HAL
open science

La sélection génomique appliquée a l'espece Vitis vinifera L. subsp. vinifera, évaluation et utilisation

Agota Fodor

To cite this version:

Agota Fodor. La sélection génomique appliquée a l'espece Vitis vinifera L. subsp. vinifera, évaluation et utilisation. Amélioration des plantes. Ecole nationale superieure agronomique de montpellier - AGRO M, 2013. Français. NNT : . tel-01001690

HAL Id: tel-01001690

https://theses.hal.science/tel-01001690v1

Submitted on 5 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

Docteur

Délivré par le

Centre international d'études supérieures en sciences agronomiques

Montpellier

Préparée au sein de l'école doctorale SIBAGHE
Et de l'unité de recherche UMT Geno-Vigne, IFV-INRA-Montpellier Supagro

Spécialité : Evolution, Ecologie, Ressources Génétiques, Paléontologie

Présentée par **Agota FODOR**

La sélection génomique appliquée à l'espèce *Vitis vinifera* L. *subsp vinifera*, évaluation et utilisation

Soutenue le 16 décembre 2013 devant le jury composé de

Mr Charles-Eric DUREL, INRA Angers	Rapporteur
Mr Gilles CHARMET, INRA Clermont-Ferrand	Rapporteur
Mr José QUERO-GARCIA, INRA Bordeaux	Examinateur
Mme Dominique THIS, Montpellier SupAgro	Examinateur
Mr Christophe SCHNEIDER, INRA, Colmar	Examinateur
Mr Patrice THIS, INRA Montpellier	Directeur de thèse
Mr Loïc Le CUNFF, IFV UMT Géno-Vigne® Montpellier	Invité

Je dédie ces travaux à mon Grand père, Dr Attila Borhidi.

A dokroti munkámat Nagypapámnak, Dr Borhidi Attilának ajánlom.

REMERCIEMENTS

Je voudrais remercier tous les membres de mon jury, Charles-Eric Durel, Gilles Charmet, José Quero-Garcia, Dominique This et Christophe Schneider d'avoir accepté d'évaluer mon travail.

Merci à l'ANRT, à Jean-Pierre Van Ruyskensvelde et aux responsables du projet CAS DAR pour la définition et le financement de cette thèse CIFRE.

Merci beaucoup à Patrice This, mon directeur de thèse, qui m'a accueillie dans son équipe et qui m'a fait confiance durant ces quatre dernières années. Merci pour ta patience et pour le raisonnement scientifique que tu m'as montré, et qui m'a beaucoup aidé lors de la rédaction d'articles et de la thèse en général.

Merci à Loïc Le Cunff, mon encadrant de l'Institut Français de la Vigne et du Vin (IFV) de m'avoir fait confiance en m'accordant ce sujet de thèse. Merci pour le temps et l'énergie que tu as investi pour le bon déroulement de ce projet de sa conception à sa réalisation.

Un grand merci à Laurent Audeguin, l'animateur de l'UMT Geno-Vigne, pour avoir eu un regard bienveillant sur mes travaux de thèse. Köszönöm Sylvie d'avoir toujours eu quelques gentils mots pour moi et un souvenir à partager sur la Hongrie. Merci pour ton aide et celle d'Isabelle, dans l'administration. Merci à Delphine qui m'a aussi beaucoup aidé dans la vie pratique de l'UMT. Merci à toute l'équipe IFV !

Je voudrais remercier tous mes co-auteurs pour leur contribution, leurs conseils scientifiques et amicaux, leurs corrections, leur regard bienveillant et leurs encouragements.
Anne-Françoise Adam-Blondon, Roberto Bacilieri, Isabelle Beccavin, Gilles Berger, Yves Bertrand, Jean-Michel Boursiquot, Philippe Chatelet, Marie Denis, Agnès Doligez, Marc Farnos, Alexandre Fournier-Level, Félix Abdel Aziz Homa, Thierry Lacombe, Valérie Laucou, Amandine Launay, Marie-Christine Le Paslier, Samuel Neuenschwander, Jean-Pierre Péros, Charles Romieu, Maryline Roques, Vincent Segura, j'ai beaucoup apprécié nos interactions et j' ai énormément appris à votre contact; cela m'a fait plaisir d'avoir travaillé avec vous !

Merci à Yves Gibon, qui m'a accueillie à la plateforme Métabolome à l'INRA de Bordeaux pendant 4 semaines pour réaliser des milliers d'analyse phénotypiques !

Un grand merci à Martine Barraud et à Joëlle Lopez, pour leur patience et disponibilité concernant la gestion des taches administratives de l'école doctorale.

Merci à tous ceux qui ont participé à mes trois campagnes de récolte sur le terrain ou au labo. Sans vous cette thèse n'aurait pas été possible. Un merci exceptionnel à Yves et Gilles, à Pierre et Thierry, à Jean-Pierre et à Patrick pour votre bonne humeur, pour les cafés et les discussions qui ont transformé le travail lourd et fatiguant du terrain en de bons moments !

Un grand merci aux trois générations de « Bocal » que j'ai eu la chance de connaitre.

Merci pour vos encouragements Greg et Maud, je suis heureuse d'avoir assisté au démarrage de votre histoire ! Merci à Yung-Fen pour les ondes positives que tu nous as transmises, ton calme, ton écoute, et de t'être libérée pour faire des pauses café avec Pilar et moi. Merci Pilar pour le lait dans le café, les petits gâteaux, mon mug (mes mugs) et pour ta grande compréhension !

L'année des stagiaires avec Alice, Claire et Ana ! Merci pour votre compagnie, les barbeuq, notre WE inoubliable en Espagne et le kiwi-vous !

Et cette dernière année, le Bocal est à rebord ! Avec une très joyeuse compagnie, qui m'ont quand même toujours laissé travailler et qui m'ont encouragé, quand j'en avais marre. Jaques, Iris, Yan, Cédric, Julia, Hajar, merci à vous !

Merci beaucoup aux autres thésards du deuxième étage : Ingrid, Constance qui nous appelle tous les jours pour aller manger, ainsi que Markus, Ratthaphon et Frederico avec qui on a également partagé de bons moments (et de bons vins !).

Merci à tous les membres de l'équipe DAVEM, que j'ai connus pendant mon séjour ! Il existe une ancienne salutation hongroise, que je vous souhaite en partant : Erőt, egészséget, bort, búzát, békességet ! (« De la force, de la bonne santé, du vin, du blé et de la paix pour vous tous !»)

Merci beaucoup à Nathalie Camus, mon mentor à la formation NCT, qui m'a aidé à développer un nouveau regard sur ma thèse et sur moi-même en m'aidant à découvrir les compétences que j'ai pu acquérir pendant cette expérience. Merci pour votre regard positif.

Mercie à Isabel Martin Grande et à Dominique This pour leur soutien et leur bienveillance.

Merci à mes amis hongrois, français et internationaux, des anciens et des nouveaux qui m'ont soutenue, détendue, qui ont cru en moi. Un grand merci à mes professeurs et l'équipe de dance Rock'n'Style; cette activité est petit à petit devenue ma « drogue », qui pouvait me détendre après une dure journée de travail. Merci à Aude et Julie qui se sont bien occupées de moi ces 6 derniers mois, qui m'ont fait découvrir les paillottes ! Merci Cecilio de m'avoir consolée, quand mon LEPSE-ien préféré est parti !

Merci beaucoup à ma famille, qui n'ont jamais cessé de croire en moi et de m'encourager à distance tout au long de ces 5 années passées en France. Cela fait 5 ans, que je suis arrivée à Montpellier, initialement pour 5 petits mois d'Erasmus. Et puis j'ai fait des rencontres, qui m'ont marquées et qui m'ont convaincues de rester et continuer ma carrière, ma vie ici en France et en français. Merci à José et Anaïs, grâce à vous je garde des souvenirs idyllique de mon séjour à Bordeaux !

Et pour finir, merci à Grégoire, sur qui j'ai pu compter dans le meilleur et le pire depuis 5 ans ! Merci pour ta compréhension, tes attentions, tes encouragements, ta tendresse. J'ai hâte de te retrouver et continuer nos aventures ensemble, cette fois dans mon pays !

SOMMAIRE

Remerciements 3

Sommaire 6

Chapitre 1 : Introduction bibliographique 11

1. La situation de la viticulture 12

1.1. Dans le monde et en France (l'importance de l'espèce, production) 12

1.2. Nouveaux enjeux socioéconomiques – l'intérêt de la création variétale chez la vigne 12

2. La création variétale chez la vigne – les atouts et les difficultés 13

2.1. Création des cépages de demain avec les outils d'aujourd'hui 13

Créer les cépages de demain avec les outils d'aujourd'hui 15

1-Introduction 15

2-La création variétale chez la vigne, les méthodes restent mais les outils changent. 15

3- Outils actuels pour créer de nouvelles variétés de vigne 19

4- Conclusion et perspectives 20

Références : 21

2.2. Les ressources génétiques disponibles chez la vigne 22

2.2.1. La composition du genre Vitis 22

2.2.2. Pools génétiques disponible dans l'espèce Vitis vinifera L. 23

2.3. Connaissances et outils moléculaires disponibles 24

2.3.1. Le génome de la vigne 24

2.3.2. Les marqueurs 24

2.4. Les programmes de création variétale chez la vigne 25

2.4.1. Avant les marqueurs moléculaires 25

2.4.2. SAM chez la vigne 27

3. Les dernières avancées de la sélection assistée par marqueurs 29

3.1. Identification des marqueurs pour la SAM via GWAS 29

3.1.1. Méthodologie de la génétique d'association 29

3.2. La Sélection Génomique (GS) 30

3.2.1. La méthodologie de la prédiction génomique 30

3.2.2. L'intérêt de la GS 34

4. Objectif de la thèse et démarche expérimentale 36

4.1. Objectifs de l'étude 36

4.2. Démarche expérimentale 36

Chapitre 2 : Evaluation de l'interêt de la sélection génomique par simulations 38

1. Introduction 39

2. Evaluation de l'intérêt de la sélection génomique par simulations 41

Genome-wide prediction methods in highly diverse and heterozygous species: proof-of-concept through simulation in grapevine. 41

Abstract 43

Introduction 44

Materials and methods 46

Simulation: 46

Linkage disequilibrium: 49

Genome-wide association: 50

Genomic prediction: 50

Test validation on standard data: 51

Results 52

Simulation: 52

Descendent populations: 53

Genome-wide association study (GWAS): 54

Prediction of phenotypes from genotypes: 54

Pine data: 56

Discussion 56

Simulated data: 56

Feasibility of GWAS in grape: 57

Prediction of phenotypes from genotypes by GEBV: 58

Prediction methods: 59

Combination of training and candidate sets: 60

Influence of trait structure: 61

Acknowledgements 61

Literature cited 62

3. Conclusion 73

Chapitre 3 : Prédiction des phénotypes à partir d'un large échantillon de diversité 74

1. Introduction 75

2. Prédiction des phénotypes à partir d'un large échantillon de diveristé 77

Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) in grape for phenotype prediction using a large diversity panel 77

Abstract 78

Introduction 79

Materials and methods 81

Plant material: 81

Field experiment: 82

Phenotyping of the studied traits: 82

Statistical tests and heritability: 82

Genotyping: 83

Genotypic data encoding: 84

Population structure and relatedness: 84

Genome-wide association: 85

Prediction methods: 85

Results 87

Genotypic data: 87

Relationship between individuals: 87

Phenotypic data: 87
GWAS: 88
Prediction of phenotypes using genotypes: 89
Discussion 90
GWAS: 91
Genomic selection: 92
References 96
Chapitre 4 : prédiction des génotypes sur une population biparentale 108
1. Introduction 109
2. Matériels et méthodes 109
3. Résultats et discussion 111
3.1. Génétique d'association 111
3.2. Prédiction des phénotypes 112
Chapitre 5 : Discussion et perspectives 114
1. Discussion 115
1.1. Contexte de l'étude 115
1.2. Simulation 117
1.3. Données réelles 119
1.4. Différence entre simulation et données réelles 122
1.5. Utilisation du Genotype « Dwarf » 122
2. Perspectives 123
2.1. Le développement d'une population d'entrainement universelle 124
2.2. La prédiction en ségrégation 124
2.3. Set de géniteurs élites 124
3. Conclusion 126
Références bibliographiques 127
Annexes 1
1. Annexe I 2

2. Annexe II 15

3. Annexe III 33

CHAPITRE 1 : INTRODUCTION BIBLIOGRAPHIQUE

1. La situation de la viticulture

1.1. Dans le monde et en France (l'importance de l'espèce, production)

La vigne (*Vitis vinifera* L.) est une des plus importantes espèces fruitières cultivées dans le monde. Selon les différentes sources officielles, en 2011 la viticulture représente la troisième production fruitière derrière les agrumes et les bananes, avec une moyenne de 69,6 millions de tonnes de raisins frais cultivé sur 7,1 million d'hectares sur les 5 continents (http://faostat3.fao.org et http://www.oiv.int/oiv/info, 2011). La majorité du raisin produit (71%) est transformé en vin, ce produit à haute valeur ajoutée possède une symbolique gastronomique, religieuse et sociale importante. Selon les années, la France et l'Italie se disputent la première place de la production viticole mondiale. En France, la filière viticole est le premier secteur contributeur à la valeur agricole du pays avec 15 % de la valeur de la production agricole totale.

Les caractéristiques de cette filière sont résumés en quelques chiffres (FranceAgriMer, 2011, http://www.franceagrimer.fr):

- ✓ 774 000 ha de vignobles (3 % des terres arables) ;
- ✓ 50 millions d'hl de vins produits ;
- ✓ 7,17 milliards d'euros d'exportations de vins (1er secteur exportateur agroalimentaire et 3ème secteur économique exportateur derrière l'aéronautique et la parfumerie) ;
- ✓ 250 000 emplois directs,
- ✓ ainsi que d'importantes retombées touristiques..

1.2. Nouveaux enjeux socioéconomiques – l'intérêt de la création variétale chez la vigne

Malgré ce cadre enviable, cette filière est en pleine mutation. En effet la viticulture française doit faire face à 3 grands défis:

- ✓ **Le respect de l'environnement.** Afin de garantir un développement durable de l'agriculture en respectant la biodiversité et l'environnement, le Grenelle de l'Environnement (rencontres politiques organisées en France en 2007) a mis en place le plan Ecophyto 2018. Ce plan vise à réduire la dépendance des exploitations agricoles aux produits phytosanitaires, tout en maintenant un niveau élevé de production agricole, en quantité et en qualité. Il avait comme ambitions la réduction de moitié de la fréquence de traitement des pesticides dans l'agriculture française en 10 ans et le retrait du marché des substances les plus préoccupantes (ces objectifs ayant été revus dernièrement). La vigne est aujourd'hui une des espèces les plus fortes

utilisatrices de produits phytosanitaires (par unité de surface) en France et en Europe, et les restrictions concernent plusieurs produits essentiels à la viticulture.

