régulation des gènes *STS*

### 1.1. Des gènes STS répondant aux stress et d'autres gènes STS de ménage ?

Vannozzi et ses collègues (2012) ont suggéré que les gènes *STS* localisés sur le chromosome 10 seraient des gènes de ménage, présentant une expression constitutive et constante dans les jeunes feuilles dans les conditions expérimentales qu'ils ont étudiées. Étant donné son rôle essentiel, on peut émettre l'hypothèse qu'un gène de ménage soit très conservé au sein des génomes du genre *Vitis*. Nos résultats montrent cependant que, sur le chromosome 10, seul le gène *STS1* est très conservé, ce gène correspondant à un pseudogène dans le génome de référence. Afin de vérifier l'hypothèse de Vannozzi et de ses collègues, mais aussi de compléter l'analyse de l'expression des gènes *STS*, il serait intéressant d'obtenir des données RNA-Seq correspondant aux parties ligneuses et racines de la vigne. Ces organes ont, en effet, été décrits comme très riches en stilbènes, et ceci de manière constitutive (Lambert *et al.*, 2013; Tisserant *et al.*, 2016). En réalisant cette analyse, il serait aussi possible de voir si la synthèse de stilbènes dans les parties ligneuses et les racines est également liée aux gènes identifiés comme très conservés et très exprimés en conditions de stress ou bien, s'il existe réellement des gènes *STS* ''de ménage'' dédiés à la synthèse constitutive de stilbènes.

### 1.2. Analyse de la réponse des gènes *STS* aux contraintes abiotiques liées au climat

Les données en conditions de stress utilisées lors de l'analyse transcriptomique se focalisaient sur les stress biotiques causés par des infections par le mildiou et l'oïdium. Il serait intéressant d'étudier d'autres conditions expérimentales comme les stress abiotiques liés au climat. Certaines études se sont déjà intéressées à ces problématiques comme illustré dans la revue en introduction (Rienth *et al.*, 2014; Corso *et al.*, 2015; Xin *et al.*, 2013). Les effets du changement climatique se faisant de plus en plus marqués, la vigne est soumise à des stress de plus en plus prononcés et doit être capable de résister à une amplitude de températures plus grande. Identifier, s'il y en a, les gènes *STS* majeurs répondant à ces stress pourrait présenter un intérêt afin d'en tenir compte dans les futurs programmes de création variétale.

### 1.3. Étude des marques épigénétiques et des éléments transposables potentiellement impliqués dans la régulation des gènes *STS*

Lors de l'analyse de l'expression des gènes *STS*, il a été possible de mettre en évidence une différence d'expression entre ces différents gènes. Bien que les mécanismes de régulation de l'expression de ces gènes aient été précédemment étudiés par Höll et ses collègues (2013), l'étude réalisée porte sur les facteurs de transcriptions et ne s'intéresse pas la régulation spécifique de chaque gène. Une piste prometteuse serait d'étudier le rôle de la méthylation d'ADN sur la régulation transcriptomique comme cela a été fait pour le génome humain (Miller and Grant, 2013). Cette analyse pourrait être effectuée en utilisant la technologie PacBio qui propose, en plus du séquençage qui sera abordé par la suite, d'identifier les bases pouvant être méthylées. Une autre possibilité serait d'étudier la nature et l'organisation des éléments transposables au sein des clusters de gènes *STS* car ces derniers pourraient jouer un rôle dans la régulation mais aussi dans la duplication des gènes. Cependant, cette analyse est à ce jour rendue difficile par une annotation automatique partielle des éléments transposables dans le génome de la vigne alors qu'une annotation fine serait requise pour ce type d'étude.

## 2. Exploitation des nouvelles techniques de séquençage produisant des lectures longues

### 2.1. Amélioration de la séquence de référence PN40024

La vigne présente un intérêt scientifique majeur avec une communauté importante de chercheurs de par le monde et en France particulièrement. Bien que de bonne qualité générale, la séquence de référence de PN40024 (Jaillon *et al.*, 2007) n'est cependant pas parfaite et certaines analyses se révèlent complexes à mettre en place. En effet, il est fréquent de retrouver des portions de génome qui n'ont pas été séquencées. Il serait intéressant, vu l'émergence et la démocratisation des nouvelles techniques de séquençage, d'améliorer cette séquence et ce en utilisant par exemple la technologie PacBio ou le nanopore sequencing permettant d'obtenir de longs fragments séquencés. Cela représenterait une avancée majeure pour l'ensemble des chercheurs travaillant sur la vigne permettant d'utiliser le plein potentiel des outils et analyses disponibles mais aussi d'élargir le champ des possibilités et le développement de nouveaux outils.

En particulier, dans cette thèse, il est fortement suspecté que certains gènes *STS* n'ont pas été identifiés du fait que certaines portions de la région contenant ces gènes n'ont pas été séquencées dans le génome de référence. La disponibilité d'une séquence de référence complétée pourrait donc permettre une analyse parfaitement exhaustive de la famille des gènes *STS* ou de toute autre famille de gènes d'intérêt comme la famille des gènes codant pour les terpènes synthases. En effet, les terpènes sont des composés majeurs impliqués dans la senteur des fleurs mais aussi dans la production des arômes des vins. Malgré leur intérêt, les études des gènes codant pour les terpènes synthases ont été rendue difficiles par la mauvaise qualité de l'assemblage de la séquence de référence particulièrement dans les régions portant ces gènes.

### 2.2. Validation des résultats de l'analyse des CNV dans quelques génomes de *Vitis vinifera*

À l'heure actuelle, une des limitations de l'analyse des CNV est le fait que nous ne disposons pas de validation pour la méthode par segmentation. Un moyen de lever cette limitation serait de séquencer et réaliser un assemblage *de novo* de certains génotypes afin de vérifier si les résultats de notre analyse sont en accord avec la réalité. Le Gewurztraminer et le Riesling, cépages majeurs dans la structuration de la diversité génétique de la vigne cultivée (Myles *et al.*, 2011), ont été sélectionnés pour être séquencés en utilisant la technologie PacBio. Les données brutes sont disponibles et l'assemblage de ces génomes est en court. Une fois assemblés, il sera possible d'analyser les gènes *STS* et *NBS* dans ces génomes. De plus, ces cépages étant fondamentalement différents, posséder leurs séquences ouvrirait de nouvelles portes quant à, par exemple, l'étude des arômes dans les vins.

### 2.3. Validation des résultats de l'analyse des CNV dans quelques génomes de *Vitis* sauvages

Notre étude sur les gènes *STS* a montré des résultats surprenants pour les génotypes *Vitis sylvestris.* En effet, ils ne présentent que peu de variations par rapport au génome de référence, ce qui mériterait d'être validé par des analyses complémentaires. Il serait envisageable de réaliser une analyse évolutive de la famille des gènes codant pour les *STS* dans les *V. sylvestris* et de l'étendre aux *Muscadinia rotundifolia* et autres *Vitis*. Ces analyses sont d'autant plus réalisables du fait que les génomes d'un *V. sylvestris* et d'une *M. rotundifolia* viennent d'être séquencés avec la technologie PacBio. Suite à l'assemblage, une validation des résultats de l'analyse des CNV pour la famille des gènes *STS* pourra être réalisée.

De plus, dans les analyses actuelles, c'est la séquence de référence PN40024 (*Vitis vinifera*), qui est utilisée pour réaliser les alignements. Il serait préférable, comme il y a des données de reséquençage de différents génotypes, d'utiliser le génotype correspondant, s'il est disponible, ou un génotype proche parmi ceux dont une séquence PacBio est disponible. En effet, de par l'éloignement de certains génotypes par rapport au génome de référence, les séquences peuvent être plus ou moins divergentes. Cela peut provoquer un biais lors d'alignements de séquences dupliquées ou similaires. Par ailleurs, les génomes assemblés avec les séquences PacBio seront probablement moins fragmentés que le génome de référence actuel. Ces deux améliorations permettront d'améliorer les alignements des séquences et les analyses qui en découlent pour permettre de comprendre avec plus d'exactitude la dynamique des familles de gènes étudiées. Au final, ces nouvelles données ne changent pas fondamentalement les approches utilisées, uniquement la méthode, mais permettront de tirer des conclusions plus précises.

### 2.4. Un intérêt pour l'étude de l'hétérozygotie

Le fait d'obtenir les séquences génomiques de différents cépages et espèces permettra aussi de s'intéresser à l'hétérozygotie de ces cépages. En effet, le génome de référence PN40024, dérivé d'un Pinot Noir, est à 93% homozygote alors que les autres génotypes ont un fort degré d'hétérozygotie (Hyma *et al.*, 2015; Fodor *et al.*, 2014). Cela provoque un biais dans le cas de génomes possédant un allèle nul et qui sont, lors de notre analyse, représentés par une profondeur plus faible que la normale. Ces cas font probablement partie de la catégorie ‘‘partiellement détectés’’.

## 3. L'amélioration de l'outil d'analyse du transcriptome

### 3.1. L'automatisation du pipeline d'analyse et l'utilisation des dernières versions de l'assemblage et de l'annotation

Comme indiqué dans la partie 1, le pipeline d'analyses développé durant cette thèse n'est pas encore entièrement automatisé. Le fait de ne devoir lancer qu'un seul script qui effectuerait, sans intervention ultérieure, toutes les étapes nécessaires à l'obtention des valeurs d'expression, pour un jeu de données spécifié, faciliterait le travail de mise à jour de cet outil.

Il est aussi à noter que l'analyse effectuée a été réalisée sur la première version de l'ancrage et de l'annotation du génome de référence. Il existe depuis une version améliorée où, concernant l'ancrage, il a été possible d'attribuer un emplacement sur les différents chromosomes des portions du génome se trouvant sur le chromosome ''unknown''. À propos de l'annotation, une seconde version est également disponible avec non seulement de nouveaux exons mais également de nouveaux gènes. Il serait intéressant d'effectuer la mise à jour de l'outil en utilisant ces dernières versions.

D'autre part, l'outil proposé sous forme de site internet n'est pas encore accessible au grand public car pas totalement finalisé. Afin de proposer une utilisation conviviale et un accès au plus grand nombre, il convient de terminer le développement de cet outil, ce qui devrait être réalisé dans les mois qui viennent. Un aperçu du design et du fonctionnement du site internet est proposé dans les perspectives de la partie 1 de cette thèse.

### 3.2.De nouvelles méthodes de normalisation

Le choix a été fait, lors du développement de l'outil transcriptomique, d'utiliser les FPKM comme méthode de normalisation. Ce choix a été motivé par le fait que nous désirons une vue globale de l'expression des gènes. Il n'est pas possible de connaitre à priori vers quels gènes l'intérêt de l'utilisateur sera porté et quel gène sera plus intéressant qu'un autre. Après avoir identifié nos besoins pour l'analyse, les FPKM se sont avérés être un choix judicieux, dans notre cas, car cela permet une comparaison de l'expression des gènes entre eux, ce que ne permettent pas certains autres outils ou méthodes qui se basent sur des contrastes ou de l'analyse différentielle d'expression. Il existe cependant d'autres outils et méthodes, par exemple edgeR (Robinson *et al.*, 2009) et DESeq2 (Love *et al.*, 2014), qui proposent différentes normalisations afin de réaliser différentes analyses comme de l'expression différentielle. Il serait intéressant de traiter nos données avec ces outils afin de proposer à l'utilisateur plusieurs manières d'observer les données selon ses besoins.