- ✓ **Les changements climatiques.** Les conditions climatiques propres à chaque région viticole ont un rôle essentiel dans la production de vins de qualité. Les changements des températures et des précipitations vont modifier le terroir, ce qui affectera la production du raisin et les caractéristiques du vin. Les différents scénarios sur l'évolution du climat suggèrent un réchauffement progressif et l'augmentation du déficit hydrique pour Europe (TURNER *et al.* 2010). Les conditions climatiques actuelles d'une région se décaleront vers le nord, nord-ouest ou vers une altitude plus élevé, avec un écart plus important dans la région méditerranéenne (MORIONDO *et al.* 2013; HANNAH *et al.* 2013). Les changements climatiques, pourraient aussi modifier la zone de répartition de certaines maladies. Les cépages tolérants ou résistants à ces stress biotiques et abiotiques (sécheresse ou température par exemple) vont devenir de plus en plus recherchés.
- ✓ **L'évolution des marchés.** La viticulture Français va devoir s'adapter à la demande des marchés internationaux. Elle doit faire face à une compétition de plus en plus soutenue de la part des autres pays notamment ceux du nouveau monde (Chili, Argentine, Etats Unis, Australie, Afrique du Sud…). Le développement de la filière vitivinicole française à l'export se fera donc en particulier avec des vins dans les segments premium et ultra premium. Cela implique d'abord d'améliorer encore ou de conserver la qualité œnologique des raisins et des vins français, et notamment la qualité des nouvelles variétés résistantes.

Même si l'organisation actuelle de la viticulture ne laisse que peu de place à la création variétale – en effet les AOCs imposent un unique lieu de production mais aussi une liste de cépages généralement réduite pour produire un vin d'appellation – et que le consommateur averti est sensibilisé aux cépages dans les vins, la création variétale doit être envisagée comme une des solutions pour répondre à ces défis. La profession est par ailleurs en attente de ces nouvelles variétés résistantes.

2. La création variétale chez la vigne – les atouts et les difficultés

2.1. Création des cépages de demain avec les outils d'aujourd'hui

Au cours de cette thèse, un article a été publié dans un guide technique «Les cépages résistants aux maladies cryptogamiques Panorama européen ; ouvrage réalisé sous la direction du groupe ICV » (GROUPE ICV 2013).

Ce guide comprend deux parties,

- ✓ une première partie présentant sous forme d'articles scientifiques vulgarisés en français les avancées de la recherche sur la création variétale chez la vigne et les dernières connaissances sur les agents pathogènes que sont le mildiou et l'oïdium.
- ✓ une deuxième partie présentant une liste non exhaustive de plusieurs cépages européens ainsi que leurs caractéristiques agronomiques, œnologiques et leur niveau de tolérance aux deux principales maladies cryptogamiques que sont l'oïdium et le mildiou.

Dans cet ouvrage, nous avions proposé un article (Le Cunff, L., Lacombe, T., Fodor, A., Farnos, M., Audeguin, L., This, P. Boursiquot J.M. 2013. Créer les cépages de demain avec les outils d'aujourd'hui. *In:* Rousseau, J., Chanfreau, S. (eds). *Les cépages résistants aux maladies: Panorama européen*. Lattes, France: Groupe ICV, p. 34-40.), relatant l'histoire de l'amélioration de la vigne jusqu'à nos jours, en présentant les étapes de l'amélioration variétale chez la vigne en relation avec les connaissances scientifiques de chaque période. Cet article est présenté ci-après.

Créer les cépages de demain avec les outils d'aujourd'hui

Le Cunff Loïc[1,2], Lacombe Thierry[2,3], Fodor Agota[1,2], Farnos Marc[2,3], Audeguin Laurent[1,2], This Patrice[2,3], Boursiquot Jean-Michel[1,2,3]

Institut Français de la Vigne et du vin, Pôle matériel végétal, Domaine de l'Espiguette, 30240 Le Grau-du-Roi.

Unité Mixte Technologique Géno-Vigne® (IFV, INRA, Montpellier SupAgro), 2 place Viala, 34060 Montpellier Cedex.

INRA – Montpellier SupAgro, UMR AGAP (Amélioration Génétique et Adaptation des Plantes) Equipe « Diversité et Adaptation de la Vigne et des Espèces Méditerranéennes » 2, Place Viala, 34060 Montpellier Cedex.

1-Introduction

La viticulture française doit faire face aujourd'hui à plusieurs défis. Elle est en effet une filière consommatrice de produits phytosanitaires et comme les autres cultures, elle est déjà confrontée aux évolutions du climat qui pourraient engendrer de profondes modifications, notamment en zone méditerranéenne. De plus, notre viticulture doit faire face à une compétition accrue de la part d'autres pays (comme par exemple : Chili, Argentine, Etats-Unis, Australie, Afrique du Sud, etc.). Afin de répondre à ces défis, nous pouvons nous appuyer sur des outils déjà disponibles comme la création de génotypes présentant des résistances aux principaux bio-agresseurs, mais également sur les avancées scientifiques permettant d'identifier les déterminismes génétiques des caractères d'intérêts. Ces connaissances constituent un socle solide pour de nouveaux programmes d'amélioration variétale chez la vigne et représentent certainement l'une des ressources pour s'adapter aux enjeux actuels.

2-La création variétale chez la vigne, les méthodes restent mais les outils changent.

La création variétale chez la vigne n'est pas une voie d'amélioration nouvelle. Elle fut utilisée consciemment ou non pour améliorer les cépages depuis la domestication de la vigne. Dans cet article, nous avons choisi de présenter la création variétale chez la vigne au travers de quatre périodes historiques majeures. Ces étapes ne se définissent pas par de nouvelles méthodes mais par une meilleure compréhension des facteurs et des acteurs de la création de nouvelles variétés. De tout temps, l'amélioration de la vigne se résume à deux méthodes encore utilisées actuellement ; seule leur mise en œuvre et les outils ont changé au cours du temps. Ces deux méthodes sont les suivantes :

La première est basée sur l'apparition naturelle de mutations dans les populations de vigne. L'impact de ces mutations est repéré morphologiquement et quand ces mutations sont intéressantes, elles sont conservées/fixées et multipliées par simple bouturage ou greffage : c'est la sélection clonale. Dans ce cas, on a une conservation de l'identité variétale de la souche initiale. L'opération peut être répétée à l'infini et par cette technique, certains cépages ont pu traverser les âges, quasiment inchangés, pour parvenir jusqu'à nous.

La seconde méthode utilise la reproduction sexuée. On sème un pépin qui est le résultat d'un croisement, naturel ou volontaire, entre deux géniteurs via le pistil de l'individu maternel et le pollen de l'individu paternel. On obtient ainsi une nouvelle plante, originale, distincte des deux parents et qui combine au hasard certains caractères parentaux.

Actuellement, la méthodologie de transfert de gènes (OGM) constitue une troisième voie potentielle d'amélioration, mais elle n'a pas encore été utilisée pour la création variétale chez la vigne.

Ces technologies sont seulement utilisées en laboratoire pour répondre à des questions d'ordre scientifique sans but commercial.

2-1 La création variétale inconsciente (première période)

La phase de domestication intervient au Néolithique, l'objectif pour les hommes de cette époque étant de disposer à proximité de leurs habitats, de vignes plus productives (grosses grappes, grosses baies), plus régulières et plus sucrées (aptitudes accrues à la fermentation). Pour réaliser cette longue opération de domestication, les premiers viticulteurs ont utilisés les deux méthodes mentionnées précédemment : sélection de vignes, propagées par multiplication végétative et création de nouveaux individus par reproduction sexuée. Ces premiers viticulteurs ont donc soit

Figure 1.1. Processus de création d'une variété. (Photos de J.P. Bruno INRA Domaine de Vassal (étapes 1,2,3), et de la Groupe ICV)

bouturé des vignes sauvages puis sélectionné des clones, soit planté des pépins et sélectionné des cépages adaptés à leurs besoins.

Un autre exemple de l'utilisation inconsciente de l'hybridation ou de la reproduction sexuée suivie d'une étape de sélection de cépages en vue d'amélioration variétale est l'expansion de la viticulture vers le nord de la France. Nous savons qu'en Gaule narbonnaise, un important vignoble a été créé par des colons romains ayant vraisemblablement importé des variétés d'Italie, au climat similaire. Mais ces cépages méridionaux se sont avérés peu adaptés aux zones extra-méditerranéennes, la viticulture s'est donc difficilement étendue au-delà de Gaillac à l'Ouest et de Valence au Nord. Dans ce contexte, l'hybridation des cépages importés avec des lambrusques locales naturellement adaptées, a sans doute permis la création rapide de variétés « gauloises » ou « gallo-romaines » et permis l'extension des vignobles en remontant les voies fluviales vers l'Ouest, le Nord et l'Est (Garonne, Loire, Rhône, Rhin).

Par la suite et durant tout le Moyen Age jusqu'au XVIIIe siècle, d'autres cépages ont été obtenus par hybridations sans doute spontanées entre cépages comme en témoignent les nombreux descendants (Chardonnay, Gamay, Melon, etc) issus de croisements entre le Gouais et le Pinot.

2-2 Premiers croisements contrôlés (deuxième période)

Le XIXe siècle voit apparaître les premiers travaux sur des croisements volontaires et contrôlés entre deux cépages dans le but de créer de nouvelles variétés mieux adaptées aux objectifs de production. Cette technique consiste tout d'abord à castrer une grappe pour éviter des autofécondations (élimination des étamines avant leur maturité de façon à éviter la rupture des sacs polliniques et donc une possible fécondation). Une fois cette étape réalisée, il s'agit, à l'aide d'un pinceau, d'appliquer le pollen de l'autre parent choisi par l'hybrideur pour réaliser la fécondation (cf. Figure 1.1). Les fleurs fécondées sont alors « ensachées » pour éviter des contaminations avec une autre source de pollen.

Près de Montpellier, en 1828, c'est-à-dire avant la connaissance des lois de Mendel sur l'hérédité (1865), Louis Bouschet de Bernard va hybrider des cépages méridionaux déficients en couleur (Grenache, Carignan, Cinsaut) avec un cépage teinturier (pulpe colorée) non adapté au Midi, pour obtenir des cépages très colorés convenant à la région. La variété la plus célèbre obtenue à cette période est l'Alicante Henri Bouschet ; elle est encore plantée sur plusieurs milliers d'hectares en

France et dans le monde. Dans ce cas, le choix des géniteurs est important et les caractères ciblés sont déterminés par le sélectionneur.

Le XIXe siècle est celui du Phylloxéra. Parallèlement à la création de porte-greffes résistants à partir d'espèces américaines, une autre voie d'investigation a cherché la solution dans des croisements complexes entre ces mêmes vignes américaines et les cépages européens, dans l'objectif d'obtenir des nouvelles variétés non greffées, résistantes aux maladies (oïdium, mildiou, phylloxéra) et capable de produire des raisins utilisables. Ces variétés étaient nommées « Hybrides Producteurs Directs ». Plusieurs centaines ont été créées. Elles ont constitué jusqu'à 30% du vignoble français dans les années 1960.

Cependant, les caractères ciblés étaient essentiellement des caractères à déterminisme génétique dit simple, définition qui correspond à des caractères contrôlés en général par une unique mutation, et facilement identifiable morphologiquement.

2-3 Connaissances des lois de l'hérédité, vers une optimisation des croisements (troisième période)

Au XXe siècle, les lois de l'hérédité sont connues et diffusées ; des modèles statistiques sont développés pour estimer l'attendu dans un croisement contrôlé. Les caractères étudiés ne sont plus à déterminisme génétique simple, mais complexe. Cette révolution permet de mener des programmes scientifiques de création de nouveaux cépages basés sur des analyses génétiques rationnelles. Ces innovations sont faites surtout par des instituts de recherche et non plus par des viticulteurs ou des amateurs éclairés. Ce sont surtout des raisins de table qui sont créés, car l'innovation dans ce secteur est plus facilement acceptée. La réussite de cette période d'hybridation est quantifiable au travers du nombre de cépages obtenus, avec des variétés comme le Cardinal, l'Italia, le Red Globe, le Centennial seedless, le Danlas ou le Prima que l'on trouve actuellement dans le commerce.

2-4 Connaissance de l'ADN, des polymorphismes moléculaires et de leurs impacts sur la variabilité des cépages (quatrième période)

Dans la période suivante, l'ADN est découvert ainsi que son rôle dans la déterminisme des caractéristiques d'une plante. Les différences observées (ou polymorphismes) entre deux individus au niveau de leurs séquences d'ADN sont directement corrélées avec la morphologie, la phénologie

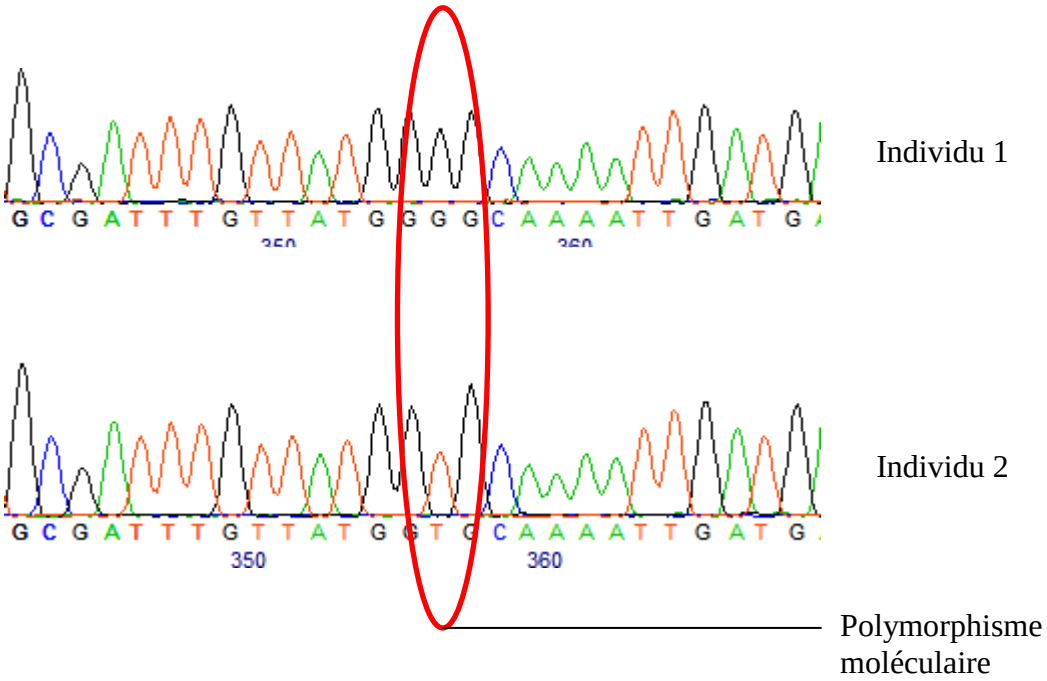


Figure 1.2. Mise en évidence du polymorphisme moléculaire existant entre deux individus.

et la composition organique de la plante. Ces polymorphismes sont visualisables grâce au développement d'une nouvelle discipline scientifique, la Biologie Moléculaire, qui permet de multiplier in vitro l'ADN et de comparer des régions du génome de plusieurs individus (cf. Figure 1.2).

Avec ces nouveaux outils moléculaires, le choix des parents et la sélection des descendants se fait sur la base de la présence ou de l'absence de certains polymorphismes, même sans avoir observé les caractères ciblés chez ces individus, mais en se référant seulement à des études menées préalablement aux croisements. Une limite de cette approche est que seuls les caractères pour lesquels des polymorphismes « fonctionnels » ont été identifiés sont sélectionnables.

3- Outils actuels pour créer de nouvelles variétés de vigne

3-1 Le décryptage complet du génome de la vigne

La morphologie d'une variété, la composition de sa baie, son niveau de résistance aux bio-agresseurs ou encore sa capacité d'adaptation aux différents environnements sont préprogrammés par la séquence « personnelle » de son ADN. Chaque variété, comme chaque organisme vivant possède en effet une séquence qui lui est propre. Cependant l'organisation globale de cette séquence est spécifique d'une espèce (« le génome »). Pour la vigne cultivée (*Vitis vinifera* L.), le décryptage du génome a été publié à deux reprises en 2007. Même si cet effort représente une avancée majeure pour la recherche et donc à plus long terme pour la création variétale, ce décryptage ne permet pas directement d'avoir accès aux mutations fonctionnelles utilisables en sélection. Il est nécessaire pour cela de réaliser des études complémentaires en vue d'obtenir une connaissance étendue de la diversité moléculaire présente chez un nombre important de cépages pour l'utiliser ultérieurement.

3-2 La diversité disponible, un réservoir d'innovation

Le décryptage du génome de la vigne permet d'accélérer l'identification de polymorphismes localisés le long du génome et de chercher des corrélations avec des caractères intéressants. Cependant, pour avoir accès à un grand nombre de mutations on doit disposer d'un grand nombre de cépages avec des origines les plus diverses possibles. Les conservatoires de vigne, comme celui du Domaine INRA de Vassal (Hérault), constituent donc naturellement des ressources indispensables aux généticiens, puisqu'ils représentent autant de réservoirs de diversité génétique.

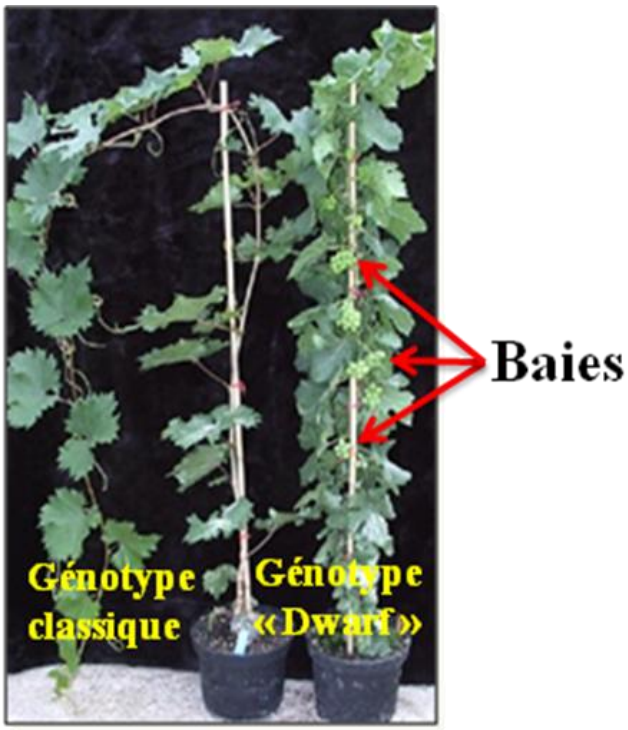


Figure 1.3. Photographie comparative de temps de mise à fruit entre un génotype classique et un génotype « Dwarf ». (source : Laurent Torrigosa)

3-3 Le génotype « Dwarf »

Afin d'accélérer les cycles de reproduction chez la vigne et de raccourcir le temps d'un programme de sélection, il est désormais possible d'utiliser un génotype particulier appelé « Dwarf » (nain). Ce génotype est le résultat d'une mutation spontanée (naturelle) apparue dans une des couches cellulaires du cépage Meunier le rendant insensible aux gibbérellines. Cette insensibilité a pour conséquence de réduire le temps de mise à fruit des génotypes porteurs de cette mutation (cf. Figure 1.3.). Lorsqu'un programme d'amélioration variétale nécessite plusieurs cycles de croisement, l'attente de mise à fruit des génotypes intermédiaires est longue chez la vigne classique (entre 2 à 3 ans minimum) alors que chez les génotypes « Dwarfs » le passage du « pépin au pépin » est de 9 à 10 mois. Ces génotypes sont donc de parfaits outils intermédiaires pour accélérer les programmes de sélection complexe chez la vigne. De plus, il est possible à chaque génération de revenir à un génotype « classique » puisque 50% des descendants d'un croisement « vigne Dwarf x vigne classique », sont morphologiquement normaux.

3-4 Vers une cinquième période, la sélection génomique

L'une des limites de la sélection utilisant l'information génétique est le nombre de polymorphismes fonctionnels connus. Pour palier ce déficit de connaissances, un nouvel outil, utilisé avec succès dans la sélection animale, est en cours d'évaluation chez la vigne : la sélection « génomique ». Cette puissante approche bio-statistique apparaît comme très novatrice dans le monde de la création variétale puisqu'elle ne nécessite pas l'identification de polymorphismes fonctionnels pour améliorer un caractère ciblé. Elle repose uniquement sur la connaissance de la présence ou de l'absence d'un grand nombre de mutations (entre 50 000 à 100 000) sans information sur leurs corrélations avec des caractères d'intérêt. Si cette approche prouve son efficacité chez la vigne, le nombre de caractères améliorables sera significativement augmenté.

4- Conclusion et perspectives

L'innovation en viticulture est depuis toujours associée à la création variétale. Les possibilités d'obtenir des variétés très proches de celles voulues ont évolué avec les avancées de la recherche scientifique, sans perdre de vue que le matériel végétal créé doit correspondre à la demande des viticulteurs, répondre à leurs préoccupations et être en adéquation avec les attentes des

consommateurs. Cette dernière motivation a d'ailleurs été le moteur de la création au cours du temps, que ce soit pour domestiquer la vigne, adapter sa culture à d'autres environnements ou pour résister à des bio-agresseurs en l'absence de produits phytosanitaires. La création variétale reste une source importante d'innovation pour répondre aux besoins de la viticulture. Elle dispose aujourd'hui de nouveaux outils qui lui permettront de créer des cépages plus rapidement et de mieux cibler les individus sélectionnés. Le dialogue entre tous les acteurs de la viticulture doit donc maintenant être approfondi afin de définir ensemble les caractéristiques des cépages de demain.

Références :

Boss P., M. Thomas. (2002) Association of dwarfism and floral induction with a grape 'green revolution' mutation. *Nature,* 416: 847-850.

Bowers J, Boursiquot JM, This P, Chu K, Johansson H, Meredith C (1999) Historical genetics: The parentage of chardonnay, gamay, and other wine grapes of northeastern France. Science 285 (5433):1562-1565.

Jaillon A, consortium FIp (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 443-467.

INRA Domaine de Vassal, Centre de Ressources Génétique de la Vigne, http://www1.montpellier.inra.fr/vassal/

Lacombe T (2009) La longue histoire des cépages, des origines à nos jours. Présenté à: Musée gallo-romain de Saint-Romain-en-Gal, Vienne, 4 juin 2009

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, et al. (2007) A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLoS ONE 2: e1326.

J. Chaïb, L. Torregrosa, D. Mackenzie, P. Corena, A. Bouquet, and M. Thomas, (2010) The grape microvine – a model system for rapid forward and reverse genetics of grapevines. Plant Journal 62(6):1083-92.

2.2. Les ressources génétiques disponibles chez la vigne

La vigne (*Vitis vinifera* L.) est une plante pérenne qui se présente sous la forme d'une liane. Selon la nouvelle classification cladistiques des plantes à fleurs proposée par les botanistes du collectif *Angiosperm Phylogeny Group* (2009) basée sur l'analyse de séquences d'ADN, elle appartient à la famille des Vitaceae qui se positionne dans la division des *Magnoliophyta* (ou Angiospermes), clade des Eudicotylédones supérieures (anglais *Core eudicots*), clade des Rosidées (anglais *Rosids*), ordre des Vitales. Parmi les espèces de la famille des Vitaceae, seules les espcèes du genre *Vitis* ont un intérêt en tant que ressources génétiques pour la création variétale.

2.2.1. La composition du genre Vitis

D'après (PEROS *et al.* 2011), l'origine du genre *Vitis* est située en Eurasie et il s'est ensuite étendu vers l'Ouest sur le continent américain. La séparation des continents et les périodes de glaciations successives du pléistocène, ont alors provoqué l'isolement de populations qui a conduit à des événements de spéciations. En effet, le genre *Vitis* se divise en deux sous-genres : *Muscadinia* et *Vitis* (ou *Euvitis*).

- ✓ Le sous-genre *Muscadinia* possède 2n=40 chromosomes et n'est représenté que par 2 à 3 espèces localisées du Sud-est des Etats-Unis jusqu'en Amérique Centrale. La seule espèce cultivée est la *Muscadinia rotundifolia*. Elle est intéressante pour l'amélioration grâce à son niveau de résistance élevé à plusieurs maladies, y compris l'oïdium et le mildiou, mais elle n'a pas une bonne qualité gustative (BOUQUET 1982).
- ✓ Le sous-genre *Vitis* possède 2n=38 chromosomes et comprend une soixantaine d'espèces. Elles sont classées selon leurs origines géographiques (GALET 1988) :
    1) Les vignes américaines : *V. aestivalis*, *V. berlandieri*, *V. cinerea*, *V. labrusca*, *V. lincecumii*, *V. riparia*, *V. rupestris,* utilisées en particulier depuis la crise phylloxérique comme porte-greffes ou en croisement avec *Vitis vinifera* L. (LEVADOUX 1956);

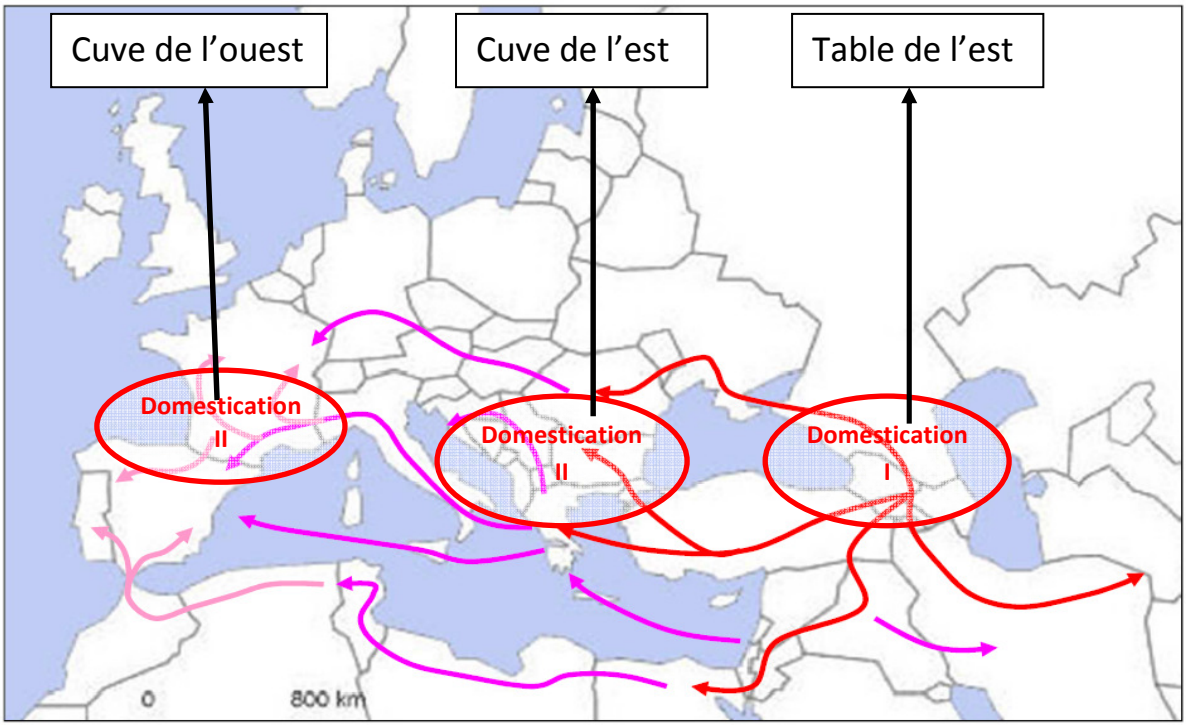


Figure 1.4. : Diffusion de la viticulture et de la vigne cultivée. Après LABRA *et al.* (2002)**.**

2) les vignes asiatiques notamment l'espèce *Vitis amurensis* qui est utilisée dans les programmes d'amélioration pour sa tolérance au froid (KOLEDA 1975) ;
3) la vigne euro-asiatique, *Vitis vinifera* L. qui regroupe l'ensemble des cépages cultivés, de cuve et de table appartenant à la sous espèce *Vitis vinifera* subsp. *vinifera,* ainsi que les vignes sauvages de la sous espèce: *Vitis vinifera* subsp. *sylvestris*.

2.2.2. Pools génétiques disponible dans l'espèce Vitis vinifera L.

Au sein de l'espèce *Vitis vinifera* subsp.*vinifera* qui correspond au compartiment cultivé de la vigne, (BACILIERI *et al.* 2013) ont décrit trois groupes de diversité selon l'origine géographique et l'utilisation des cépages. Cette étude a été menée sur 2096 individus de la collection de Vassal (la plus grande collection mondiale de l'espèce *Vitis vinifera*) en utilisant 20 marqueurs microsatellites (SSR).