### 3.3. Une veille afin de favoriser la mise à jour de la base de données

Sur le long terme, un travail de mise à jour de l'outil et d'intégration des nouvelles données disponibles devra être envisagé. Il serait intéressant de donner la possibilité à un utilisateur de contacter la personne en charge du maintien à jour du site internet et ce, afin de proposer un ou des jeux de données pour qu'ils soient ajoutés à notre base de données. Cela permettrait d'augmenter la puissance de cet outil en proposant toujours plus de conditions expérimentales pour des analyses transcriptomiques les plus complètes possibles.

### 3.4. Des analyses de co-expression de gènes facilitées

Grâce au pipeline d'analyse d'expression réalisée sur le génome entier, les analyses de co-expression pourraient s'avérer facilitées. L'information étant disponible, il s'agirait d'isoler les gènes ayant un profil d'expression similaire. Cela permettrait d'identifier des mécanismes pouvant potentiellement interagir de manière positive pour, par exemple, favoriser les défenses de la plante. Il serait aussi possible d'étudier la régulation des gènes d'intérêt en analysant leur co-expression éventuelle avec des facteurs de transcriptions, et également de vérifier si des sites de liaison sont présents, dans le génome d'intérêt, non loin du ou des gènes étudiés. Cela pourrait permettre de proposer des hypothèses pour expliquer, dans le cas de gènes se ressemblant fortement, pourquoi différents niveau d'expression sont observés, comme c'est le cas pour les gènes *STS*.

### 3.5. Dans le cas d'expression différentielle

Le mildiou et l'oïdium étant des fléaux toujours présents pour la vigne, il serait intéressant d'approfondir l'analyse de l'expression différentielle en sélectionnant quelques espèces présentant une tolérance à ces pathogènes et d'autres présentant une sensibilité. Cela implique de réaliser plus d'expériences de RNA-Seq dans diverses conditions. Avec ces nouvelles données et les nouvelles séquences génomiques décrites précédemment, il serait donc possible d'identifier, de manière certaine, les gènes différentiellement exprimés en cas de sensibilité ou de tolérance. Il serait possible d'étudier les différences d'expression entre les génotypes plus tolérants comparés à ceux sensibles mais aussi d'identifier les gènes répondant, de manière spécifique, à un pathogène donné.

La possibilité d'étudier les réponses de la vigne à des pathogènes autres que le mildiou et l'oïdium est aussi envisageable en réalisant d'autres expériences de RNA-Seq. Cela permettrait, à terme, de pouvoir sélectionner les plantes résistantes en fonction d'une analyse précoce du génome afin d'identifier la présence de gènes importants pour la défense de la vigne.

## 4. Quid d'autres familles de gènes ?

Il serait envisageable d'étudier de nouvelles familles de gènes ou des familles de gènes dont une amplification en nombre dans la vigne a déjà été montrée. L'étude pourrait porter sur des familles de gènes importantes pour la défense de la plante, les arômes ou des mécanismes autres. Cela permettrait de savoir si les mêmes mécanismes observés dans cette thèse, en particulier dans l'étude des gènes *STS*, se retrouvent dans d'autres familles de gènes. Par exemple, il serait intéressant de s'intéresser aux facteurs de transcription MYB et WRKY mais aussi aux gènes codant pour les callose synthases.

# Références bibliographiques générales


**Abbà, S., Galetto, L., Carle, P., Carrère, S., Delledonne, M., Foissac, X., Palmano, S., Veratti, F. and Marzachì, C.** (2014) RNA-Seq profile of flavescence dorée phytoplasma in grapevine. *BMC Genomics*, **15**, 1088.

**Adams, D.J.** (2004) Fungal cell wall chitinases and glucanases. *Microbiology*, **150**, 2029–2035.

**Agrawal, G.K. and Rakwal, R.** (2006) Rice proteomics: a cornerstone for cereal food crop proteomes. *Mass Spectrom Rev*, **25**, 1–53.

**Ahuja, I., Kissen, R. and Bones, A.M.** (2012) Phytoalexins in defense against pathogens. *Trends Plant Sci.*, **17**, 73–90.

**Akita, M. and Valkonen, J.P.T.** (2002) A novel gene family in moss (Physcomitrella patens) shows sequence homology and a phylogenetic relationship with the TIR-NBS class of plant disease resistance genes. *J. Mol. Evol.*, **55**, 595–605.

**Allen, J.E. and Salzberg, S.L.** (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–603.

**Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J.** (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

**Anastasiou, E., Kenz, S., Gerstung, M., MacLean, D., Timmer, J., Fleck, C. and Lenhard, M.** (2007) Control of Plant Organ Size by KLUH/CYP78A5-Dependent Intercellular Signaling. *Dev. Cell*, **13**, 843–856.

**Anders, S., Pyl, P.T. and Huber, W.** (2015b) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

**Andersen, E.J., Ali, S., Reese, R.N., Yen, Y. and Neupane, S.** (2016) Diversity and Evolution of Disease Resistance Genes in Barley ( Hordeum vulgare L .). , 99–108.

**Andersen, M.D.** (2000) Cytochromes P-450 from Cassava (Manihot esculenta Crantz) Catalyzing the First Steps in the Biosynthesis of the Cyanogenic Glucosides Linamarin and Lotaustralin. *J. Biol. Chem.*, **275**, 1966–1975.

**Andolfo, G., Jupe, F., Witek, K., et al.** (2014) Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC Plant Biol.*, **14**, 120.

**Andolfo, G., Sanseverino, W., Rombauts, S., Peer, Y. Van de, Bradeen, J.M., Carputo, D., Frusciante, L. and Ercolano, M.R.** (2013) Overview of tomato (Solanum lycopersicum) candidate pathogen recognition genes reveals important Solanum R locus dynamics. *New Phytol.*, **197**, 223–237.

**Anon** (2015) Food and Agriculture Organisation fo the United Nations. Food and Agricultural commodities production / Commodities by region.

**Aquea, F., Vega, A., Timmermann, T., Poupin, M.J. and Arce-Johnson, P.** (2011) Genome-wide analysis of the SET DOMAIN GROUP family in Grapevine. *Plant Cell Rep.*, **30**, 1087–1097.

**Arlorio, M., Ludwig, A., Boller, T. and Bonfante, P.** (1992) Inhibition of fungal growth by plant chitinases andβ-1,3-glucanases. *Protoplasma*, **171**, 34–43.

**Ayabe, S. and Akashi, T.** (2006) Cytochrome P450s in flavonoid metabolism. *Phytochem. Rev.*, **5**, 271–282.

**Aziz, A., Poinssot, B., Daire, X., Adrian, M., Bézier, A., Lambert, B., Joubert, J.-M. and Pugin, A.** (2003) Laminarin elicits defense responses in grapevine and induces protection against Botrytis cinerea and Plasmopara viticola. *Mol. Plant Microbe Interact.*, **16**, 1118–1128.

**Baginsky, S. and Gruissem, W.** (2006) Arabidopsis thaliana proteomics: from proteome to genome. *J Exp Bot*, **57**, 1485–1491.

**Bai, J., Pennill, L.A., Ning, J., et al.** (2002) Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. *Genome Res.*, **12**, 1871–1884.

**Bairoch, A. and Boeckmann, B.** (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **19 Suppl**, 2247–2249.

**Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S. and Werck-Reichhart, D.** (2011) Cytochromes P450. *Arab. B.*, **9**, e0144.

**Balasubramanian, V., Vashisht, D., Cletus, J. and Sakthivel, N.** (2012) Plant β-1,3-glucanases: their biological functions and transgenic expression against phytopathogenic fungi. *Biotechnol. Lett.*, **34**, 1983–1990.

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

**Beaudoin, G. a W. and Facchini, P.J.** (2013) Isolation and characterization of a cDNA encoding (S)-cis-N-methylstylopine 14-hydroxylase from opium poppy, a key enzyme in sanguinarine biosynthesis. *Biochem. Biophys. Res. Commun.*, **431**, 597–603.

**Beffa, R. and Meins, F.** (1996) Pathogenesis-related functions of plant beta-1,3-glucanases investigated by antisense transformation--a review. *Gene*, **179**, 97–103.

**Belkhadir, Y., Subramaniam, R. and Dangl, J.L.** (2004) Plant disease resistance protein signaling: NBS-LRR proteins and their partners. *Curr. Opin. Plant Biol.*, **7**, 391–399.

**Belli Kullan, J., Lopes Paim Pinto, D., Bertolini, E., et al.** (2015) miRVine: a microRNA expression atlas of grapevine based on small RNA sequencing. *BMC Genomics*, **16**, 393.

**Belval, L., Marquette, A., Mestre, P., Piron, M.-C., Demangeat, G., Merdinoglu, D. and Chich, J.-F.** (2015) A fast and simple method to eliminate Cpn60 from functional recombinant proteins produced by E. coli Arctic Express. *Protein Expr. Purif.*, **109**, 29–34.

**Bent, A.F. and Mackey, D.** (2007) Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions. *Annu. Rev. Phytopathol.*, **45**, 399–436.

**Bergelson, J., Kreitman, M., Stahl, E. a and Tian, D.** (2001) Evolutionary dynamics of plant R-genes. *Science*, **292**, 2281–2285.

**Bézier, A., Lambert, B. and Baillieul, F.** (2002) Cloning of a grapevine Botrytis- responsive gene that has homology to the tobacco hypersensitivity- related hsr203J. *J. Exp. Bot.*, **53**, 2279–2280.

**Biazzi, E., Carelli, M., Tava, A., Abbruscato, P., Losini, I., Avato, P., Scotti, C. and Calderini, O.** (2015) CYP72A67 catalyses a key oxidative step in Medicago truncatula hemolytic saponin biosynthesis. *Mol. Plant*, **8**, 1493–1506.

**Bickhart, D.M., Hou, Y., Schroeder, S.G., et al.** (2012) Copy number variation of individual cattle genomes using next-generation sequencing. , 778–790.

**Bindschedler, L. V, Panstruga, R. and Spanu, P.D.** (2016) Mildew-Omics: How Global Analyses Aid the Understanding of Life and Evolution of Powdery Mildews. *Front. Plant Sci.*, **7**, 123.

**Bisson, L.F., Waterhouse, A.L., Ebeler, S.E., Walker, M.A. and Lapsley, J.T.** (2002) The present and future of the international wine industry. *Nature*, **418**, 696–699.

**Blanc, S., Wiedemann-Merdinoglu, S., Dumas, V., Mestre, P. and Merdinoglu, D.** (2012) A reference genetic map of Muscadinia rotundifolia and identification of Ren5, a new major locus for resistance to grapevine powdery mildew. *Theor. Appl. Genet.*, **125**, 1663–1675.

**Boachon, B., Junker, R.R., Miesch, L., et al.** (2015) CYP76C1 (Cytochrome P450)-Mediated Linalool Metabolism and the Formation of Volatile and Soluble Linalool Oxides in Arabidopsis Flowers: A Strategy for Defense against Floral Antagonists. *Plant Cell*, **27**, 2972–90.

**Bomblies, K. and Weigel, D.** (2010) Arabidopsis and relatives as models for the study of genetic and genomic incompatibilities. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **365**, 1815–1823.

**Bönisch, F., Frotscher, J., Stanitzek, S., Rühl, E., Wüst, M., Bitz, O. and Schwab, W.** (2014) A UDP-Glucose:Monoterpenol Glucosyltransferase Adds to the Chemical Diversity of the Grapevine Metabolome. *Plant Physiol.*, **165**, 561–581.

**Bonomelli, A., Mercier, L., Franchel, J., Baillieul, F., Benizri, E. and Mauro, M.-C.** (2004) Response of Grapevine Defenses to UV—C Exposure. *Am. J. Enol. Vitic.*, **55**, 51–59.

**Booker, J., Sieberer, T., Wright, W., et al.** (2005) MAX1 Encodes a Cytochrome P450 Family Member that Acts Downstream of MAX3/4 to Produce a Carotenoid-Derived Branch-Inhibiting Hormone. *Dev. Cell*, **8**, 443–449.

**Borad, V. and Sriram, S.** (2008) Pathogenesis-related proteins for the plant protection. *Asian J Exp Sci*.

**Borneman, A.R., Schmidt, S.A. and Pretorius, I.S.** (2013) At the cutting-edge of grape and wine biotechnology. *Trends Genet.*, **29**, 263–271.

**Britto, D.S., Pirovani, C.P., Andrade, B.S., Santos, T.P. Dos, Pungartnik, C., Cascardo, J.C.M., Micheli, F. and Gesteira, A.S.** (2013) Recombinant β-1,3-1,4-glucanase from Theobroma cacao impairs Moniliophthora perniciosa mycelial growth. *Mol. Biol. Rep.*, **40**, 5417–5427.

**Bulcke, M. V, Bauw, G., Castresana, C., Montagu, M. Van and Vandekerckhove, J.** (1989) Characterization of vacuolar and extracellular beta(1,3)-glucanases of tobacco: Evidence for a strictly compartmentalized plant defense system. *Proc. Natl. Acad. Sci. USA*, **86**, 2673–2677.

**Busso, D., Delagoutte-Busso, B. and Moras, D.** (2005) Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in Escherichia coli. *Anal. Biochem.*, **343**, 313–321.

**Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L.** (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 1–9.

**Cândido, E. de S., Pinto, M.F.S., Pelegrini, P.B., Lima, T.B., Silva, O.N., Pogue, R., Grossi-de-Sá, M.F. and Franco, O.L.** (2011) Plant storage proteins with antimicrobial activity: novel insights into plant defense mechanisms. *FASEB J*, **25**, 3290–3305.

**Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D. and May, G.** (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol.*, **4**, 10.

**Cánovas, F.M., Dumas-Gaudot, E., Recorbet, G., Jorrin, J., Mock, H.-P. and Rossignol, M.** (2004) Plant proteome analysis. *Proteomics*, **4**, 285–298.

**Cao, J., Schneeberger, K., Ossowski, S., et al.** (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.*, **43**, 956–963.

**Carelli, M., Biazzi, E., Panara, F., et al.** (2011) Medicago truncatula CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. *Plant Cell*, **23**, 3070–81.

**Castresana, J.** (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.

**Causse, M., Desplat, N., Pascual, L., et al.** (2013) Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics*, **14**, 791–805.

**Chai, L., Li, Y., Chen, S., Perl, A., Zhao, F. and Ma, H.** (2014) RNA sequencing reveals high resolution expression change of major plant hormone pathway genes after young seedless grape berries treated with gibberellin. *Plant Sci.*, **229**, 215–24.

**Chakrabarti, M., Zhang, N., Sauvage, C., et al.** (2013) A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 17125–30.

**Chen, Q., Han, Z., Jiang, H., Tian, D. and Yang, S.** (2010) Strong positive selection drives rapid diversification of R-Genes in arabidopsis relatives. *J. Mol. Evol.*, **70**, 137–148.

**Cheng, D.W., Lin, H., Takahashi, Y., Walker, M.A., Civerolo, E.L. and Stenger, D.C.** (2010) Transcriptional regulation of the grape cytochrome P450 monooxygenase gene CYP736B expression in response to Xylella fastidiosa infection. *BMC Plant Biol.*, **10**, 135.

**Chin, C., Peluso, P., Sedlazeck, F.J., et al.** (2016) Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing.

**Chong, J., Poutaraud, A. and Hugueney, P.** (2009) Metabolism and roles of stilbenes in plants. *Plant Sci.*, **177**, 143–155.

**Christ, B., Sussenbacher, I., Moser, S., Bichsel, N., Egert, A., Muller, T., Krautler, B. and Hortensteiner, S.** (2013) Cytochrome P450 CYP89A9 Is Involved in the Formation of Major Chlorophyll Catabolites during Leaf Senescence in Arabidopsis. *Plant Cell*, **25**, 1868–1880.

**Christie, N., Tobias, P.A., Naidoo, S. and Külheim, C.** (2016) The Eucalyptus grandis NBS-LRR Gene Family: Physical Clustering and Expression Hotspots. *Front. Plant Sci.*, **6**, 1238.

**Collu, G., Unver, N., Peltenburg-Looman, A.M.G., Heijden, R. van der, Verpoorte, R. and Memelink, J.** (2001) Geraniol 10-hydroxylase, a cytochrome P450 enzyme involved in terpenoid indole alkaloid biosynthesis. *FEBS Lett.*, **508**, 215–220.

**Corso, M., Vannozzi, A., Maza, E., et al.** (2015) Comprehensive transcript profiling of two grapevine rootstock genotypes contrasting in drought susceptibility links the phenylpropanoid pathway to enhanced tolerance. *J. Exp. Bot.*, **66**, 5739–5752.

**Cramer, G.R., Ghan, R., Schlauch, K.A., et al.** (2014) Transcriptomic analysis of the late stages of grapevine (Vitis vinifera cv. Cabernet Sauvignon) berry ripening reveals significant induction of ethylene signaling and flavor pathways in the skin. *BMC Plant Biol.*, **14**, 370.

**Cuadros-Inostroza, A., Ruíz-Lara, S., González, E., Eckardt, A., Willmitzer, L. and Peña-Cortés, H.** (2016) GC–MS metabolic profiling of Cabernet Sauvignon and Merlot cultivars during grapevine berry development and network analysis reveals a stage- and cultivar-dependent connectivity of primary metabolites. *Metabolomics*, **12**, 39.

**Cucurachi, M., Busconi, M., Morreale, G., Zanetti, A., Bavaresco, L. and Fogher, C.** (2012) Characterization and differential expression analysis of complete coding sequences of Vitis vinifera L. sirtuin genes. *Plant Physiol. Biochem.*, **54**, 123–132.

**Cui, H., Tsuda, K. and Parker, J.E.** (2014) Effector-Triggered Immunity: From Pathogen Perception to Robust Defense. *Annu. Rev. Plant Biol.*

**Czemmel, S., Galarneau, E.R., Travadon, R., McElrone, A.J., Cramer, G.R. and Baumgartner, K.** (2015) Genes expressed in grapevine leaves reveal latent wood infection by the fungal pathogen Neofusicoccum parvum. *PLoS One*, **10**, 1–21.

**Dadakova, K., Havelkova, M., Kurkova, B., Tlolkova, I., Kasparovsky, T., Zdrahal, Z. and Lochman, J.** (2015) Proteome and transcript analysis of Vitis vinifera cell cultures subjected to Botrytis cinerea infection. *J. Proteomics*, **119**, 143–153.

**Daines, B., Wang, H., Li, Y., Han, Y., Gibbs, R. and Chen, R.** (2009) High-throughput multiplex sequencing to discover copy number variants in Drosophila. *Genetics*, **182**, 935–941.

**Dal Santo, S., Vannozzi, A., Tornielli, G.B., Fasoli, M., Venturini, L., Pezzotti, M. and Zenoni, S.** (2013) Genome-Wide Analysis of the Expansin Gene Superfamily Reveals Grapevine-Specific Structural and Functional Characteristics. *PLoS One*, **8**.

**Dang, T.-T.T. and Facchini, P.J.** (2014) CYP82Y1 Is N-Methylcanadine 1-Hydroxylase, a Key Noscapine Biosynthetic Enzyme in Opium Poppy. *J. Biol. Chem.*, **289**, 2013–2026.

**Dangl, J.L. and Jones, J.D.** (2001) Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.

**Danilova, N.** (2006) The evolution of immune mechanisms. *J. Exp. Zool. B Mol. Dev. Evol.*, **306**, 496–520.

**Das, M. and Das, D.K.** (2010) Resveratrol and cardiovascular health. *Mol. Aspects Med.*, **31**, 503–512.

**Davies, G. and Henrissat, B.** (1995) Structures and mechanisms of glycosyl hydrolases. *Structure*, **3**, 853–859.

**Degu, A., Hochberg, U., Sikron, N., et al.** (2014) Metabolite and transcript profiling of berry skin during fruit development elucidates differential regulation between Cabernet Sauvignon and Shiraz cultivars at

branching points in the polyphenol pathway. *BMC Plant Biol.*, **14**, 1–20.

**Degu, A., Morcia, C., Tumino, G., et al.** (2015) Metabolite profiling elucidates communalities and differences in the polyphenol biosynthetic pathways of red and white Muscat genotypes. *Plant Physiol. Biochem.*, **86**, 24–33.

**Derckel, J.-P., Audran, J.-C., Haye, B., Lambert, B. and Legendre, L.** (1998) Characterization, induction by wounding and salicylic acid, and activity against Botrytis cinerea of chitinases and β-1,3-glucanases of ripening grape berries. *Physiol. Plant.*, **104**, 56–64.

**Dereeper, A., Guignon, V., Blanc, G., et al.** (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465--469.

**Dereeper, A., Nicolas, S., Cunff, L. Le, Bacilieri, R., Doligez, A., Peros, J.-P., Ruiz, M. and This, P.** (2011) SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics*, **12**, 134.

**Dèrozier, S., Samson, F., Tamby, J.-P., et al.** (2011) Exploration of plant genomes in the FLAGdb++ environment. *Plant Methods*, **7**, 8.

**Desikan, R., Reynolds, A., Hancock, J.T. and Neill, S.J.** (1998) Harpin and hydrogen peroxide both initiate programmed cell death but have differential effects on defence gene expression in Arabidopsis suspension cultures. *Biochem. J.*, **330 ( Pt 1**, 115–120.

**Diaz-Chavez, M.L., Moniodis, J., Madilao, L.L., et al.** (2013) Biosynthesis of Sandalwood Oil: Santalum album CYP76F Cytochromes P450 Produce Santalols and Bergamotol. *PLoS One*, **8**, e75053.

**Domingos, S., Fino, J., Paulo, O.S., Oliveira, C.M. and Goulao, L.F.** (2016) Molecular candidates for early-stage flower-to-fruit transition in stenospermocarpic table grape (Vitis vinifera L.) inflorescences ascribed by differential transcriptome and metabolome profiles. *Plant Sci.*, **244**, 40–56.

**Dong, X., Mindrinos, M., Davis, K.R. and Ausubel, F.M.** (1991) Induction of Arabidopsis defense genes by virulent and avirulent Pseudomonas syringae strains and by a cloned avirulence gene. *Plant Cell*, **3**, 61–72.