Le premier groupe différencié, appelé « table de l'est – TE », correspond aux raisins de table d'aujourd'hui qui aurait pour origine le centre de domestication primaire, localisé dans le Caucase (LEVADOUX 1956; ARADHYA *et al.* 2003; GRASSI *et al.* 2003; ARROYO-GARCIA *et al.* 2006). Les deux autres groupes regroupent les variétés de cuve, qui ont été sélectionnées pour donner du vin. Concernant le groupe appelé « cuve de l'est – CE », il est possible qu'une partie des cépages qui le composent soit le résultat d'un événement de domestication secondaire dans la région de la Péninsule balkanique (Lacombe com. pers.). En effet, la culture de la vigne et du vin suivait les migrations humaines. Les variétés du centre de domestication primaire ont été croisées avec des variétés locales ou des individus sauvages locaux formant un nouveau groupe de diversité, un centre domestication secondaire. Le troisième groupe, « cuve de l'ouest – CO » formé en Europe de l'ouest serait le résultat d'une domestication secondaire par croisements entre les cépages des balkans et des individus sauvages de ces zones géographiques (Nord de l'Espagne, Sud de la France; ARROYO-GARCIA *et al.* 2006). Ces trois groupes sont très proches de ceux caractérisés par (NEGRUL 1938) sur des observations phénotypiques: Orientalis pour TE, Pontica pour CE et Occidentalis pour CO. Ces trois groupes peuvent être distingués par les marqueurs moléculaires (Figure 1.4.).

La vigne cultivée compte actuellement prés de 6 000 cépages à travers le monde (GALET 2000). La diversité génétique présente au sein du compartiment cultivé chez la vigne est aussi très importante avec près de 17 allèles par loci SSR (LAUCOU *et al.* 2011) et un SNP tous les 49 nucléotides (LE CUNFF *et al.* 2008). De plus des études récentes montrent que peu de généalogies complexes ont

pu être identifiés (faible nombre de générations et un grand nombre de cépages sans apparentement) (MYLES *et al.* 2011; LACOMBE *et al.* 2012). Cependant malgré cette diversité disponible, 20 cépages représentent à eux seuls plus du tiers (37%) des vignobles mondiaux (Boursiquot J.M., pers. com.). Cet encépagement très peu diversifié rend la viticulture actuelle particulièrement vulnérable face aux maladies et aux changements climatiques. De plus parmi les cépages le plus cultivés, beaucoup sont apparentés, souvent avec une seule génération d'écart ; par exemple Pinot noir – Chardonnay – Gouais blanc, Cabernet franc – Merlot ; ((BOWERS *et al.* 1999; BOURSIQUOT *et al.* 2009).

2.3. Connaissances et outils moléculaires disponibles

2.3.1. Le génome de la vigne

La vigne est une espèce diploïde, sa composition génomique est de 2n = 38 chromosomes pour une taille de 470 Mb (JAILLON *et al.* 2007). Cette espèce est très hétérozygote (He=0,74), caractérisée par une diversité génétique importante (LIJAVETZKY *et al.* 2007; MYLES *et al.* 2010; LAUCOU *et al.* 2011; CARRIER *et al.* 2012). Plusieurs études ont été menées sur le déséquilibre de liaisons (DL). La première a été réalisée avec 38 marqueurs SSRs, dans une core-collection de 141 cultivars de *V. vinifera* dans cinq régions du génome correspondant toutes à des chromosomes différents et dans lesquelles avaient été détectés des QTL (BARNAUD *et al.* 2006). Le DL était significatif jusqu'à 16,8 cM, mais le r^2 a chuté à 0.1 sur 5 cM qui correspond en moyenne sur le génome à 650-1080 kb. Des études menées sur des échantillons différents avec des SNPs en région intra-génique ont montrées une diminution du DL nettement plus rapide, avec un r^2=0,2 pour une distance allant de 200 à 700 nucléotides (THIS *et al.* 2007; LIJAVETZKY *et al.* 2007). Une étude utilisant 3349 SNPs répartis au long du génome (MYLES *et al.* 2010) estiment une taille de DL sur environ 10Kb, pour un r^2 égal à 0,2. En se basant sur cette dernière étude, le nombre de marqueurs nécessaires pour marquer l'ensemble du génome de la vigne (470 MB) est d'au moins 47 000 SNPs.

2.3.2. Les marqueurs

La séquence complète du génome (JAILLON *et al.* 2007 ; VELASCO *et al.* 2007) facilite grandement la mise en place de méthodologies haut débit d'identification de marqueurs. Notamment grâce à l'utilisation des nouvelles technologies de séquençage qui permettent de découvrir de nombreux polymorphismes et de diminuer le prix du génotypage. MYLES *et al.* (2010) ont utilisé une méthode rapide et relativement simple, pour détecter 469 470 SNP sur le génome entier. Ces données ont servie de base au développement d'une puce de génotypage de 9K SNP. Depuis, une nouvelle puce 18K a été réalisée par LE PASLIER *et al.* (2013) en analysant 43 génotypes *Vitis vinifera ssp. vinifera* et des espèces apparentées (*Vitis vinifera ssp vinifera,* 4 *V. vinifera ssp sylvestris,* 3 *V. cinerea*, 3 *V.*

berlandieri, 3 *V. aestivalis* , 3 *V. labrusca*, *1 V. lincecumii, 5 M. rotundifolia*). Les polymorphismes représentés sur cette puce sont repartis sur le génome entier, principalement dans des régions géniques.

Ces marqueurs peuvent être utiles pour mieux comprendre la structure du génome et l'évolution de l'espèce – en étudiant le DL au long du génome et l'apparentement entre individus – et ils peuvent également contribuer à l'amélioration variétale par la sélection assistée par marqueurs (SAM). En effet, les informations génomiques, obtenues à moindre coût par rapport aux données phénotypiques, peuvent être valorisées par différentes méthodes pour prédire au stade plantule les capacités phénotypiques de la plante adulte.

2.4. Les programmes de création variétale chez la vigne

Les étapes historiques de la création variétale chez la vigne sont présentées dans la chapitre 2.1 (LE CUNFF *et al.* 2013). Ici nous résumerons brièvement les plus grandes étapes et focaliserons notre analyse sur la période actuelle avec l'apparition des marqueurs moléculaires.

2.4.1. Avant les marqueurs moléculaires

Plusieurs périodes peuvent être définies, pendant lesquels la création variétale a été active. De la domestication au 19ème siècle, les variétés ont évoluées, l'encépagement s'est diversifié, au travers notamment des croisements avec les vignes sauvages afin de permettre une adaptation des cépages à des conditions nouvelles : expansion de la vigne vers l'ouest ainsi que vers l'est et le nord,. Elle s'est faite essentiellement par les producteurs et de façon non intentionnelle. C'est à cette période que sont ainsi apparus les cépages renommés tels que le Cabernet-Sauvignon (BOWERS and MEREDITH 1996), et le Chardonnay (BOWERS *et al.* 1999).

Avec la colonisation d'Amérique du Nord, des nouvelles *Vitis* ont été découvertes, comme *V riparia* et *V. labrusca*, qui poussait dans un environnement humide, sous les arbres. A cause de leur mauvaise qualité, les *Vitis* américaines n'ont pas été cultivées pour leur fruits, ils étaient utilisés dans des croisements en tant que géniteurs donneur de résistances aux pathogènes locaux (mildiou, oïdium et phylloxéra). Les croisements précoces ont donné des variétés comme l'Isabella, le Catawba et le Norton en utilisant *V. aestivalis* et *V. labrusca*. Plus tard ces espèces cousines (*V. riparia* et *V. labrusca)* ont été utilisées comme source de tolérances au froid (PINNEY 1989). Les espèces *V. rupestris* et *V. lincecumii* étaient moins résistantes que les autres, mais en revanche elles avaient une meilleure qualité gustative. Ces hybrides sont devenus importants en Europe lors de la crise du phylloxéra, ce qui a provoqué une augmentation de la demande pour ces porte-greffes résistants au phylloxéra (HUSMANN 1880).

Le 19ème siècle avec la crise phylloxérique et l'arrivée des maladies cryptogamiques que sont l'oïdium et le mildiou, a été une période de création très intense en Europe aussi. C'est en effet dès 1860, que sont créés les porte-greffes par hybridation interspécifique puis le greffage qui permet de lutter efficacement contre le phylloxéra. La lutte contre le mildiou et l'oïdium a été chimique (utilisation « massive » de produits phytosanitaires, notamment la bouillie bordelaise et le souffre) mais aussi génétique avec la création des hybrides producteurs directs (HPD) en croisant des accessions d'espèces sauvages américaines et des variétés de *Vitis vinifera* de bonnes qualités gustatives afin de créer des variétés résistantes à la fois au phylloxera, au mildiou et à l'oïdium (THIS *et al.* 2006). Des sélectionneurs français ont été très actifs à cette période : notamment Victor Gazin qui a travaillé avec *V. rupestris* alors qu'un amateur sélectionneur Eugène Contassot a planté des boutures d'hybrides importés d'Amériques issus des travaux d'Hermann Jaeger dont la variété Jaeger 70 (GALET 1988). Les descendants issus du croisement entre cet Hybride et des individus de l'espèce *V. vinifera* ont donné des nombreux HPD et porte-greffes résistants au mildiou, qui ont connus une certaines renommée, obtenus grâce aux travaux de Georges Couderc et Albert Siebel (PAUL 1996). Ces variétés hybrides ont connu un grand succès entre 1875 et 1940, mais à cause de leur qualité gustative insatisfaisante et dans certains cas d'une teneur en méthanol supérieure à la moyenne, elles ont était abandonnées en France et dans l'Europe de l'ouest.

Des programmes ont également été repris dans les années 60, en Allemagne, par croisement d'hybrides producteurs anciens avec des variétés de *Vitis vinifera* et ont conduit à la création de quelques cépages résistants à qualité organoleptiques acceptables et un niveau plus ou moins important de résistance au mildiou : Regent, crée en 1967, a ainsi été inscrit au catalogue allemand en 1996. D'autres croisements ont également été réalisé ayant conduit à divers cépages plus ou moins résistants (Cabernet Cortis, Muscaris). En France, l'INRA de Bordeaux a également réalisé de tels croisements qui ont abouti à des variétés classées dans la catégorie d'Agrément, utilisables par les amateurs (Aladin, Amandin, Candin, Perdin) (CHANFREAU and ROUSSEAU 2013). Pendant cette même période, les Russes ont déjà réalisé des croisement entre *V. amurensis* et *V. vinifera* pour apporter de nouvelles sources de résistance à l'oïdium et au mildiou mais également contre le froid (VENUTI *et al.* 2013). Les premiers hybrides contenant dans leur généalogie des espèces asiatiques et américaines (Bronner, Solaris) ont été créés en Allemagne (Freiburg WI, Allemagne) (CHANFREAU and ROUSSEAU 2013). Cependant aucune donnée moléculaire ne valide le pyramidage de plusieurs sources de résistances.

En France, à cette même époque, A. Bouquet, chercheur à l'INRA de Bordeaux puis de Montpellier a misé sur l'utilisation de résistances monogéniques provenant de *Muscadinia rotondifolia*. A partir d'un hybride NC6-15, produit par le Pr Olmo (Californie), il a croisé différents

individus des recroisements successifs, totalement résistants à l'oïdium et au mildiou avec différentes variétés de *Vitis vinifera* pour produire en 5ème et 6ème retro-croisements des individus totalement résistants et avec un niveau de qualité organoleptique acceptable (BOUQUET *et al.* 2000).

Une des résistances issue des Vitis américaines a été contournée (Rpv3 contenu notamment dans les variétés hybrides Bianca et Regent). Ce fait a poussé les sélectionneurs à pyramider (accumuler) plusieurs sources de résistance, venant de différentes espèces *Vitis* pour rendre le contournement plus difficile et la résistance plus durable (VENUTI *et al.* (2013), SCHWANDER *et al.* (2012)). C'est cette même option qui a été prise par l'INRA de Colmar qui a lancé un programme de création de variétés à résistances polygéniques, qui seront présenté dans le paragraphe suivant.

En parallèle, des croisements ont été également réalisés pour la création de nouvelles variétés de raisins de table. Dans un premier temps, afin de diversifier la gamme de couleurs, de précocité, de forme et taille de la baie ou d'arôme. Parmi les exemples les plus connus, on trouve la variété Italia, issue d'un croisement entre les variétés Bicane et Muscat de Hambourg (CATALOGUE DES VARIÉTÉS ET CLONES DE VIGNE 2007), ou la variété Muscat de Hambourg, un crosiement entre le Muscat d'Alexandrie et Frankenthal (CATALOGUE DES VARIÉTÉS ET CLONES DE VIGNE 2007). Les programmes se sont ensuite orientés vers la création de variétés sans pépins, dites apyrènes. L'apyrénie étant un caractère assez difficile à travailler, des innovations techniques tels que le sauvetage d'embryons ont été mise au point (JI *et al.* 2013), mais ces méthodes sont très lourdes. L'avènement des marqueurs moléculaires a facilité leur création.

En conclusion, la création variétale chez la vigne a été active, même si les barrières règlementaires, notamment en AOC limitent leur dissémination, ce qui rend le « turn-over » des variétés de vigne moins rapide que chez les autres espèces fruitières. Elle diffère également de la création variétale chez les espèces annuelles de grande culture comme le blé ou le maïs, dans le sens où la sélection récurente n'est pas utilisée, pas plus que des schémas de pré-breeding ou la création de lignées élites issues des schémas de croisements complexes.

2.4.2. SAM chez la vigne

Avec l'arrivée des marqueurs moléculaires, les chercheurs ont initié la recherche de marqueurs liés à la variation des phénotypes d'intérêt d'abord dans des populations biparentales (cartographie des QTL), puis sur des matériels plus diversifiés (génétique d'association). Le nombre de marqueur étant faible, les études se limitaient sur des régions génomiques précises pour identifier des gènes candidats. La première étude qui a utilisé un marquage couvrant le génome entier est celle de MYLES *et al.* (2011) utilisant 5110 SNP génotypes sur 289 individus.

Durant ces 15 dernières années plusieurs caractères monogéniques ou peu complexes – contrôlés par un seul ou quelques gènes à effet fort – ont été étudié avec ces outils et méthodes et des marqueurs intéressants pour la mise en place des programmes de SAM ont été identifiés (et même utilisés). Les caractères mono- et oligo-géniques peuvent être facilement repérés dans le génome par une cartographie de QTL. Le rôle de la génétique d'association se limite dans ce cas à la cartographie fine des gènes candidats. Les marqueurs ainsi identifiés peuvent être utilisés en sélection pour suivre les allèles favorables dans des croisements. Chez la vigne ce genre d'études a été réalisé sur la couleur de la baie (THIS *et al.* 2007; FOURNIER-LEVEL *et al.* 2009), le goût muscat (EMANUELLI *et al.* 2010), l'apyrénie (MEJIA *et al.* 2011), la résistance aux maladies cryptogamiques (PAUQUET *et al.* 2001; MARGUERIT *et al.* 2009; DI GASPERO *et al.* 2012). Cependant pour répondre aux demandes actuelles des viticulteurs, les caractères à améliorer sont souvent complexes et structurées (comme la qualité du raisin, la taille de la baie, la résistance aux stress abiotiques). Ainsi, la taille de la baie est un caractère structuré et son phénotype varie en fonction du pool génétique du compartiment cultivé que l'on observe (HOUEL 2011). Nous devons donc tester et appliquer des outils et des méthodes plus performantes pour pouvoir continuer à avancer. Par exemple le rendement, la qualité du raisin, (composition des acides, teneur en sucre) et la tolérance aux stress abiotiques sont des caractères complexes contrôlés vraisemblablement par de nombreux QTLs à faible effet (FANIZZA *et al.* 2005; HUANG *et al.* 2012).