**Doxey, A.C., Yaish, M.W.F., Moffatt, B.A., Griffith, M. and McConkey, B.J.** (2007) Functional divergence in the Arabidopsis beta-1,3-glucanase gene family inferred by phylogenetic reconstruction of expression states. *Mol. Biol. Evol.*, **24**, 1045–1055.

**Duan, D., Fischer, S., Merz, P., Bogs, J., Riemann, M. and Nick, P.** (2016) An ancestral allele of grapevine transcription factor MYB14 promotes plant defence. *J. Exp. Bot.*, **67**, 1795–1804.

**Duchêne, E., Huard, F., Dumas, V., Schneider, C. and Merdinoglu, D.** (2010) The challenge of adapting grapevine varieties to climate change. *Clim. Res.*, **41**, 193–204.

**Duchêne, E. and Schneider, C.** (2005) Grapevine and climatic changes: a glance at the situation in Alsace. *Agron. Sustain. Dev.*, **25**, 93–99.

**Ebrahim, S., Usha, K. and Singh, B.** (2011) Science against microbial pathogens: communicating current research and technological advances A. Méndez-Vilas, ed. , **3**, 1043–1054.

**Eddy, S.R.** (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.

**Edgar, R.C.** (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

**Ehlting, J., Hamberger, B., Million-Rousseau, R. and Werck-Reichhart, D.** (2006) Cytochromes P450 in phenolic metabolism. *Phytochem. Rev.*, **5**, 239–270.

**Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E.** (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, 1–10.

**Falginella, L., Castellarin, S.D., Testolin, R., Gambetta, G. a, Morgante, M. and Gaspero, G. Di** (2010) Expansion and subfunctionalisation of flavonoid 3',5'-hydroxylases in the grapevine lineage. *BMC Genomics*, **11**, 562.

**Falginella, L., Gaspero, G. Di and Castellarin, S.D.** (2012) Expression of flavonoid genes in the red grape berry of "Alicante Bouschet" varies with the histological distribution of anthocyanins and their chemical composition. *Planta*, **236**, 1037–51.

**Fang, L., Hou, Y., Wang, L., Xin, H., Wang, N. and Li, S.** (2014) Myb14, a direct activator of STS, is associated with resveratrol content variation in berry skin in two grape cultivars. *Plant Cell Rep.*, **33**, 1629–1640.

**Farrow, S.C., Hagel, J.M., Beaudoin, G. a W., Burns, D.C. and Facchini, P.J.** (2015) Stereochemical inversion of (S)-reticuline by a cytochrome P450 fusion in opium poppy. *Nat. Chem. Biol.*, **11**, 728–732.

**Fasoli, M., Dal Santo, S., Zenoni, S., et al.** (2012) The Grapevine Expression Atlas Reveals a Deep Transcriptome Shift Driving the Entire Plant into a Maturation Program. *Plant Cell*, **24**, 3489–3505.

**Feyereisen, R.** (2011) Arthropod CYPomes illustrate the tempo and mode in P450 evolution. *Biochim. Biophys. Acta - Proteins Proteomics*, **1814**, 19–28.

**Finn, R.D., Bateman, A., Clements, J., et al.** (2014) Pfam: The protein families database. *Nucleic Acids Res.*, **42**.

**Fodor, A., Segura, V., Denis, M., et al.** (2014) Genome-wide prediction methods in highly diverse and heterozygous species: Proof-of-concept through simulation in grapevine. *PLoS One*, **9**.

**Francia, E., Morcia, C., Pasquariello, M., Mazzamurro, V., Milc, J.A., Rizza, F., Terzi, V. and Pecchioni, N.** (2016) Copy number variation at the HvCBF4–HvCBF2 genomic segment is a major component of frost resistance in barley. *Plant Mol. Biol.*, **92**, 161–175.

**Friedman, A.R. and Baker, B.J.** (2007) The evolution of resistance genes in multi-protein plant resistance systems. *Curr. Opin. Genet. Dev.*, **17**, 493–499.

**Galtier, N., Gouy, M. and Gautier, C.** (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*, **12**, 543–548.

**Genova, A. Di, Almeida, A.M., Muñoz-Espinoza, C., et al.** (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.*, **14**, 7.

**George, I.S. and Haynes, P.A.** (2014) Current perspectives in proteomic analysis of abiotic stress in Grapevines. *Front. Plant Sci.*, **5**, 686.

**Ghan, R., Sluyter, S.C. Van, Hochberg, U., et al.** (2015) Five omic technologies are concordant in differentiating the biochemical characteristics of the berries of five grapevine (Vitis vinifera L.) cultivars. *BMC Genomics*, **16**, 946.

**Giacomelli, L., Nanni, V., Lenzi, L., et al.** (2012) Identification and Characterization of the Defensin-Like Gene Family of Grapevine. *Mol. Plant-Microbe Interact.*, **25**, 1118–1131.

**Giannuzzi, G., D'Addabbo, P., Gasparro, M., Martinelli, M., Carelli, F.N., Antonacci, D. and Ventura, M.** (2011) Analysis of high-identity segmental duplications in the grapevine genome. *BMC Genomics*, **12**, 436.

**Giribaldi, M. and Giuffrida, M.G.** (2010) Heard it through the grapevine: proteomic perspective on grape and wine. *J Proteomics*, **73**, 1647–1655.

**Giribaldi, M., Purrotti, M., Pacifico, D., et al.** (2011) A multidisciplinary study on the effects of phloem-limited viruses on the agronomical performance and berry quality of Vitis vinifera cv. Nebbiolo. *J Proteomics*, **75**, 306–315.

**González-Agüero, M., García-Rojas, M., Genova, A. Di, Correa, J., Maass, A., Orellana, A. and Hinrichsen, P.** (2013) Identification of two putative reference genes from grapevine suitable for gene expression analysis in berry and related tissues derived from RNA-Seq data. *BMC Genomics*, **14**, 878.

**Gouy, M., Guindon, S. and Gascuel, O.** (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol. Biol. Evol.*, **27**, 221–224.

**Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W. and Dangl, J.L.** (1995) Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. *Science*, **269**,

843–6.

**Gray, D.J., Li, Z.T. and Dhekney, S.A.** (2014) Precision breeding of grapevine (Vitis vinifera L.) for improved traits. *Plant Sci.*, **228**, 3–10.

**Greer, S., Wen, M., Bird, D., Wu, X., Samuels, L., Kunst, L. and Jetter, R.** (2007) The cytochrome P450 enzyme CYP96A15 is the midchain alkane hydroxylase responsible for formation of secondary alcohols and ketones in stem cuticular wax of Arabidopsis. *Plant Physiol.*, **145**, 653–667.

**Gresele, P., Cerletti, C., Guglielmini, G., Pignatelli, P., Gaetano, G. de and Violi, F.** (2011) Effects of resveratrol and other wine polyphenols on vascular function: An update. *J. Nutr. Biochem.*, **22**, 201–211.

**Grimplet, J., Adam-Blondon, A.-F., Bert, P.-F., et al.** (2014) The grapevine gene nomenclature system. *BMC Genomics*, **15**, 1077.

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

**Grimplet, J., Martínez-zapater, J.M. and Carmona, M.J.** (2016) Structural and functional annotation of the MADS-box transcription factor family in grapevine. *BMC Genomics*, 1–23.

**Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O.** (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.

**Guo, J., Zhou, Y.J., Hillwig, M.L., et al.** (2013) CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts. *Proc. Natl. Acad. Sci.*, 1–6.

**Guo, L., Gao, Z. and Qian, Q.** (2014) Application of resequencing to rice genomics, functional genomics and evolutionary analysis. *Rice (N. Y).*, **7**, 4.

**Guo, Y.-L., Fitz, J., Schneeberger, K., Ossowski, S., Cao, J. and Weigel, D.** (2011) Genome-Wide Comparison of Nucleotide-Binding Site-Leucine-Rich Repeat-Encoding Genes in Arabidopsis. *Plant Physiol.*, **157**, 757–769.

**Hamberger, B. and Bak, S.** (2013) Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **368**, 20120426.

**Han, Y., Gao, S., Muegge, K., Zhang, W. and Zhou, B.** (2015) Advanced applications of RNA sequencing and challenges. *Bioinform. Biol. Insights*, **9**, 29–46.

**Hannah, L., Roehrdanz, P.R., Ikegami, M., Shepard, A. V, Shaw, M.R., Tabor, G., Zhi, L., Marquet, P. a and Hijmans, R.J.** (2013) Climate change, wine, and conservation. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 6907–12.

**Hansen, C.H.** (2001) Cytochrome P450 CYP79F1 from Arabidopsis Catalyzes the Conversion of Dihomomethionine and Trihomomethionine to the Corresponding Aldoximes in the Biosynthesis of Aliphatic Glucosinolates. *J. Biol. Chem.*, **276**, 11078–11085.

**Hara-Nishimura, I. and Hatsugai, N.** (2011) The role of vacuole in plant cell death. *Cell Death Differ.*, **18**, 1298–1304.

**Hatlestad, G.J., Sunnadeniya, R.M., Akhavan, N. a, Gonzalez, A., Goldman, I.L., McGrath, J.M. and Lloyd, A.M.** (2012) The beet R locus encodes a new cytochrome P450 required for red betalain production. *Nat. Genet.*, **44**, 816–20.

**Hatsugai, N. and Hara-Nishimura, I.** (2010) Two vacuole-mediated defense strategies in plants. *Plant Signal. Behav.*, **5**, 1568–1570.

**Haudenschild, C., Schalk, M., Karp, F. and Croteau, R.** (2000) Functional expression of regiospecific cytochrome P450 limonene hydroxylases from mint (Mentha spp.) in Escherichia coli and Saccharomyces cerevisiae. *Arch. Biochem. Biophys.*, **379**, 127–36.

**Helliwell, C.A., Sheldon, C.C., Olive, M.R., Walker, A.R., Zeevaart, J.A.D., Peacock, W.J. and Dennis,**

**E.S.** (1998) Cloning of the Arabidopsis ent-kaurene oxidase gene GA3. *Proc. Natl. Acad. Sci.*, **95**, 9019–9024.

**Henk, A.D., Warren, R.F. and Innes, R.W.** (1999) A new Ac-like transposon of arabidopsis is associated with a deletion of the RPS5 disease resistance gene. *Genetics*, **151**, 1581–1589.

**Hofer, R., Boachon, B., Renault, H., et al.** (2014) Dual function of the cytochrome P450 CYP76 family from Arabidopsis thaliana in the metabolism of monoterpenols and phenylurea herbicides. *Plant Physiol.*, **166**, 1149–1161.

**Hofius, D., Munch, D., Bressendorff, S., Mundy, J. and Petersen, M.** (2011) Role of autophagy in disease resistance and hypersensitive response-associated cell death. *Cell Death Differ.*, **18**, 1257–1262.

**Höll, J., Vannozzi, A., Czemmel, S., et al.** (2013) The R2R3-MYB transcription factors MYB14 and MYB15 regulate stilbene biosynthesis in Vitis vinifera. *Plant Cell*, **25**, 4135–49.

**Hong, S., Chen, X., Jin, L. and Xiong, M.** (2013) Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.*, **41**, 1–15.

**Hrmova, M. and Fincher, G.** (1993) Purification and properties of three (1--\textgreater3)-beta-D-glucanase isoenzymes from young leaves of barley (Hordeum vulgare). *Biochem J*, **289 ( Pt 2**, 453–461.

**Huang, H.C., Niu, Y. and Qin, L.X.** (2015) Differential expression analysis for RNA-Seq: An overview of statistical methods and computational software. *Cancer Inform.*, **14**, 57–67.

**Hückelhoven, R.** (2007) Cell wall-associated mechanisms of disease resistance and susceptibility. *Annu. Rev. Phytopathol.*, **45**, 101–127.

**Hughes, R.K., Domenico, S. De and Santino, A.** (2009) Plant cytochrome CYP74 family: Biochemical features, endocellular localisation, activation mechanism in plant defence and improvements for industrial applications. *ChemBioChem*, **10**, 1122–1133.

**Hyma, K.E., Barba, P., Wang, M., Londo, J.P., Acharya, C.B., Mitchell, S.E., Sun, Q., Reisch, B. and Cadle-Davidson, L.** (2015) *Heterozygous Mapping Strategy (HetMappS) for High Resolution Genotyping-By-Sequencing Markers: A Case Study in Grapevine*,.

**Ilc, T., Halter, D., Miesch, L., et al.** (2016) A grapevine cytochrome P450 generates the precursor of wine lactone, a key odorant in wine. *New Phytol.*

**Iovene, M., Zhang, T., Lou, Q., Buell, C.R. and Jiang, J.** (2013) Copy number variation in potato - an asexually propagated autotetraploid species. *Plant J.*, **75**, 80–89.

**Irmler, S., Schröder, G., St-Pierre, B., Crouch, N.P., Hotze, M., Schmidt, J., Strack, D., Matern, U. and Schröder, J.** (2000) Indole alkaloid biosynthesis in Catharanthus roseus: New enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. *Plant J.*, **24**, 797–804.

**Ito, T. and Meyerowitz, E.M.** (2000) Overexpression of a gene encoding a cytochrome P450, CYP78A9, induces large and seedless fruit in arabidopsis. *Plant Cell*, **12**, 1541–50.

**Jacob, F., Vernaldi, S. and Maekawa, T.** (2013) Evolution and conservation of plant NLR functions. *Front. Immunol.*, **4**.

**Jacobs, Dry and Robinson** (1999) Induction of different pathogenesis-related cDNAs in grapevine infected with powdery mildew and treated with ethephon. *Plant Pathol.*, **48**, 325–336.

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

**Jeandet, P., Douillet-Breuil, A.-C., Bessis, R., Debord, S., Sbaghi, M. and Adrian, M.** (2002) Phytoalexins from the Vitaceae: biosynthesis, phytoalexin gene expression in transgenic plants, antifungal activity, and metabolism. *J. Agric. Food Chem.*, **50**, 2731–2741.

**Jiao, Y., Xu, W., Duan, D., Wang, Y. and Nick, P.** (2016) A stilbene synthase allele from a Chinese wild grapevine confers resistance to powdery mildew by recruiting salicylic acid signalling for efficient defence. *J. Exp. Bot.*, **67**, erw351.

**Jin, X., Feng, D., Wang, H. and Wang, J.** (2007) A novel tissue-specific plantain beta-1,3-glucanase gene that is regulated in response to infection by Fusarium oxysporum fsp. cubense. *Biotechnol. Lett.*, **29**, 1431–1437.

**Jones, J.D.G. and Dangl, J.L.** (2006) The plant immune system. *Nature*, **444**, 323–329.

**Jones, L., Riaz, S., Morales-Cruz, A., Amrine, K.C.H., McGuire, B., Gubler, W.D., Walker, M.A. and Cantu, D.** (2014) Adaptive genomic structural variation in the grape powdery mildew pathogen, Erysiphe necator. *BMC Genomics*, **15**, 1081.

**Jørgensen, K., Morant, A.V., Morant, M., Jensen, N.B., Olsen, C.E., Kannangara, R., Motawia, M.S., Møller, B.L. and Bak, S.** (2011) Biosynthesis of the cyanogenic glucosides linamarin and lotaustralin in cassava: isolation, biochemical characterization, and expression pattern of CYP71E7, the oxime-metabolizing cytochrome P450 enzyme. *Plant Physiol.*, **155**, 282–92.

**Karvonen, M.J. and Somersalo, O.** (1952) A note on the determination of fructose and glucose by the Somogyi method. *Ann. Med. Exp. Biol. Fenn.*, **30**, 31–34.

**Kauffmann, S., Legrand, M., Geoffroy, P. and Fritig, B.** (1987) Biological function of ;pathogenesis-related' proteins: four PR proteins of tobacco have 1,3-beta-glucanase activity. *EMBO J.*, **6**, 3209–3212.

**Keen, N.T., Yoshikawa, M. and Wang, M.C.** (1983) Phytoalexin Elicitor Activity of Carbohydrates from Phytophthora megasperma f.sp. glycinea and Other Sources. *Plant Physiol.*, **71**, 466–471.

**Kikuchi, T., Shibuya, H. and Jones, J.T.** (2005) Molecular and biochemical characterization of an endo-beta-1,3-glucanase from the pinewood nematode Bursaphelenchus xylophilus acquired by horizontal gene transfer from bacteria. *Biochem. J.*, **389**, 117–125.

**Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L.** (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

**Kim, H.-S., Park, K.-G., Baek, S.-B. and Kim, J.-G.** (2011) Inheritance of (1–3)(1–4)-beta-D-glucan content in barley (Hordeum vulgare L.). *J. Crop Sci. Biotechnol.*, **14**, 239–245.

**Kim, T.-S., He, Q., Kim, K.-W., et al.** (2016) Genome-wide resequencing of KRICE_CORE reveals their potential for future breeding, as well as functional and evolutionary studies in the post-genomic era. *BMC Genomics*, **17**, 408.

**Klarzynski, O., Plesse, B., Joubert, J.M., Yvin, J.C., Kopp, M., Kloareg, B. and Fritig, B.** (2000) Linear beta-1,3 glucans are elicitors of defense responses in tobacco. *Plant Physiol.*, **124**, 1027–1038.

**Kobayashi, S., Goto-yamamoto, N. and Hirochika, H.** (2004) Mutations in Grape Skin Color. , **304**, 8602.

**Kortekamp, A.** (2006) Expression analysis of defence-related genes in grapevine leaves after inoculation with a host and a non-host pathogen. *Plant Physiol. Biochem.*, **44**, 58–67.

**Kraus, P.F. and Kutchan, T.M.** (1995) Molecular cloning and heterologous expression of a cDNA encoding berbamunine synthase, a C--O phenol-coupling cytochrome P450 from the higher plant Berberis stolonifera. *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 2071–2075.

**Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A.** (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–45.

**Kumar, S., Stecher, G. and Tamura, K.** (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.*, **33**, msw054.

**Laemmli, U.K.** (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680–685.

**Lambert, C., Richard, T., Renouf, E., Bisson, J., Waffo-T??guo, P., Bordenave, L., Ollat, N., M??rillon, J.M. and Cluzet, S.** (2013) Comparative analyses of stilbenoids in canes of major Vitis vinifera L. cultivars. *J. Agric. Food Chem.*, **61**, 11392–11399.

**Langridge, P. and Fleury, D.** (2011) Making the most of "omics" for crop breeding. *Trends Biotechnol.*, **29**, 33–40.

**Larbat, R., Hehn, A., Hans, J., Schneider, S., Jugdé, H., Schneider, B., Matern, U. and Bourgaud, F.** (2009) Isolation and functional characterization of CYP71AJ4 encoding for the first P450 monooxygenase of angular furanocoumarin biosynthesis. *J. Biol. Chem.*, **284**, 4776–85.

**Larkin, M., Blackshields, G., Brown, N., et al.** (2007) ClustalW and ClustalX version 2. *Bioinformatics*, **23**, 2947–2948.

**Laudert, D., Pfannschmidt, U., Lottspeich, F., Holländer-Czytko, H. and Weiler, E.W.** (1996) Cloning, molecular and functional characterization of Arabidopsis thaliana allene oxide synthase (CYP 74), the first enzyme of the octadecanoid pathway to jasmonates. *Plant Mol. Biol.*, **31**, 323–335.

**Lee, S., Badieyan, S., Bevan, D.R., Herde, M., Gatz, C. and Tholl, D.** (2010) Herbivore-induced and floral homoterpene volatiles are biosynthesized by a single P450 enzyme (CYP82G1) in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 21205–10.

**Leister, D.** (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet.*, **20**, 116–122.

**Letunic, I. and Bork, P.** (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

**Leubner-Metzger, G.** (2005) beta-1,3-Glucanase gene expression in low-hydrated seeds as a mechanism for dormancy release during tobacco after-ripening. *Plant J.*, **41**, 133–145.

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

**Li, H., Xia, N. and Förstermann, U.** (2012) Cardiovascular effects and molecular targets of resveratrol. *Nitric Oxide*, **26**, 102–110.

**Licausi, F., Giorgi, F.M., Zenoni, S., Osti, F., Pezzotti, M. and Perata, P.** (2010) Genomic and transcriptomic analysis of the AP2/ERF superfamily in Vitis vinifera. *BMC Genomics*, **11**, 719.

**Lin, F. and Chen, X.M.** (2007) Genetics and molecular mapping of genes for race-specific all-stage resistance and non-race-specific high-temperature adult-plant resistance to stripe rust in spring wheat cultivar Alpowa. *Theor. Appl. Genet.*, **114**, 1277–1287.

**Liu, B., Xue, X., Cui, S., et al.** (2010) Cloning and characterization of a wheat beta-1,3-glucanase gene induced by the stripe rust pathogen Puccinia striiformis f. sp. tritici. *Mol. Biol. Rep.*, **37**, 1045–1052.

**Liu, J., Chen, N., Chen, F., Cai, B., Dal Santo, S., Tornielli, G.B., Pezzotti, M. and Cheng, Z.-M.M.** (2014) Genome-wide analysis and expression profile of the bZIP transcription factor gene family in grapevine (Vitis vinifera). *BMC Genomics*, **15**, 281.

**Liu, J., Liu, X., Dai, L. and Wang, G.** (2007) Recent Progress in Elucidating the Structure , Function and Evolution of Disease Resistance Genes in Plants. *J. Genet.*, **34**, 765–776.

**Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M. and Henrissat, B.** (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**, D490--495.

**Loon, L.C. van, Rep, M. and Pieterse, C.M.J.** (2006) Significance of inducible defense-related proteins in infected plants. *Annu. Rev. Phytopathol.*, **44**, 135–162.

**Loon, L.C. van and Strien, E.A. van** (1999) The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins. *Physiol. Mol. Plant Pathol.*, **55**, 85–97.

**Love, M.I., Huber, W. and Anders, S.** (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

**Luo, P., Wang, Y.H., Wang, G.D., Essenberg, M. and Chen, X.Y.** (2001) Molecular cloning and functional identification of (+)-delta-cadinene-8-hydroxylase, a cytochrome P450 mono-oxygenase (CYP706B1) of cotton sesquiterpene biosynthesis. *Plant J.*, **28**, 95–104.