Des programmes de sélection assistée par marqueurs pour créer des cépages résistants ont été initiés ces quinze dernières années. En effet grâce aux avancées de la recherche, on a pu identifier plusieurs sources de résistance dans des espèces américaines (contre le mildiou par exemple: Rpv3 par BELLIN *et al.* (2009)), ou asiatiques (contre le mildiou par exemple: Rpv10 par SCHWANDER *et al.* (2012) ; Rpv12 par VENUTI *et al.* (2013)), et dans le sous-genre *Muscadinia* (contre l'oïdium : Run1 par PAUQUET *et al.* (2001) et contre le mildiou Rpv1 par Merdinoglu et al. (2003). Des génotypes de *Vitis vinifera* ont été récemment identifiés (COLEMAN *et al.* 2009) comme résistants à l'oïdium et porteur du gène Ren1. L'utilisation des marqueurs liés à ces gènes a permis de mettre en place des programmes de SAM, comme ceux menés à l'INRA de Colmar. L'utilisation de ces marqueurs permet de valider la présence de plusieurs gènes de résistance et donc le pyramidage. Cependant un contrôle phénotypique est toujours effectué en condition contrôlé (plateforme de l'INRA de Colmar) et au vignoble.

3. Les dernières avancées de la sélection assistée par marqueurs

Des programmes de sélection assistée par marqueurs (SAM) sont en cours dans un grand nombre d'espèce (XU and CROUCH 2008; COLLARD and MACKILL 2008). Grâce au développement rapide des outils de génotypage haut débit comme le GBS (Genotyping by Sequencing ; ELSHIRE *et al.* 2011; DAVEY *et al.* 2011), les études d'associations peuvent être étendues sur tout le génome (Genome-wide association studies ; GWAS ; ATWELL *et al.* 2010). Mais aussi de nouvelles méthodologies ont pu être développées pour utiliser cette information génétique abondante, comme la sélection génomique (GS ; MEUWISSEN *et al.* 2001), qui prédit la valeur génétique (Genetic-Estimated Breeding Value ; GEBV) des individus génotypés. Dans cette partie nous allons présenter ces stratégies avec un focus sur la sélection génomique, qui semble être la plus prometteuse pour l'amélioration des caractères à architecture complexe.

3.1. Identification des marqueurs pour la SAM via GWAS

3.1.1. Méthodologie de la génétique d'association

Les études de génétique d'associations (GA) cherchent à identifier des polymorphismes (marqueurs) liés au caractère d'intérêt, dans le cas idéal, des polymorphismes causaux. L'idée est de tester la corrélation entre des classes génotypiques différentes (marqueurs moléculaires) et la variation des phénotypes observés à chaque locus. Si les marqueurs recouvrent tout le génome et pas seulement des régions ciblées (gène candidat, ou QTL), on parle de la génétique d'association « genome-wide » (Genome-wide association study – GWAS). Le GWAS scanne donc tout le génome dans le but d'identifier des marqueurs expliquant une partie de la variabilité phénotypique observée pour des caractères d'intérêt (ATWELL *et al.* 2010; HUANG *et al.* 2010; TIAN *et al.* 2011). Un méthode récemment publié – « mlmm » pour multi-locus mixed-model (SEGURA *et al.* 2012) – utilise un algorithme de type « stepwise forward-backward », permettant d'intégrer des associations pertinentes dans le modèle d'association comme cofacteur fixe. Cette approche augmente la puissance de l'étude d'association, et fournit le set de marqueurs qui explique le mieux la variabilité observée.

Puisque l'étude a été menée en comparant un grand nombre de fonds génétiques différents, les marqueurs significatifs peuvent être plus aisément utilisés dans des programmes de SAM, ce qui est un grand avantage de cette méthode. La puissance de détection en GWA est liée à la densité des marqueurs (nombre idéal en fonction de la taille du déséquilibre de liaison) mais aussi au nombre d'individus observés (ZHAO *et al.* 2007; BUCKLER *et al.* 2009; WANG *et al.* 2012). Afin de diminuer les fausses associations (« faux-positifs ») liés à la structure de la population, différentes méthodes ont été proposées. La plus répandue est celle proposée par YU *et al.* (2006), qui via un modèle mixte

prend en compte l'apparentement en effet aléatoire (kinship) et la structure en effet fixe. Ainsi les erreurs de type I et II (faux-positifs et faux-négatifs) sont moins nombreuses. Afin de pallier à cette difficulté, il est possible de réaliser les études dans des sous-échantillons non structurés. Une alternative à cette correction *a posteriori* est l'utilisation de populations ne présentant pas de structure. Ainsi des populations de cartographie dite NAM (Nested association Mapping ; (YU *et al.* 2008; WALLACE *et al.* 2013) ou encore les populations MAGIC (Multiparent Advanced Generation Inter-Cross ; CAVANAGH *et al.* 2008) ont été proposées pour limiter l'effet de la structure. Ce genre de population est cependant très difficile et très long à obtenir chez les espèces pérennes.

Le GWAS est une approche puissante pour travailler sur des caractères peu complexes dont la variabilité est causée par des gènes à effet forts (majeurs). Mais cette méthodologie perd en efficacité pour d'autres types d'architecture génétique des caractères:

- ✓ les caractères polygéniques (complexes) contrôlés par de nombreux QTLs à effet faible, caractéristiques que l'on retrouve chez de nombreux caractères d'intérêt (lié par exemple à la phénologie ou à la qualité),
- ✓ les caractères covariant avec un gradient environnemental ou avec une sélection humaine, introduisant ainsi un effet de confusion dans des tests d'association (CARDON and PALMER 2003; MARCHINI *et al.* 2004).

3.2. La Sélection Génomique (GS)

3.2.1. La méthodologie de la prédiction génomique

La sélection génomique est une méthode de sélection qui, au lieu de sélectionner des marqueurs liés au caractère étudié, utilise l'ensemble de l'information de nombreux marqueurs provenant du génome entier pour expliquer la variance génétique totale. Avec cette approche nous cherchons à prédire la valeur génétique de l'individu observé pour un caractère donné sans identifier les mutations causales. On appelle ces valeurs prédites les « Genetic estimated breeding values » (GEBVs ; (MEUWISSEN *et al.* 2001). L'efficacité de la prédiction se mesure le plus souvent par la précision, qui correspond au coefficient de corrélation de Pearson entre les GEBVs et le phénotype réel. Cette valeur peut être divisée par la racine-carrée de l'héritabilité pour obtenir un estimateur comparable entre caractères (« accuracy »). Pour calculer les GEBVs des individus en cours de sélection (« population candidate ») – dont nous connaissons uniquement les génotypes –, nous utilisons une modèle statistique (« modèle de prédiction ») qui assigne une valeur à chaque

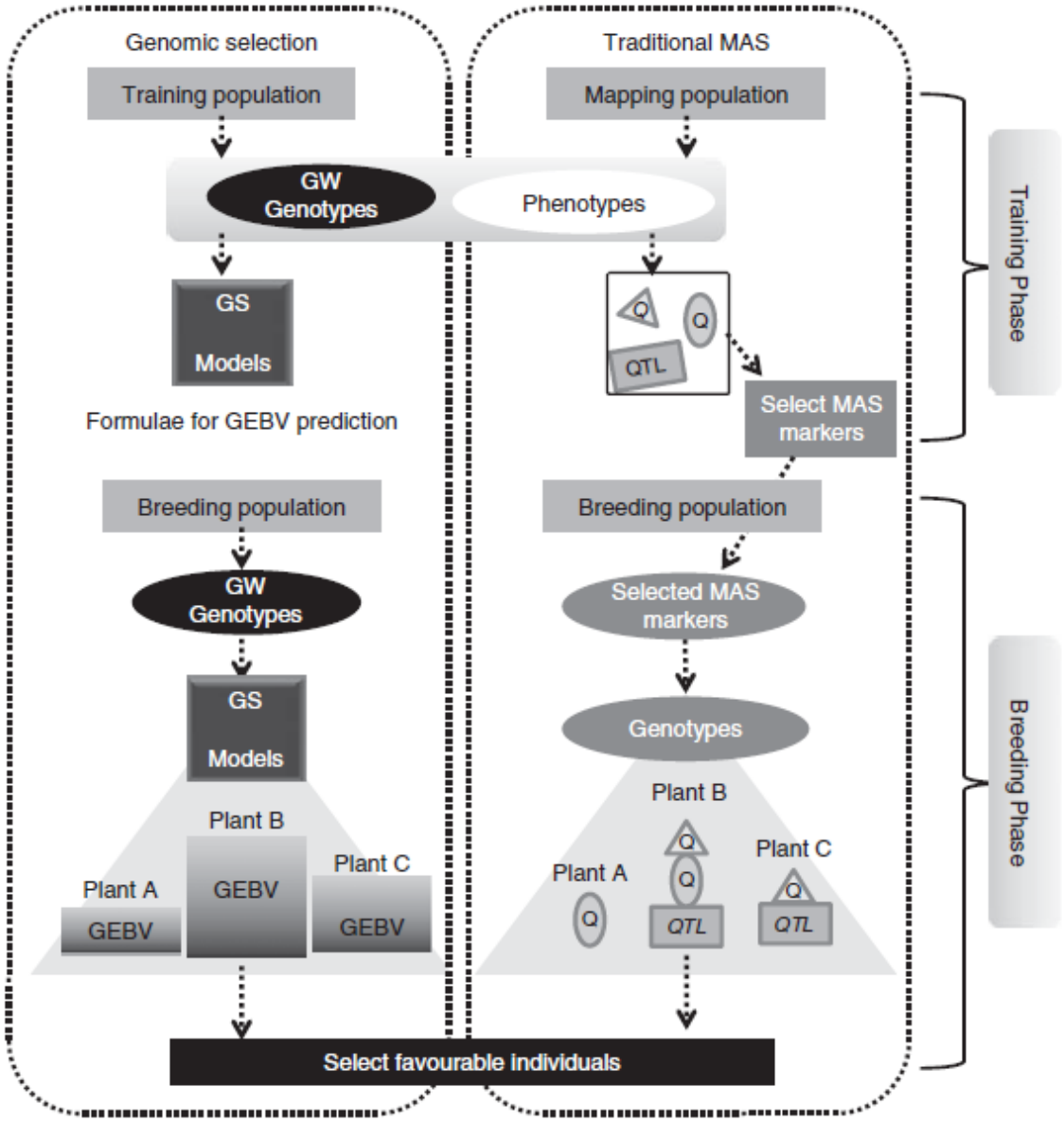


Figure 1.5. Un schéma de sélection génomique (GS ; à gauche) et de la sélection assisté par marqueur (SAM) classique (à droite) issu de NAKAYA and ISOBE (2012). Les deux méthodes sont composées d'une partie « training » (entrainement ou apprentissage) et d'une partie « breeding » (amélioration). Dans la SAM classique, la partie « training » comprend l'identification des associations (QTLs) et dans la GS c'est la définition d'un modèle statistique permettant de prédire des « Genomic Estimated Breeding Values » (GEBVs). La partie « breeding » représente la sélection des individus sur la base de leurs génotypes aux marqueurs identifié pour la SAM classique ou selon leurs GEBVs pour la GS.

marqueur (ou individus avec le modèle GBLUP) à l'aide de paramètres estimées sur une autre population phénotypée et génotypée (« population d'entraînement » ; HEFFNER *et al.* 2009).

Le déroulement d'un cycle de sélection avec la SAM basé sur les marqueurs identifié en GWAS (« SAM classique ») et avec le GS a été comparé par NAKAYA and ISOBE (2012 ; Figure 1.5). Ce schéma met en parallèle la population d'entrainement (« training population ») utilisée pour le GS et le panel de cartographie utilisé pour les études de GWAS (« mapping population ») mais les caractéristiques idéales pour ces deux matériaux de départ ne sont pas forcement les mêmes. A ce sujet, HAMBLIN *et al.* (2011) ont réalisé une étude bibliographique qui analyse les facteurs de « génétique des populations» ayant une influence potentielle sur l'efficacité de ces deux méthodes. Ils montrent que les critères pour définir un échantillon pour réaliser une étude de GWA suffisamment puissante ne correspondent pas toujours aux critères optimaux pour la définition d'une population d'entrainement en GS. La notion de population d'entrainement optimale, est cependant difficile à définir, et c'est justement une des questions clef de la GS actuellement. De nombreux études de simulation, et sur les données réelles, ont identifié et caractérisé des facteurs influençant l'efficacité de la GS. Certains des paramètres sont liés à l'espèce et au caractère étudié – l'**héritabilité** du caractère d'intérêt et son **architecture génétique**, la **portée du DL** dans le génome. D'autres, comme la composition de la **population d'entrainement** (y compris la **taille** et les **relations d'apparentement**), la **densité de marqueur** et le **modèle statistique** pour estimer les GEBVs, le sont moins. Il est difficile d'analyser ces paramètres séparément car finalement ils sont tous relié par des lois de la génétique des populations (discuté dans HEFFNER *et al.* 2009; JANNINK *et al.* 2010; HAMBLIN *et al.* 2011). Nous présenterons maintenant quelques un de ces paramètres.

La **densité de marquage** est un critère important en GS (comme en GWAS). Elle dépend du **niveau de DL** au tour des QTLs et dans le génome en général. Le but est d'avoir un marqueur en DL avec le polymorphisme causal (QTL) qui représentera son effet sur le phénotype. Plusieurs études théoriques montrent, que plus le DL est fort entre SNP et QTL, plus la prédiction est fiable (MEUWISSEN *et al.* 2001; CALUS *et al.* 2008; SOLBERG *et al.* 2008). Ces résultats suggèrent que le cas idéal serait d'avoir autant de marqueur que d'unité de génome en DL complet, ou de « fragment chromosomique indépendant » (Me ; aucun individu ne présente de recombinaison dans ce fragment) dans la population d'entrainement (GODDARD 2008; HAYES *et al.* 2009c). Le niveau du DL dans une population dépend de la taille efficace (Ne) de cette population (SVED 1971). SOLBERG *et al.* (2008) et de DAETWYLER *et al.* (2010) ont proposé des modèles pour estimer le nombre de marqueur nécessaire pour la sélection génomique en fonction des paramètres Me et Ne.