**Luo, S., Yu, J.A. and Song, Y.S.** (2016) Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads. , 1–21.

**Luo, S., Zhang, Y., Hu, Q., Chen, J., Li, K., Lu, C., Liu, H., Wang, W. and Kuang, H.** (2012) Dynamic Nucleotide-Binding Site and Leucine-Rich Repeat-Encoding Genes in the Grass Family. *Plant Physiol.*, **159**, 197–210.

**Lupas, A., Dyke, M. Van and Stock, J.** (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–4.

**Luz, L. de A., Silva, M.C.C., Ferreira, R. da S., Santana, L.A., Silva-Lucca, R.A., Mentele, R., Oliva, M.L.V., Paiva, P.M.G. and Coelho, L.C.B.B.** (2013) Structural characterization of coagulant Moringa oleifera Lectin and its effect on hemostatic parameters. *Int. J. Biol. Macromol.*, **58**, 31–38.

**Magome, H., Nomura, T., Hanada, A., et al.** (2013) CYP714B1 and CYP714B2 encode gibberellin 13-oxidases that reduce gibberellin activity in rice. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 1947–52.

**Mahanil, S., Ramming, D., Cadle-Davidson, M., Owens, C., Garris, A., Myles, S. and Cadle-Davidson, L.** (2012) Development of marker sets useful in the early selection of Ren4 powdery mildew resistance and seedlessness for table and raisin grape breeding. *Theor. Appl. Genet.*, **124**, 23–33.

**Majoul, T., Bancel, E., Tribo"\i, E., Hamida, J. Ben and Branlard, G.** (2003) Proteomic analysis of the effect of heat stress on hexaploid wheat grain: Characterization of heat-responsive proteins from total endosperm. *Proteomics*, **3**, 175–183.

**Malacarne, G., Perazzolli, M., Cestaro, A., Sterck, L., Fontana, P., Peer, Y. van de, Viola, R., Velasco, R. and Salamini, F.** (2012) Deconstruction of the (paleo)polyploid grapevine genome based on the analysis of transposition events involving NBS resistance genes. *PLoS One*, **7**.

**Martin, D.M., Aubourg, S., Schouwey, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T. and Bohlmann, J.** (2010) Functional annotation, genome organization and phylogeny of the grapevine (Vitis vinifera) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.*, **10**, 226.

**Martin, G.B., Bogdanove, A.J. and Sessa, G.** (2003) Understanding the functions of plant disease resistance proteins. *Annu. Rev. Plant Biol.*, **54**, 23–61.

**Matus, J.T., Aquea, F., Espinoza, C., et al.** (2014) Inspection of the grapevine BURP superfamily highlights an expansion of RD22 genes with distinctive expression features in berry development and ABA-mediated stress responses. *PLoS One*, **9**.

**McDowell, J.M. and Simon, S.A.** (2006) Recent insights into R gene evolution. *Mol. Plant Pathol.*, **7**, 437–448.

**McHale, L., Tan, X., Koehl, P. and Michelmore, R.W.** (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biol.*, **7**, 212.

**Meyers, B.C., Dickerman, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sobral, B.W. and Young, N.D.** (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.*, **20**, 317–332.

**Meyers, B.C., Kozik, A., Griego, A., Kuang, H. and Michelmore, R.W.** (2003) Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis. *Plant Cell Online*, **15**, 809–834.

**Mica, E., Piccolo, V., Delledonne, M., et al.** (2009) High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera. *BMC Genomics*, **10**, 558.

**Michelmore, R.W. and Meyers, B.C.** (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.*, **8**, 1113–1130.

**Miettinen, K., Dong, L., Navrot, N., et al.** (2014) The seco-iridoid pathway from Catharanthus roseus. *Nat. Commun.*, **5**, 3606.

**Migicovsky, Z., Sawler, J., Money, D., et al.** (2016) Genomic ancestry estimation quantifies use of wild species in grape breeding. *BMC Genomics*, **17**, 478.

**Mikkelsen, M.D., Hansen, C.H., Wittstock, U. and Halkier, B. a** (2000) Cytochrome P450 CYP79B2 from Arabidopsis catalyzes the conversion of tryptophan to indole-3-acetaldoxime, a precursor of indole glucosinolates and indole-3-acetic acid. *J. Biol. Chem.*, **275**, 33712–7.

**Miller, J.L., Grant, P.A.** (2013) The role of DNA methylation and histone modifications in transcriptional regulation in humans. *Subcell Biochem*, **61**, 289–317.

**Miller, M.A., Pfeiffer, W. and Schwartz, T.** (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *2010 Gateway Computing Environments Workshop (GCE)*. IEEE, pp. 1–8.

**Milli, A., Cecconi, D., Bortesi, L., et al.** (2012) Proteomic analysis of the compatible interaction between Vitis vinifera and Plasmopara viticola. *J. Proteomics*, **75**, 1284–1302.

**Mittler, R., Vanderauwera, S., Suzuki, N., Miller, G., Tognetti, V.B., Vandepoele, K., Gollery, M., Shulaev, V. and Breusegem, F. Van** (2011) ROS signaling: the new wave? *Trends Plant Sci.*, **16**, 300–309.

**Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., Shinozaki, K. and Tran, L.S.P.** (2013) TreeTFDB: An integrative database of the transcription factors from six economically important tree crops for functional predictions and comparative and functional genomics. *DNA Res.*, **20**, 151–162.

**Mondragón-Palomino, M., Meyers, B.C., Michelmore, R.W. and Gaut, B.S.** (2002) Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. *Genome Res.*, **12**, 1305–1315.

**Moretto, M., Sonego, P., Pilati, S., et al.** (2016) VESPUCCI: Exploring Patterns of Gene Expression in Grapevine. *Front. Plant Sci.*, **7**, 1–11.

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

**Morohashi, Y. and Matsushima, H.** (2000) Development of beta-1,3-glucanase activity in germinated tomato seeds. *J. Exp. Bot.*, **51**, 1381–1387.

**Moses, T., Pollier, J., Faizal, A., Apers, S., Pieters, L., Thevelein, J.M., Geelen, D. and Goossens, A.** (2014) Unravelling the Triterpenoid Saponin Biosynthesis of the African Shrub Maesa lanceolata. *Mol. Plant*.

**Mutterer, J. and Zinck, E.** (2013) Quick-and-clean article figures with FigureJ. *J. Microsc.*, **252**, 89–91.

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

**Myles, S., Boyko, A.R., Owens, C.L., 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, G.Y., Buckler, E. and Ware, D.** (2010) Rapid genomic characterization of the genus Vitis. *PLoS One*, **5**.

**Nafisi, M., Goregaoker, S., Botanga, C.J., Glawischnig, E., Olsen, C.E., Halkier, B. a and Glazebrook, J.** (2007) Arabidopsis cytochrome P450 monooxygenase 71A13 catalyzes the conversion of indole-3-acetaldoxime in camalexin synthesis. *Plant Cell*, **19**, 2039–52.

**Naithani, S., Raja, R., Waddell, E.N., Elser, J., Gouthu, S., Deluc, L.G. and Jaiswal, P.** (2014) VitisCyc: a metabolic pathway knowledgebase for grapevine (Vitis vinifera). *Front. Plant Sci.*, **5**, 644.

**Nascimento-Gavioli, M.C.A., Agapito-Tenfen, S.Z., Nodari, R.O., Welter, L.J., Sanchez Mora, F.D., Saifert, L., Silva, A.L. da and Guerra, M.P.** (2016) Proteome of Plasmopara viticola-infected Vitis vinifera provides insights into grapevine Rpv1/Rpv3 pyramided resistance to downy mildew. *J Proteomics*.

**Nelson, D. and Werck-Reichhart, D.** (2011) A P450-centric view of plant evolution. *Plant J.*, **66**, 194–211.

**Nelson, D.R.** (2009) The cytochrome P450 homepage. *Hum. Genomics*, **4**, 59–65.

**Nelson, D.R., Ming, R., Alam, M. and Schuler, M.A.** (2008) Comparison of Cytochrome P450 Genes from Six Plant Genomes. *Trop. Plant Biol.*, **1**, 216–235.

**Nelson, D.R., Schuler, M. a, Paquette, S.M., Werck-Reichhart, D. and Bak, S.** (2004) Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. *Plant Physiol.*, **135**, 756–772.

**Neuhaus, J.M., Flores, S., Keefe, D., Ahl-Goy, P. and Meins F, J.** (1992) The function of vacuolar beta-1,3-glucanase investigated by antisense transformation. Susceptibility of transgenic Nicotiana sylvestris plants

to Cercospora nicotianae infection. *Plant Mol. Biol.*, **19**, 803–813.

**Neuhoff, V., Arold, N., Taube, D. and Ehrhardt, W.** (1988) Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis*, **9**, 255–262.

**Nishimura, M.T., Stein, M., Hou, B.-H., Vogel, J.P., Edwards, H. and Somerville, S.C.** (2003) Loss of a callose synthase results in salicylic acid-dependent disease resistance. *Science (80-. ).*, **301**, 969–972.

**Nützmann, H.-W. and Osbourn, A.** (2014) Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.*, **26**, 91–9.

**Nwafor, C.C., Gribaudo, I., Schneider, A., Wehrens, R., Grando, M.S. and Costantini, L.** (2014) Transcriptome analysis during berry development provides insights into co-regulated and altered gene expression between a seeded wine grape variety and its seedless somatic variant. *BMC Genomics*, **15**, 1030.

**Oide, S., Bejai, S., Staal, J., Guan, N., Kaliff, M. and Dixelius, C.** (2013) A novel role of PR2 in abscisic acid (ABA) mediated, pathogen-induced callose deposition in Arabidopsis thaliana. *New Phytol.*, **200**, 1187–1199.

**Ono, E., Nakai, M., Fukui, Y., et al.** (2006) Formation of two methylenedioxy bridges by a Sesamum CYP81Q protein yielding a furofuran lignan, (+)-sesamin. *Proc. Natl. Acad. Sci.*, **103**, 10116–10121.

**Østergaard, O., Melchior, S., Roepstorff, P. and Svensson, B.** (2002) Initial proteome analysis of mature barley seeds and malt. *Proteomics*, **2**, 733–739.

**Paim Pinto, D.L., Brancadoro, L., Dal Santo, S., Lorenzis, G. De, Pezzotti, M., Meyers, B.C., Pè, M.E. and Mica, E.** (2016) The Influence of Genotype and Environment on Small RNA Profiles in Grapevine Berry. *Front. Plant Sci.*, **7**.

**Panchy, N., Lehti-Shiu, M.D. and Shiu, S.-H.** (2016) Evolution of gene duplication in plants. *Plant Physiol.*, **171**, pp.00523.2016.

**Pantaleo, V., Vitali, M., Boccacci, P., Miozzi, L., Cuozzo, D., Chitarra, W., Mannini, F., Lovisolo, C. and Gambino, G.** (2016) Novel functional microRNAs from virus-free and infected Vitis vinifera plants under water stress. *Sci. Rep.*, **6**, 20167.

**Parage, C., Tavares, R., Rety, S., et al.** (2012) Structural, Functional, and Evolutionary Analysis of the Unusually Large Stilbene Synthase Gene Family in Grapevine. *Plant Physiol.*, **160**, 1407–1419.

**Park, J.H., Halitschke, R., Kim, H.B., Baldwin, I.T., Feldmann, K. a. and Feyereisen, R.** (2002) A knock-out mutation in allene oxide synthase results in male sterility and defective wound signal transduction in Arabidopsis due to a block in jasmonic acid biosynthesis. *Plant J.*, **31**, 1–12.

**Pauli, H.H. and Kutchan, T.M.** (1998) Molecular cloning and functional heterologous expression of two alleles encoding (S)-N-methylcoclaurine 3'-hydroxylase (CYP80B1), a new methyl jasmonate-inducible cytochrome P-450-dependent mono-oxygenase of benzylisoquinoline alkaloid biosynthesis. *Plant J.*, **13**, 793–801.

**Peck, S.C.** (2005) Update on proteomics in Arabidopsis. Where do we go from here? *Plant Physiol*, **138**, 591–599.

**Perazzolli, M., Moretto, M., Fontana, P., Ferrarini, A., Velasco, R., Moser, C., Delledonne, M. and Pertot, I.** (2012b) Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. *BMC Genomics*, **13**, 660.

**Peressotti, E., Wiedemann-Merdinoglu, S., Delmotte, F., Bellin, D., Gaspero, G. Di, Testolin, R., Merdinoglu, D. and Mestre, P.** (2010) Breakdown of resistance to grapevine downy mildew upon limited deployment of a resistant variety. *BMC Plant Biol.*, **10**, 147.

**Pervaiz, T., Haifeng, J., Salman Haider, M., Cheng, Z., Cui, M., Wang, M., Cui, L., Wang, X. and Fang, J.** (2016) Transcriptomic Analysis of Grapevine (cv. Summer Black) Leaf, Using the Illumina Platform. *PLoS One*, **11**, e0147369.

**Pfalz, M., Vogel, H. and Kroymann, J.** (2009) The Gene Controlling the Indole Glucosinolate Modifier1

Quantitative Trait Locus Alters Indole Glucosinolate Structures and Aphid Resistance in Arabidopsis. *Plant Cell*, **21**, 985–999.

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

**Pinosio, S., Giacomello, S., Faivre-Rampant, P., et al.** (2016) Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Mol. Biol. Evol.*, **33**, 2706–19.

**Porter, B.W., Paidi, M., Ming, R., Alam, M., Nishijima, W.T. and Zhu, Y.J.** (2009) Genome-wide analysis of Carica papaya reveals a small NBS resistance gene family. *Mol. Genet. Genomics*, **281**, 609–626.

**Potenza, E., Racchi, M.L., Sterck, L., Coller, E., Asquini, E., Tosatto, S.C.E., Velasco, R., Peer, Y. Van de and Cestaro, A.** (2015) Exploration of alternative splicing events in ten different grapevine cultivars. *BMC Genomics*, **16**, 706.

**Pulvirenti, A., Giugno, R., Distefano, R., et al.** (2015) A knowledge base for Vitis vinifera functional analysis. *BMC Syst. Biol.*, **9 Suppl 3**, S5.

**Quinlan, A.R. and Hall, I.M.** (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–2.

**R Development Core Team, R.** (2011) R: A Language and Environment for Statistical Computing R. D. C. Team, ed. *R Found. Stat. Comput.*, **1**, 409.

**Ramos, M.J.N., Coito, J.L., Silva, H.G., Cunha, J., Costa, M.M.R. and Rocheta, M.** (2014b) Flower development and sex specification in wild grapevine. *BMC Genomics*, **15**, 1095.

**Reina-Pinto, J.J. and Yephremov, A.** (2009) Surface lipids and plant defenses. *Plant Physiol. Biochem.*, **47**, 540–549.

**Renaud, S. and Lorgeril, M. de** (1992) Wine, alcohol, platelets, and the French paradox for coronary heart disease. *Lancet*, **339**, 1523–1526.

**Riaz, S., Tenscher, a C., Smith, B.P., Ng, D. a and Walker, M. a** (2008) Use of SSR markers to assess identity, pedigree, and diversity of cultivated muscadine grapes. *J. Am. Soc. Hortic. Sci.*, **133**, 559–568.

**Richly, E., Kurth, J. and Leister, D.** (2002) Mode of amplification and reorganization of resistance genes during recent Arabidopsis thaliana evolution. *Mol. Biol. Evol.*, **19**, 76–84.

**Rienth, M., Torregrosa, L., Luchaire, N., Chatbanyong, R., Lecourieux, D., Kelly, M.T. and Romieu, C.** (2014) Day and night heat stress trigger different transcriptomic responses in green and ripening grapevine (vitis vinifera) fruit. *BMC Plant Biol.*, **14**, 108.

**Robinson, M.D., McCarthy, D.J. and Smyth, G.K.** (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

**Rose, J.K.C., Ham, K.-S., Darvill, A.G. and Albersheim, P.** (2002) Molecular cloning and characterization of glucanase inhibitor proteins: coevolution of a counterdefense mechanism by plant pathogens. *Plant Cell*, **14**, 1329–1345.

**Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B.** (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

**Salekdeh, G.H., Siopongco, J., Wade, L.J., Ghareyazie, B. and Bennett, J.** (2002) Proteomic analysis of rice leaves during drought stress and recovery. *Proteomics*, **2**, 1131–1145.

**Santos-Rosa, M., Poutaraud, A., Merdinoglu, D. and Mestre, P.** (2008) Development of a transient expression system in grapevine via agro-infiltration. *Plant Cell Rep.*, **27**, 1053–1063.

**Schiex, T., Moisan, A. and Rouzé, P.** (2001) EugÉne: An eukaryotic gene finder that combines several sources of evidence. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 111–125.

**Schneider, C.A., Rasband, W.S. and Eliceiri, K.W.** (2012) NIH Image to ImageJ: 25 years of image analysis.

*Nat Meth*, **9**, 671–675.

**Scofield, S.R., Tobias, C.M., Rathjen, J.P., Chang, J.H., Lavelle, D.T., Michelmore, R.W. and Staskawicz, B.J.** (1996) Molecular Basis of Gene-for-Gene Specificity in Bacterial Speck Disease of Tomato. *Science (80-. ).*, **274**, 2063–2065.

**Seki, H., Sawai, S., Ohyama, K., et al.** (2011) Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. *Plant Cell*, **23**, 4112–23.

**Seo, J., Gordish-Dressman, H. and Hoffman, E.P.** (2006) An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics*, **22**, 808–14.

**Seshan, V.E. and Olshen, A.B.** (2014) DNAcopy : A Package for Analyzing DNA Copy Data. *Bioconductor Vignette*, 1–7.

**Shangguan, L., Kayesh, E., Leng, X., Sun, X., Korir, N.K., Mu, Q. and Fang, J.** (2013) Whole genome identification and analysis of FK506-binding protein family genes in grapevine (Vitis vinifera L.). *Mol. Biol. Rep.*, **40**, 4015–4031.

**Shao, Z.-Q., Xue, J.-Y., Wu, P., Zhang, Y.-M., Wu, Y., Hang, Y.-Y., Wang, B. and Chen, J.-Q.** (2016) Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiol.*, pp.01487.2015.

**Shen, J., Araki, H., Chen, L., Chen, J.Q. and Tian, D.** (2006) Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in Arabidopsis thaliana. *Genetics*, **172**, 1243–1250.

**Shiratake, K. and Suzuki, M.** (2016) Omics studies of citrus, grape and rosaceae fruit trees. *Breed Sci*, **66**, 122–138.

**Sievers, F., Wilm, A., Dineen, D., et al.** (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

**Silva, C. Da, Zamperin, G., Ferrarini, A., 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.

**Skylas, D.J., Copeland, L., Rathmell, W.G. and Wrigley, C.W.** (2001) The wheat-grain proteome as a basis for more efficient cultivar identification. *Proteomics*, **1**, 1542–1546.

**Slater, G.S.C. and Birney, E.** (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

**Somogyi, M.** (1952) Notes on sugar determination. *J. Biol. Chem.*, **195**, 19–23.

**Sonnhammer, E.L.L. and Durbin, R.** (1995) A Dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, 1–10.

**Sparvoli, F., Martin, C., Scienza, A., Gavazzi, G., Tonelli, C. and Celoria, V.** (1994) Cloning and molecular analysis of structural genes involved in flavonoid and stilbene biosynthesis in grape (Vitis vinifera L.). *Plant Mol Biol*, **75969**, 743–755.

**Springer, N.M., Ying, K., Fu, Y., et al.** (2009) Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genet.*, **5**, e1000734.

**Stamatakis, A.** (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

**Stintzi, A., Heitz, T., Prasad, V., Wiedemann-Merdinoglu, S., Kauffmann, S., Geoffroy, P., Legrand, M. and Fritig, B.** (1993) Plant "pathogenesis-related" proteins and their role in defense against pathogens. *Biochimie*, **75**, 687–706.

**Sun, X., Fan, G., Su, L., Wang, W., Liang, Z., Li, S. and Xin, H.** (2015) Identification of cold-inducible microRNAs in grapevine. *Front. Plant Sci.*, **6**, 1–13.

**Sun, X., Xie, Z., Zhang, C., Mu, Q., Wu, W., Wang, B. and Fang, J.** (2016) A characterization of grapevine of GRAS domain transcription factor gene family. *Funct. Integr. Genomics*.

**Swaminathan, S., Morrone, D., Wang, Q., Fulton, D.B. and Peters, R.J.** (2009) CYP76M7 is an ent-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell*, **21**, 3315–25.

**Sweetman, C., Wong, D.C., Ford, C.M. and Drew, D.P.** (2012) Transcriptome analysis at four developmental stages of grape berry (Vitis vinifera cv. Shiraz) provides insights into regulated and coordinated gene expression. *BMC Genomics*, **13**, 691.

**Takeuchi, Y., Yoshikawa, M., Takeba, G., Tanaka, K., Shibata, D. and Horino, O.** (1990) Molecular Cloning and Ethylene Induction of mRNA Encoding a Phytoalexin Elicitor-Releasing Factor, beta-1,3-Endoglucanase, in Soybean. *Plant Physiol.*, **93**, 673–682.

**Takos, A.M. and Rook, F.** (2012) Why biosynthetic genes for chemical defense compounds cluster. *Trends Plant Sci.*, **17**, 383–388.

**Tavares, S., Wirtz, M., Beier, M.P., Bogs, J., Hell, R. and Amâncio, S.** (2015) Characterization of the serine acetyltransferase gene family of Vitis vinifera uncovers differences in regulation of OAS synthesis in woody plants. *Front. Plant Sci.*, **6**, 74.

**Teoh, K.H., Polichuk, D.R., Reed, D.W., Nowak, G. and Covello, P.S.** (2006) Artemisia annua L. (Asteraceae) trichome-specific cDNAs reveal CYP71AV1, a cytochrome P450 with a key role in the biosynthesis of the antimalarial sesquiterpene lactone artemisinin. *FEBS Lett.*, **580**, 1411–6.

**The angiosperm phylogeny group** (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.*, **161**, 105–121.

**The Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–814.

**Thomson, D.W., Bracken, C.P. and Goodall, G.J.** (2011) Experimental strategies for microRNA target identification. *Nucleic Acids Res.*, **39**, 6845–6853.