L'efficacité de la sélection génomique repose également sur les **relations d'apparentement** entre les individus (HABIER *et al.* 2007, 2010). Il est connu que les liens de parentés structurent le génome ce qui génère des associations entre le phénotype et des marqueurs non-liés (ATWELL *et al.* 2010). Ces associations sont indésirables en GWAS, mais en GS elles fournissent de l'information sur le fond génétique de la population d'entrainement. Cette information améliore la précision de la prédiction quand la population candidate est apparentée avec la population d'entrainement. Mais elle devient moins pertinente quand les deux populations sont éloignées (MEUWISSEN 2009; DE ROOS *et al.* 2009; CLARK *et al.* 2012). Les études de prédiction entre populations éloignés ont montré que la précision peut être améliorée en augmentant la **densité de marquage** (HAYES *et al.* 2009b; IBANEZ-ESCRICHE *et al.* 2009; DE ROOS *et al.* 2009; TOOSI *et al.* 2010; KIZILKAYA *et al.* 2010). Ce résultat peut être expliqué par la **porté du DL** qui diminue en intégrant de la diversité dans le panel étudié (DE ROOS *et al.* 2008, 2009). En effet, la portée du DL dans une population dépend de sa taille efficace (SVED 1971), quand la taille efficace est plus grande, la diversité génétique est aussi plus importante, cependant le niveau d'apparentement entre individus baisse ainsi que la porté du DL (FALCONER and MACKAY 1996).

Plusieurs études ont montré que la prédiction devient plus précise en augmentant le **nombre d'individu dans la population d'entrainement** – sans trop s'éloigner du fond génétique de la population candidate – (WONG and BERNARDO 2008; LORENZANA and BERNARDO 2009; VANRADEN and SULLIVAN 2010). En effet, l'accumulation de l'information phénotypique rend plus robuste l'estimation des effets des marqueurs (HAYES *et al.* 2009a). Mais le nombre d'individu d'entrainement n'est pas un critère suffisant, sa composition est aussi très importante. Ils doivent être bien représentatifs de la population candidate en terme **d'apparentement** – comme décrit dans le paragraphe précédant – et de diversité allèlique. Des concepts et des méthodes ont été proposés pour optimiser la composition de la population d'entrainement en prenant en compte de ces informations (ALBRECHT *et al.* 2011; PSZCZOLA *et al.* 2012; RINCENT *et al.* 2012).

Les premières études de simulation ont montré que la prédiction des caractères peu héritables était moins précise que pour les caractères à **héritabilité** élevée (BERNARDO and YU 2007). Cependant, comparée à la SAM classique, la GS est plus performante pour améliorer les caractères complexes à faible héritabilité (HEFFNER *et al.* 2009; HAYES *et al.* 2009a; NAKAYA and ISOBE 2012). La précision peut être améliorée en utilisant un plus grand échantillon d'individus phénotypé (augmenter le **nombre d'individu dans la population d'entrainement**). HAYES *et al.* (2009a) ont montré que quand la taille efficace de la population d'entrainement est grande (Ne=1000), la diminution de l'héritabilité de 0,6 à 0,2 nécessite 4 fois plus d'individus dans la population d'entrainement pour obtenir la même précision (entre 0,6 et 0,8).

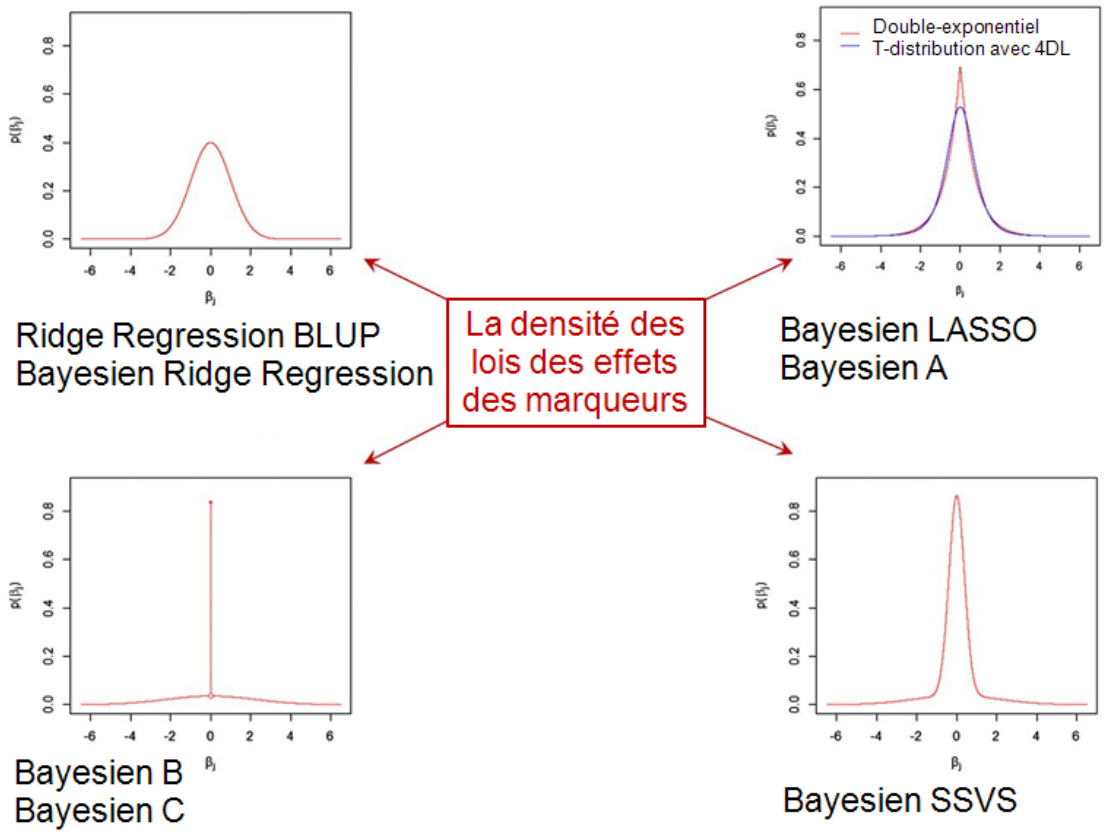


Figure 1.6. Les lois (« priors ») les plus souvent utilisées dans les modèles de la sélection génomique (d'après DE LOS CAMPOS *et al.* 2012). En haut à gauche se trouve la densité de la loi Gaussienne. Dans le sens des aiguilles dd'une montre : la loi double exponentielle (Bayesien LASSO) avec une loi de t à 4 degré de liberté (Bayes A), une loi composée (comme dans Bayes Stochastic Search Variable Selection) et une loi présentant une masse à zéro et quelques effets forts (Bayes B et C).

Depuis la première étude sur le GS de nombreux **modèles de prédictions** ont été développés et testés sur des jeux de données simulées et réelles (MOSER *et al.* 2009; LUND *et al.* 2009; HESLOT *et al.* 2012; DE LOS CAMPOS *et al.* 2012a). Le défi statistique de la sélection génomique est d'estimer l'effet de tous les marqueurs dans le même modèle de régression, sachant que le nombre de marqueurs (les variables explicatives, *p*) est nettement plus important que le nombre d'individus (observations, *n*). Ces conditions – appelé aussi *p>>n* – mènent à un manque de degré de liberté qui peut être résolu i. par la sélection de variables ou ii. en appliquant un « shrinkage[1] » sur les effets estimés, ou iii. une combinaison de ces deux. Les GEBVs sont le plus souvent estimés avec un modèle linéaire de la façon suivante :

$$GEBV = \mu + \sum_{j=1}^{p} x_{ij}\, \beta_j,,$$

où μ est la moyenne, x_{ij} est le génotype de l'individu *i* pour le marqueur *j* (*j*=1,…*p*) et β_j est l'effet de ce marqueur. Mais des solutions semi-paramétriques ont également été proposées (régression Reproducing Kernel Hilbert Space « RKHS », ou Neutral Networks « NN ») ; de plus elles semblent être très performantes (GIANOLA *et al.* 2010). Les méthodes estimant des GEBVs par régression linéaire sont cependant plus utilisées pour le moment. La revue de DE LOS CAMPOS *et al.* (2012) propose une description et une comparaison détaillée de ces méthodes avec l'explication des approches statistiques.

D'une manière générale, les effets attribués aux marqueurs (β_j) sont issus d'une loi (« prior ») dont les paramètres sont estimés sur la population d'entrainement. Les modèles imputent les effets d'après des lois autorisant différents type d'effets allant de fort à faible, et seulement dans certains modèles à zéro. La Figure 1.6 (DE LOS CAMPOS *et al.* 2012a) présente les principaux types de loi et le nom de la méthode qui les implémente. Brièvement, l'utilisation d'une distribution normale autorise beaucoup d'effets faibles et moyens mais très peu d'effets forts (extrême). Elle est implémentée dans la méthode Rigde Regression-BLUP (HOERL and KENNARD 1970) avec une approche fréquentiste, et dans la méthode Bayesien Rigde Regression (BRR) par une approche probabiliste (Bayesien). Une autre méthode souvent utilisée est le G-BLUP (Genome enabled Best Linear Unbiaised Prediction) qui est statistiquement équivalente au RR-BLUP sous les conditions présentées par HABIER *et al.* (2007) et HAYES *et al.* (2009d). Les autres distributions présentées sur la Figure 1.6, suivent une logique d'application d'un « shrinkage » vers zéro crescendo (c'est-à-dire qu'ils augmentent le nombre de marqueurs à effets très faibles, tout en autorisant quelques marqueurs à effets forts (extrêmes)). Le modèle de Bayes A (MEUWISSEN *et al.* 2001) utilise une distribution de type t-distribution qui permet

[1]Processus pour resserer la distribution des effets le plus possible autour de 0.

une capacité de « shrinkage » similaire à une distribution double-exponentielle utilisé dans le cadre du modèle de type LASSO (Least Absolute Shrinkage and Selection Operator) implémenté par un approche Bayesienne (PEREZ *et al.* 2010). A la différence du modèle Bayes A, les méthodes de type Bayesien LASSO (PARK and CASELLA 2008), Bayes B (MEUWISSEN *et al.* 2001) et Bayes C (HABIER *et al.* 2011) cherchent aussi à réaliser un « shrinkage » important mais combiné à une sélection des variables afin de répondre aux contraintes du problème statistique p>>n. Pour cela, ils fixent l'effet d'un certain nombre (ou pourcentage) de marqueurs à zéro, et pour le reste les effets sont issus d'une loi double exponentielle en Bayesien LASSO, d'une loi normale en Bayes B et d'une t-distribution en Bayes C. Dans le cadre de la méthode Bayes SSVS (Stochastic Search Variable Selection ; CALUS *et al.* 2008), c'est la combinaison de deux lois (deux lois normales ou deux distributions-t) qui permet de régler plus précisément la proportion des effets faibles par rapport aux effets forts

Le **choix du modèle** dépend principalement du caractère étudié, car le but est de trouver la méthode qui modélise le mieux possible la distribution réelle des effets des QTLs, c'est-à-dire l'**architecture génétique du caractère**. S'il s'agit d'un caractère complexe contrôlé par de nombreux QTLs à effet faible, les méthodes de RR-BLUP et BRR sont les plus adaptées, mais si on suppose l'effet de gènes majeurs, Bayes B ou C sont conseillées. Cette théorie a été confirmée par les résultats des simulations, mais sur les données réelles – disponibles depuis ces deux dernières années – très peu de différences ont été observées entre les méthodes (HESLOT *et al.* 2012; RESENDE *et al.* 2012a). En conclusion, DE LOS CAMPOS *et al.* (2012) recommandent l'utilisation des méthodes RR-BLUP (ou GBLUP), Bayesien LASSO et la Bayes A, éventuellement le Bayes B, si des gènes majeurs sont supposés et si la densité de marquage est élevée.

3.2.2. L'intérêt de la GS

La GS a fait ces premières armes dans la sélection des vaches Holstein (HAYES *et al.* 2009a), elle a été mise en pratique et elle a révolutionné l'amélioration des bovins laitiers (PRYCE and DAETWYLER 2012; BOUQUET and JUGA 2013). Elle a été évaluée pour autres espèces animales – par exemple les poules pondeuses (SITZENSTOCK *et al.* 2013), les moutons (DAETWYLER *et al.* 2010a) – et pour l'explication de la « missing heritability[2] » de la taille chez l'humain (MAKOWSKY *et al.* 2011). Le potentiel de la sélection génomique a été également étudié chez de nombreuses plantes autogames et allogames (plus hétérozygotes), sur des populations biparentales et des familles connectées. D'abord sur les données simulées – par exemple chez le maïs (BERNARDO and YU 2007; BERNARDO

[2] C'est-à-dire l'incapacité de la génétique d'association « genome-wide » d'expliquer une part très importante de la variabilité des caractères (>80%).

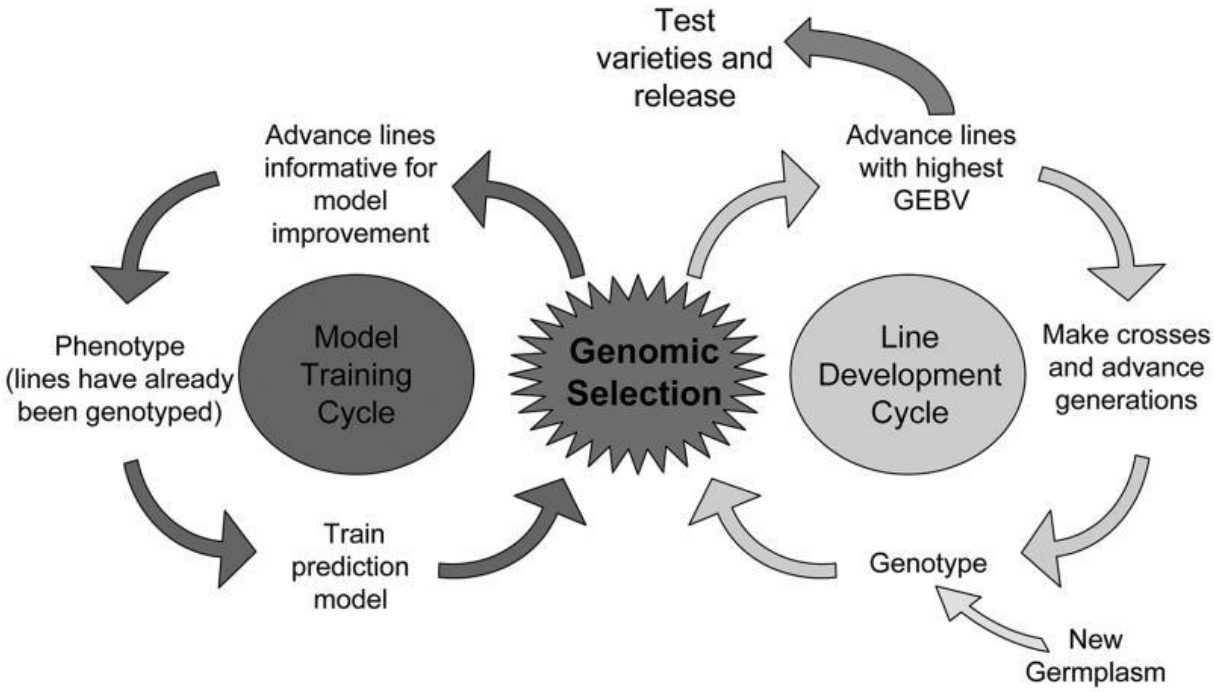


Figure 1.7. Un schéma de sélection génomique avec actualisation du modèle de prédiction (issu de HEFFNER *et al.* 2009). La partie gauche présente l'étape d'entrainement (création puis actualisation) du modèle de prédiction sur des individus génotypés et phénotypés. La partie droite présente la sélection dans le matériel végétal à améliorer, dont nous ne connaissons que les génotypes. Après l'application du modèle de prédiction nous obtenons des GEBVs. Certains génotypes vont être phénotypés (vers la partie gauche) pour actualiser le modèle statistique. Les individus avec un GEBV satisfaisant peuvent ensuite participer à des croisements ou être testés comme candidats de nouvelles variétés.