**Tisserant, L.-P., Aziz, A., Jullian, N., Jeandet, P., Clément, C., Courot, E. and Boitel-Conti, M.** (2016) Enhanced Stilbene Production and Excretion in Vitis vinifera cv Pinot Noir Hairy Root Cultures. *Molecules*, **21**, 1703.

**Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M.J. van, Salzberg, S.L., Wold, B.J. and Pachter, L.** (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–5.

**Vannozzi, A., Dry, I.B., Fasoli, M., Zenoni, S. and Lucchin, M.** (2012a) Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. *BMC Plant Biol.*, **12**, 130.

**Vannozzi, A., Dry, I.B., Fasoli, M., Zenoni, S. and Lucchin, M.** (2012b) Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. *BMC Plant Biol.*, **12**, 130.

**Velasco, R., Zharkikh, A., Troggio, M., et al.** (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, **2**.

**Venturini, L., Ferrarini, A., Zenoni, S., Tornielli, G.B., Fasoli, M., Dal Santo, S., et al.** (2013) De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. *BMC Genomics*, **14**, 41.

**Vitulo, N., Forcato, C., Carpinelli, E., 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.

**Vivier, M.A. and Pretorius, I.S.** (2002) Genetically tailored grapevines for the wine industry. *Trends Biotechnol.*, **20**, 472–478.

**Wan, H., Yuan, W., Ye, Q., et al.** (2012) Analysis of TIR- and non-TIR-NBS-LRR disease resistance gene analogous in pepper: characterization, genetic variation, functional divergence and expression patterns. *BMC Genomics*, **13**, 502.

**Wan, Y., Schwaninger, H.R., Baldo, A.M., Labate, J.A., Zhong, G.-Y. and Simon, C.J.** (2013) A

phylogenetic analysis of the grape genus (Vitis L.) reveals broad reticulation and concurrent diversification during neogene and quaternary climate change. *BMC Evol. Biol.*, **13**, 141.

**Wang, C., Han, J., Liu, C., Kibet, K., Kayesh, E., Shangguan, L., Li, X. and Fang, J.** (2012) Identification of microRNAs from Amur grape (vitis amurensis Rupr.) by deep sequencing and analysis of microRNA variations with bioinformatics. *BMC Genomics*, **13**, 122.

**Wang, C., Wang, X., Kibet, N.K., Song, C., Zhang, C., Li, X., Han, J. and Fang, J.** (2011) Deep sequencing of grapevine flower and berry short RNA library for discovery of novel microRNAs and validation of precise sequences of grapevine microRNAs deposited in miRBase. *Physiol. Plant.*, **143**, 64–81.

**Wang, M., Vannozzi, A., Wang, G., Liang, Y.-H., Tornielli, G.B., Zenoni, S., Cavallini, E., Pezzotti, M. and Cheng, Z.-M. (Max)** (2014) Genome and transcriptome analysis of the grapevine (Vitis vinifera L.) WRKY gene family. *Hortic. Res.*, **1**, 16.

**Wang, Q., Hillwig, M.L., Okada, K., Yamazaki, K., Wu, Y., Swaminathan, S., Yamane, H. and Peters, R.J.** (2012) Characterization of CYP76M5-8 indicates metabolic plasticity within a plant biosynthetic gene cluster. *J. Biol. Chem.*, **287**, 6159–68.

**Wang, Y., Xiong, G., Hu, J., et al.** (2015) Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet.*, **47**, 944–948.

**Ward, E.R., Payne, G.B., Moyer, M.B., et al.** (1991) Differential Regulation of beta-1,3-Glucanase Messenger RNAs in Response to Pathogen Infection. *Plant Physiol.*, **96**, 390–397.

**Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J.** (2009) Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

**Wellesen, K., Durst, F., Pinot, F., Benveniste, I., Nettesheim, K., Wisman, E., Steiner-Lange, S., Saedler, H. and Yephremov, A.** (2001) Functional analysis of the LACERATA gene of Arabidopsis provides evidence for different roles of fatty acid ω-hydroxylation in development. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 9694–9699.

**Wen, J., Xiong, Z., Nie, Z.L., et al.** (2013) Transcriptome Sequences Resolve Deep Relationships of the Grape Family. *PLoS One*, **8**, 1–9.

**Wessels, J.G.H.** (1994) Developmental Regulation of Fungal Cell Wall Formation. *Annu. Rev. Phytopathol.*, **32**, 413–437.

**Wickham, H.** (2009) *ggplot2*,.

**Widholm, J.M.** (1972) The use of fluorescein diacetate and phenosafranine for determining viability of cultured plant cells. *Stain Technol*, **47**, 189–194.

**Wienkoop, S., Baginsky, S. and Weckwerth, W.** (2010) Arabidopsis thaliana as a model organism for plant proteome research. *J Proteomics*, **73**, 2239–2248.

**Wijk, K.J. van** (2001) Challenges and prospects of plant proteomics. *Plant Physiol*, **126**, 501–508.

**Wittstock, U.** (2000) Cytochrome P450 CYP79A2 from Arabidopsis thaliana L. Catalyzes the Conversion of L-Phenylalanine to Phenylacetaldoxime in the Biosynthesis of Benzylglucosinolate. *J. Biol. Chem.*, **275**, 14659–14666.

**Wong, D.C.J., Schlechter, R., Vannozzi, A., Höll, J., Hmmam, I., Bogs, J., Tornielli, G.B., Castellarin, S.D. and Matus, J.T.** (2016) A systems-oriented analysis of the grapevine R2R3-MYB transcription factor family uncovers new insights into the regulation of stilbene accumulation. *DNA Res.*, **0**, dsw028.

**Wong, D.C.J., Sweetman, C., Drew, D.P. and Ford, C.M.** (2013) VTCdb: a gene co-expression database for the crop species Vitis vinifera (grapevine). *BMC Genomics*, **14**, 882.

**Worrall, D., Hird, D.L., Hodge, R., Paul, W., Draper, J. and Scott, R.** (1992) Premature dissolution of the microsporocyte callose wall causes male sterility in transgenic tobacco. *Plant Cell*, **4**, 759–771.

**Wu, T.D. and Nacu, S.** (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

**Wu, T.D. and Watanabe, C.K.** (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–75.

**Xie, Y.-R., Raruang, Y., Chen, Z.-Y., Brown, R.L. and Cleveland, T.E.** (2015) ZmGns, a maize class I β-1,3-glucanase, is induced by biotic stresses and possesses strong antimicrobial activity. *J Integr Plant Biol*, **57**, 271–283.

**Xin, H., Zhu, W., Wang, L., et al.** (2013) Genome wide transcriptional profile analysis of Vitis amurensis and Vitis vinifera in response to cold stress. *PLoS One*, **8**, e58740.

**Xu, W., Li, R., Zhang, N., Ma, F., Jiao, Y. and Wang, Z.** (2014) Transcriptome profiling of Vitis amurensis, an extremely cold-tolerant Chinese wild Vitis species, reveals candidate genes and events that potentially connected to cold stress. *Plant Mol. Biol.*, **86**, 527–541.

**Yang, S., Feng, Z., Zhang, X., Jiang, K., Jin, X., Hang, Y., Chen, J.Q. and Tian, D.** (2006) Genome-wide investigation on the genetic variations of rice disease resistance genes. *Plant Mol. Biol.*, **62**, 181–193.

**Yang, S., Fresnedo-Ramírez, J., Sun, Q., et al.** (2016) Next generation mapping of enological traits in an F2 interspecific grapevine hybrid family. *PLoS One*, **11**, 1–19.

**Yang, S., Zhang, X., Yue, J.-X., Tian, D. and Chen, J.-Q.** (2008) Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics*, **280**, 187–198.

**Yang, X. and Wang, J.** (2015) Genome-wide analysis of NBS-LRR genes in sorghum genome revealed several events contributing to NBS-LRR gene evolution in grass species. *Evol. Bioinforma.*, **12**, 9–21.

**Yang, Y., He, M., Zhu, Z., Li, S., Xu, Y., Zhang, C., Singer, S.D. and Wang, Y.** (2012) Identification of the dehydrin gene family from grapevine species and analysis of their responsiveness to various forms of abiotic and biotic stress. *BMC Plant Biol.*, **12**, 140.

**Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J.** (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

**Young, N.D., Zhou, P. and Silverstein, K.A.T.** (2016) Exploring structural variants in environmentally sensitive gene families. *Curr. Opin. Plant Biol.*, **30**, 19–24.

**Yu** (2013) Callose Synthase Family Genes Involved in the Grapevine Defense Response to Downy Mildew Disease. *Phytopathol.*, 56–64.

**Yuan, B., Zhai, C., Wang, W., Zeng, X., Xu, X., Hu, H., Lin, F., Wang, L. and Pan, Q.** (2011) The Pik-p resistance to Magnaporthe oryzae in rice is mediated by a pair of closely linked CC-NBS-LRR genes. *Theor. Appl. Genet.*, **122**, 1017–1028.

**Yue, J.X., Meyers, B.C., Chen, J.Q., Tian, D. and Yang, S.** (2012) Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes. *New Phytol.*, **193**, 1049–1063.

**Zamboni, A., Carli, M. Di, Guzzo, F., et al.** (2010) Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks. *Plant Physiol*, **154**, 1439–1459.

**Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., Bellin, D., Pezzotti, M. and Delledonne, M.** (2010) Characterization of Transcriptional Complexity during Berry Development in Vitis vinifera Using RNA-Seq. *Plant Physiol.*, **152**, 1787–1795.

**Zhang, K., Han, Y.-T., Zhao, F.-L., Hu, Y., Gao, Y.-R., Ma, Y.-F., Zheng, Y., Wang, Y.-J. and Wen, Y.-Q.** (2015) Genome-wide Identification and Expression Analysis of the CDPK Gene Family in Grape, Vitis spp. *BMC Plant Biol.*, **15**, 164.

**Zhang, Z., Qi, S., Tang, N., et al.** (2014) Discovery of Replicating Circular RNAs by RNA-Seq and Computational Algorithms. *PLoS Pathog.*, **10**.

**Zhao, S., Guo, Y., Sheng, Q. and Shyr, Y.** (2014) Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinformatics*, **15**, P16.

**Zhao, T., Xia, H., Liu, J. and Ma, F.** (2014) The gene family of dehydration responsive element-binding transcription factors in grape (Vitis vinifera): Genome-wide identification and analysis, expression profiles, and involvement in abiotic stress resistance. *Mol. Biol. Rep.*, **41**, 1577–1590.

**Zheng, F., Wu, H., Zhang, R., Li, S., He, W., Wong, F.-L., Li, G., Zhao, S. and Lam, H.-M.** (2016) Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family. *BMC Genomics*, **17**, 402.

**Zmieńko, A., Samelak, A., Kozłowski, P. and Figlerowicz, M.** (2014) Copy number polymorphism in plant genomes. *Theor. Appl. Genet.*, **127**, 1–18.

**Zohary, D. and Hopf, M.** (1973) Domestication of Pulses in the Old World. *Science (80-. ).*, **182**, 887–894.

**Zohary, D. and Spiegel-Roy, P.** (1972) Beginnings of Fruit Growing in the Old World. *Science (80-. ).*, **187**, 319–327.