2009), l'orge (ZHONG *et al.* 2009; IWATA and JANNINK 2011), le palmier à l'huile (WONG and BERNARDO 2008), les arbres forestiers (GRATTAPAGLIA and RESENDE 2010), l'eucalyptus (DENIS and BOUVET 2012) , – et actuellement sur des données réelles – par exemple le maïs (ALBRECHT *et al.* 2011), le blé (CROSSA *et al.* 2010; HEFFNER *et al.* 2011; STORLIE and CHARMET 2013), arabidopsis et l'orge (LORENZANA and BERNARDO 2009), la canne à sucre (GOUY *et al.* 2013), le pommier (KUMAR *et al.* 2012), le pin à l'encens – *Pinus taeda* L. – (RESENDE *et al.* 2012a), l'eucalyptus (RESENDE *et al.* 2012b). Ces résultats suggèrent tous que la sélection génomique à un potentiel supérieur par rapport à la SAM « classique », notamment pour les caractères complexes (même si le nombre d'individus dans la population d'entrainement et le nombre de marqueur doivent être encore augmentés dans certains cas). Plusieurs études sur des données réelles et simulées ont montré que l'intégration de la GS dans des schémas de sélection permettrait un gain du temps considérable sur les cycles de sélection (WONG and BERNARDO 2008; HEFFNER *et al.* 2010; GRATTAPAGLIA and RESENDE 2010; HAYES *et al.* 2013). Ces études soulignent également l'importance d'actualiser le modèle de prédiction : à cause des recombinaisons d'une génération à l'autre, l'effet estimé de DL entre SNP et QTL peut en effet changer ainsi que la structure d'apparentement (MUIR 2007; HEFFNER *et al.* 2009). Un schéma avec actualisation du modèle de prédiction est présenté à la Figure 1.7 (HEFFNER *et al.* 2009).

Actuellement la GS pose encore deux grands défis qui demandent des recherches plus approfondies pour améliorer ses performances. Le premier est le développement de modèles de prédiction prenant en compte les effets de dominance et d'épistasie (DENIS and BOUVET 2012; SUN *et al.* 2012). Le second est la prise en compte de l'interaction entre les génotypes (ou QTLs) et l'environnement (GODDARD and HAYES 2007; HEFFNER *et al.* 2009; CROSSA *et al.* 2010; STORLIE and CHARMET 2013).

En conclusion, la GS est un concept prometteur qui permet de traiter avec efficacité les caractères complexes. Compte tenu des caractéristiques biologiques et génétiques de la vigne, de l'importance des caractères quantitatifs complexes et souvent structurés chez cette espèces, de la difficulté et du coût souvent élevé du phénotypage notamment si l'on veut accéder au vin, des attentes actuelles de la profession pour des nouvelles variétés résistantes aux maladies, la mise en pratique de la sélection génomique chez cette espèce nous semble des plus pertinentes.

4. Objectif de la thèse et démarche expérimentale

4.1. Objectifs de l'étude

L'objectif principal de cette thèse est de tester de nouvelles méthodologies, basées sur les connaissances et les derniers outils de la recherche, pour développer la création variétale chez la vigne. Pour ce faire nous avons testé et comparé deux méthodes : la sélection génomique (GS) et la sélection assistée par marqueur (SAM) « classique » basée sur les marqueurs identifiés par la génétique d'association « genome-wide » (GWAS).

4.2. Démarche expérimentale

Les travaux se sont organisés sur deux axes : i. une étude de simulation et ii. une analyse sur des données réelles.

- ✓ Le but de l'étude de simulation était d'évaluer le potentiel théorique de la GS dans le contexte vigne, sur des caractères simples ou complexes, structurés ou non structurés. Nous avons cherché à identifier les conditions idéales et les facteurs limitants pour la mise en place de la GS chez la vigne.
- ✓ Le but du deuxième axe était d'évaluer quels résultats nous pouvions obtenir avec le GS et le SAM « classique » avec les données et outils actuellement disponibles chez *Vitis vinifera* L. subsp. *vinifera*.

Les travaux réalisés au cours de cette thèse seront présenté dans les trois chapitres suivants du manuscrit.

Le **chapitre 2** correspond au premier axe de la thèse et est présenté sous forme d'un article soumis à la revue Genetics (« Genome-wide prediction methods in highly diverse and heterozygous species: proof-of-concept through simulation in grapevine. »). Il s'agit d'une étude de simulation qui reproduit le génome et la diversité génétique connue chez la vigne afin d'explorer le potentiel des méthodes GWAS et GS dans ce contexte. Cette étude se focalise sur l'effet de la structure de population et la capacité à prédire des caractères structurées, sans la contrainte du nombre d'individus dans la population d'entrainement et du nombre de marqueurs. L'article présente une nouvelle méthode combinant le GS et le GWAS et un scénario de simulation permettant la reproduction des populations de la vigne cultivé.

Les chapitres 3 et 4 correspondent à l'axe deux de la thèse.

Dans le **chapitre 3** (présenté sous forme d'article en préparation ; « Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) in grape for phenotype prediction using a large diversity

panel » qui sera soumis à la revue Theoretical and Applied Genetics), la population d'entrainement est un panel couvrant la diversité connue chez *Vitis vinitera* L. subsp *vinifera*, initialement pensé et dévéloppé pour des études de GWAS (NICOLAS *et al*, *in prep*). La population candidate est une descendance de 23 individus issus d'un croisement biparentale (Syrah x Grenache). Le génotypage est réalisé en utilisant la puce SNP la plus dense existante actuellement chez la vigne qui comprend 18 000 SNPs (LE PASLIER *et al.* 2013)

Le **chapitre 4** présente une étude de cross-validation sur 189 descendants du croisement Syrah x Grenache, génotypés sur 127 marqueurs microsatelites (SSR).

Enfin, le **chapitre 5** correspond à la discussion générale du travail et aux perspectives et le **chapitre 6** regroupe la liste des références bibliographiques, y comrpis celles figurant dans les articles.

Ce manuscrit comprend aussi **3 annexes** :

- ✓ **L'Annexe I.** contient les données supplémentaires de l'article présenté dans le chapitre I et soumis à Genetics (« Genome-wide prediction methods in highly diverse and heterozygous species: proof-of-concept through simulation in grapevine »).
- ✓ **L'Annexe II.** contient les données supplémentaires de l'article en préparation présenté dans le chapitre II (« Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) in grape for phenotype prediction using a large diversity panel »).
- ✓ **L'Annexe III.** contient un rapport rédigé dans le cadre d'une formation de l'école doctorale « Valorisation des compétences – un Nouveau Chapitre de la Thèse® ». Cette formation, de 10 jours, a pour objectif de valoriser la préparation du doctorat comme une première expérience professionnelle de gestion de projet.

CHAPITRE 2 : EVALUATION DE L'INTERET DE LA SELECTION GENOMIQUE PAR SIMULATIONS

1. Introduction

Au début de cette thèse la sélection génomique avait déjà été utilisée chez les animaux, notamment les bovin laitiers, mais pas encore chez les plantes. Les études de simulations ont commencé à évaluer l'efficacité de la GS surtout pour les espèces de grandes cultures – maïs, orge, blé (BERNARDO and YU 2007; HEFFNER *et al.* 2009; ZHONG *et al.* 2009; LORENZANA and BERNARDO 2009) – pour lesquelles les lignées élites sont issues de schémas de croisement complexes et de larges pédigrées sont disponibles. Mais aucun test n'avait été réalisé sur des espèces comme la vigne, le café ou les citrus, pour lesquelles les sélectionneurs gèrent une vaste diversité (souvent structurée) pour créer des nouvelles variétés qui sont souvent issues de croisements simples entre 2 variétés existantes sélectionnées sur leurs qualités.

Dans la mesure où aucune donnée moléculaire n'était disponible sur quelque matériel végétal que ce soit, il est vite apparu indispensable d'avoir recours à la simulation, comme cela avait été le cas pour les nombreuses espèces déjà citées. Notre étude de simulation avait pour principal objectif de tester les avantages et les limites de la GS chez une plante comme la vigne qui se caractérise par une forte hétérozygotie, une large diversité et l'existence de pools génétiques différenciés structurés autour de caractéristiques morphologiques propres. Elle avait également des objectifs plus appliqués, à savoir quel nombre de marqueurs moléculaires étaient nécessaires pour envisager une étude avec des résultats suffisamment intéressants, et quels types de caractères allaient pouvoir être travaillés (en termes de structuration, complexité et héritabilité notamment).

La première étape de ce travail a été de simuler des données génotypiques et phénotypiques correspondant au matériel végétal disponible dans l'équipe de recherche et sur lequel pourraient être réalisés ultérieurement des études de GWAs et GS. L'équipe DAVEM de l'UMR AGAP, a la responsabilité scientifique de la collection de ressources génétiques du domaine de Vassal, la plus grande collection mondiale de ressources de la vigne. Des études récentes sur cette collection ont montré que la collection est structurée en 3 pools génétiques relativement bien différenciés (BACILIERI *et al.* 2013), et que la diversité de cette collection reflétait bien l'histoire de la vigne.

Partant de ce résultats, nous avons donc simulé (une version possible de) l'histoire de la vigne, impliquant la domestication dans la région Caspienne, il y a 7 000 ans, puis des étapes de migration, telle que la vigne les a connus avec les différentes civilisations grecques, romaines, étrusque, arabes et chrétiennes (THIS *et al.* 2006) et des étapes de domestications secondaires avec participation du pool sauvage (*V. vinifera* subsp *sylvestris*). Ce sont les données issues de la simulation (données génotypiques de type SNPs et données phénotypiques avec 4 caractères présentant des caractéristiques bien distinctes) qui ont servi de point de départ pour tester la sélection génomique.

Les programmes de création en cours, ou qui devraient démarrer en lien avec les professionnels de la filière, seront dans un premier temps des croisements simples impliquant des variétés spécifiques de différentes régions (Cabernet-Sauvignon en Bordelais, Ugni blanc en Cognac, Chardonnay en Champagne, Pinot N en Bourgogne) et du matériel résistant issus des programmes de sélection et qui se rapproche fortement des pools *Vitis vinifera*. Il nous a donc semblé pertinent de tester par simulation des populations candidates issues de croisement entre individus appartenant à l'un des 3 pools génétiques, permettant aussi de vérifier l'effet de l'apparentement entre populations candidates et d'entrainement.

Les résultats de cette étude ont donné lieu à une publication soumise au journal Genetics, et à laquelle ont collaboré, Samuel Neuenschwander, qui a développé le logiciel quantiNEMO, Vincent Ségura et Alexandre Fournier-Level et divers collègues de l'UMR AGAP.

2. Evaluation de l'intérêt de la sélection génomique par simulations

Premier article de la thèse, soumis pour publication dans la revue Genetics

Genome-wide prediction methods in highly diverse and heterozygous species: proof-of-concept through simulation in grapevine.

Agota Fodor[*, §], Vincent Segura[†], Marie Denis[‡], Samuel Neuenschwander[**,§§], Alexandre Fournier-Level[††], Philippe Chatelet[§], Félix Abdel Aziz Homa[‡], Thierry Lacombe[§], Patrice This[*, §], Loic Le Cunff [*,§]

[*] UMT Geno-Vigne[®], IFV-INRA-Montpellier Supagro, Montpellier, France, 34060, [§] UMR AGAP, INRA, Montpellier, France, 34060, [†] UR 588 AGPF, INRA, Orleans, France,45075, [‡] UMR AGAP, CIRAD, Montpellier, France, 34398, [**] University of Lausanne, Department of Ecology and Evolution, Lausanne, Switzerland, 1015, [§§] University of Lausanne, Swiss Institute of Bioinformatics, Vital-IT, Lausanne, Switzerland, 1015, [††] Department of Genetics, The University of Melbourne, Parkville, Australia, 3010

Short running title: Combined genome-wide prediction

Key words: Genome-wide association, Genomic selection, Structured trait, Highly diverse training population, Grapevine.

Corresponding author: Loïc Le Cunff

Address: INRA, UMT Geno-Vigne, 2 place Viala, 34060 Montpellier, France

Tel: +33 4 99 61 30 97

E-mail: loic.lecunff@supagro.inra.fr

ABSTRACT

Nowadays, genome-wide association studies (GWAS) and genomic selection (GS), methods which use genome-wide marker data for phenotype prediction, are of much potential interest in plant breeding. However, to our knowledge, no studies have been performed yet on the predictive ability of these methods for structured traits when using training populations with high level of genetic diversity. Such an example of a highly heterozygous, perennial species is grapevine. The present study compares the accuracy of models based on GWAS or GS alone, or in combination, for predicting simple or complex traits, linked or not with population structure. In order to explore the relevance of these methods in this context, we performed simulations using approx 90,000 SNPs on a population of 3,000 individuals structured into three groups and corresponding to published diversity grapevine data. To estimate the parameters of the prediction models, we defined four training populations of 1,000 individuals, corresponding to these three groups and a core collection. Finally, to estimate the accuracy of the models, we also simulated four breeding populations of 200 individuals. Although prediction accuracy was low when breeding populations were too distant from the training populations, high accuracy levels were obtained using the sole core-collection as training population. The highest prediction accuracy was obtained (up to 0.9) using the combined GWAS-GS model. We thus recommend using the combined prediction model and a core-collection as training population for grapevine breeding or for other important economic crops with the same characteristics such as coffee or *Citrus* species.

INTRODUCTION

Thanks to new sequencing technologies (NGS), use of molecular markers is nowadays much less expensive, allowing the development of genome-wide approaches for characterizing the genetic architecture of complex traits, or for marker assisted selection, such as genome-wide association studies (GWAS) or genomic selection (GS).

Recently, GWAS has been widely used in plant genetics to understand genetic architecture and identify molecular polymorphisms explaining part of the variation for traits of agricultural interest (ATWELL *et al.* 2010; HUANG *et al.* 2010; TIAN *et al.* 2011). These markers can then be used in marker-assisted selection (MAS) programs. GWAS has identified many common alleles of major effect. However GWAS is less efficient to detect associations for structured traits (CARDON and PALMER 2003; MARCHINI *et al.* 2004). Indeed, traits of agricultural interest may be correlated with environmental gradients and lead to confounding effects in the association tests. In a similar way, the impact of human selection may also strengthen population structure, all the "elite" breeds sharing a narrow genetic base, thus leading to false positives (type II errors) in association tests. Moreover the efficiency of GWAS is also impacted by the genetic architecture of the studied trait: indeed, the detection of linked molecular markers in polygenic traits strongly depends both on the size of the sample and on the density of molecular marker used (ZHAO *et al.* 2007; BUCKLER *et al.* 2009; WANG *et al.* 2012).

Genomic selection (GS) is a more recent methodology to make a more efficient use of whole genome information in MAS. In contrast to GWAS methodology which identifies molecular polymorphisms linked to the variation for selected traits, GS allows the prediction of a breeding value – genomic estimated breeding values (GEBV) – for the genotypes tested (MEUWISSEN *et al.* 2001) based on large sets of markers. Previous studies on animal and plant models, based on both simulated and real data demonstrated the interest of GS, especially for capturing small-effect quantitative trait loci (BERNARDO and YU 2007; WONG and BERNARDO 2008; HAYES *et al.* 2009a; GRATTAPAGLIA and RESENDE 2010; HAMBLIN *et al.* 2011). In breeding programs, GS could significantly reduce costs by limiting both size and number of field experiments and facilitating early selection through an efficient use of molecular information. Genotype-based prediction also allows selection in breeding schemes when phenotyping breeding candidates is impossible or difficult (GODDARD and HAYES 2007; HEFFNER *et al.* 2010; JANNINK *et al.* 2010; NAKAYA and ISOBE 2012).

In GS, as the number of markers greatly exceeds the number of individuals, advanced statistical methods are definitely required. In recent years, many different methods were developed to realize these predictions (reviewed and compared in MOSER *et al.* 2009; JANNINK *et al.* 2010; DE LOS CAMPOS *et*

al. 2012). To take into account a large variety of genetic architectures, some models assume that all genomic segments equally affect phenotype, whereas others assume heterogeneity among SNP effects and consider different shapes of the prior distribution for marker effects (Bayesian approaches).

Today, most studies have concentrated on animal models or annual plants, with large pedigrees or complex breeding schemes. However, in several economically important species, such as coffee, orange and grapevine, this type of information and breeding material are not available (no pre-breeding population) due to the biological characteristics of these crops. Grapevine is one of the earliest domesticated fruit crop (ZOHARY 1996) that has been widely cultivated for its fruits and wine. Studying molecular data of a very large set of *Vitis vinifera* L. subsp. *vinifera*, (BACILIERI *et al.* 2013) identified three groups of varieties based on their geographical origin and their use. The most commonly acknowledged scenario (LEVADOUX 1956; ARADHYA *et al.* 2003; GRASSI *et al.* 2003; ARROYO-GARCÍA *et al.* 2006) dates grape domestication back to circa 5000 years BC in the Eastern Caspian region (primary domestication center). Through selection, mostly targeted at large-sized, clear-colored berries and hermaphrodite flowers, a coherent sub-population emerged (denoted "Table-East", TE). Due to human migrations, domesticated varieties were introduced in the Balkans around 4,000 BC where they crossed with local wild individuals and were then selected for small berries to produce wine, forming the group denoted "Wine-East" (WE) group (BACILIERI *et al.* 2013). Finally viticulture arrived in Western Europe around 1,000 BC and wine varieties from the Balkans crossed with local wild individuals forming the "Wine-West" (WW) group.

In grapevine, no advanced breeding lines from complex schemes are available. Instead, breeders are handling a large parental panel with a high diversity both at morphological and molecular level. This material is highly heterozygous (He = 0.76(LAUCOU *et al.* 2011), as a result of a strong inbreeding depression and the predominance of vegetative propagation which maintained a high level of molecular diversity (LIJAVETZKY *et al.* 2007; MYLES *et al.* 2010; LAUCOU *et al.* 2011; CARRIER *et al.* 2012). This panel is also characterized by a low level of linkage disequilibrium (LD) between marker loci (r^2 ~ 0.2 at 5-10 Kb, (LIJAVETZKY *et al.* 2007; MYLES *et al.* 2010). Most of the cultivars are interconnected by a series of first-degree relationships (for example, Pinot noir – Chardonnay – Gouais blanc, Cabernet franc – Merlot ; (BOWERS *et al.* 1999; BOURSIQUOT *et al.* 2009), but the number of connected generations is rather low (MYLES *et al.* 2011; LACOMBE *et al.* 2012). Furthermore some major agricultural traits (for example berry size) are linked to population structure, making association studies difficult (HOUEL 2011).

Since the demand for new grapevine cultivars is increasing for cultivars with sustainable resistance/tolerance traits and well adapted to climate changes (OLLAT *et al.* 2011; MORIONDO *et al.* 2013; HANNAH *et al.* 2013), and since the number of molecular tools available for this species is soaring, GWAS and GS are indeed becoming relevant in this crop. The first set of high density genome-wide molecular markers, developed on eight *Vitis* species by (MYLES *et al.* 2010) comprised 9K SNP (Vitis9KSNP array) and was successfully used for preliminary assessment of germplasm collections. A new 18K genotyping chip is already available (LE PASLIER *et al.* 2013) but will only increase the number of markers available for *Vitis vinifera* L. up to 20K. Because of the rapid decay of LD observed in grapevine (MYLES *et al.* 2010) hundreds of thousands of markers would be necessary to perform efficient GWAS and GS. Such number would only be reached by resequencing hundreds of cultivars. Since developing the resources enabling marker-assisted selection at the whole genome level in grape will still require some heavy work, it is indispensable to perform a preliminary assessment of the feasibility of MAS, targeting structured or unstructured traits using GS in a broad pool of unrelated genetic resources. This will allow testing the limitations and potential uses of GWAS and GS in grapevine through simulated data sets.

In this work we simulated genomic and phenotypic data for a large set of individuals to obtain highly polymorphic, heterozygous, structured populations similar to the present population of cultivated *Vitis vinifera* L. Using these virtual populations, we performed both GWAS and GS for traits of different complexity using a large set of markers compatible with the extent of LD in this species. The objectives were i) to test GWAS ability to detect simulated quantitative trait loci ii) to analyze and to compare the performance of a prediction based on markers identified through GWAS (classic MAS) with all marker using GS methods iii) and to estimate the influence of trait complexity and structure on prediction accuracy, using different combination of training and candidate sets defined in a structured population.

MATERIALS AND METHODS

Simulation: We simulated a population of 3,000 individuals representing the genetic diversity of *Vitis vinifera* L., based on the knowledge available on the history of this species (LEVADOUX 1956; ARADHYA *et al.* 2003; GRASSI *et al.* 2003; ARROYO-GARCIA *et al.* 2006; THIS *et al.* 2006; MYLES *et al.* 2011; LAUCOU *et al.* 2011; BACILIERI *et al.* 2013; EMANUELLI *et al.* 2013).

Simulated genomes comprised the typical 19 chromosomes, each of 79 cM, for a total of 1,500 cM corresponding to the genetic map of grapevine published by (DOLIGEZ *et al.* 2006). Ten thousands markers were randomly positioned on each chromosome, for a total of 189,500 bi-allelic markers

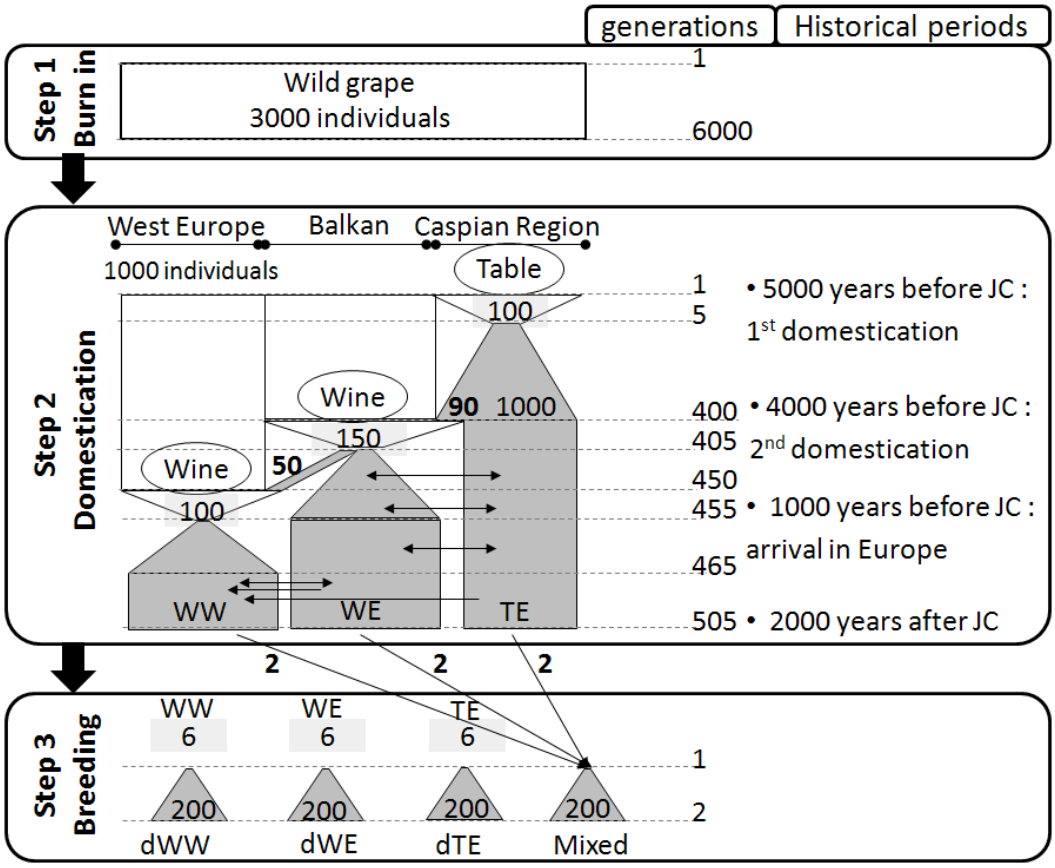


FIGURE 2.**1. Scheme of the demographical scenario based on our working hypothesis on grapevine evolution.** This scheme, implemented with quantiNemo, is composed of three steps: burn in, domestication and breeding. Burn in and domestication steps had the purpose to obtain grapevine diversity groups corresponding to Western Europe wine group (WW), Eastern Europe and Balkan wine group (WE) and Eastern Europe and Caucasus table group (TE) as described by **Bacilieri** ***et al.*** **(2013)**. Breeding step models crosses between selected individuals of these groups. At the right side of the figure are represented generation numbers and historical events with dates. White area represents wild grape, after domestication it is showed in grey. "Wine" and "Table" symbolize the two different definitions of selection applied on the trait under selection (selection optima and intensity). Black arrows show the direction of migration and its intensity is indicated by boldface numbers, specifying the number of migrating individuals. The intensities of the bottlenecks are indicated by normal numbers specifying the number of selected individuals.

(SNP), and 500 multi-allelic markers (SSR, 20 alleles per locus) with a mutation rate of 10e-6 and 10e-4 per generation, respectively (VIGOUROUX *et al.* 2002; DE MITA *et al.* 2013). Considering that genome length in grapevine is 470 Mb (JAILLON *et al.* 2007), one simulated cM corresponds to 300 Kb. We simulated four independents quantitative traits: i) structured simple trait (10 QTL), ii) non-structured simple trait (10 QTL), iii) structured complex trait (100 QTL), iv) non-structured complex trait (100 QTL, under the assumption of strict additivity. QTLs were bi-allelic loci, randomly positioned on the genome. One of the two possible alleles had an effect of zero (no effect on the trait), while the other had an effect randomly sampled from a normal distribution (with mean = 0 and variance=1).

Simulations were carried out with a modified version of quantiNEMO, an individual-based program developed for the analysis of quantitative traits with explicit genetic architecture potentially under selection in a structured population (NEUENSCHWANDER *et al.* 2008). We based our demographic scenario (Figure 2.1) on grapevine domestication history and our goal was to define a scenario matching the published population data (F_{ST}, LD, heterozygosity and population structure ; (MYLES *et al.* 2010; LAUCOU *et al.* 2011; LACOMBE 2012; BACILIERI *et al.* 2013). This demographic scenario consisted in two steps (burn in and domestication) to obtain presently existing material and a third step (breeding) to simulate a breeding program.

In order to simulate a wild, pre-domestication population with realistic allele frequencies and LD between neutral loci at mutation-drift equilibrium, we ran a burn in step as a common starting point for the ten replicates of the domestication step. A single population was simulated with a census population size and carrying capacity of 3,000. It was run for 6,000 generations with random mating to obtain the needed LD level (r^2 value of 0.2 observed at the distance of 10 kb) between neutral markers and to generate enough segregating sites for the next analyses. At the end of the burn in step, fixed loci were removed and individuals were randomly organized in three groups (sub-populations) of 1,000 individuals, forming a meta-population.

Step 2 consisted in the domestication step. It was established to obtain the three diversity groups of the cultivated compartment of *Vitis vinifera* L. subsp. *vinifera* described by (BACILIERI *et al.* 2013) in the Vassal collection : the "Table-East" group (TE) corresponding to the table grape varieties originated from the primary domestication center, localized in the Caucasus, the "Wine-East" group (WE) of wine varieties from the Balkans and Eastern Europe, and the "Wine-West" group (WW) of wine varieties from Western and Central Europe.

It is difficult to estimate the number of generations throughout grape domestication history as grape is a long-lived perennial species. Propagation type varied greatly between vegetative and generative methods at different times and in the different grapevine-growing areas. Based on historical data and

personal communication by J.M. Boursiquot and T. Lacombe, we chose to run the domestication step for about 500 generations. Simulating 505 generations allowed recreating a population structure (F_{ST} and structure) and linkage disequilibrium (LD) pattern similar to what is currently observed in cultivated grape.

The migration rate between each pair of population was set to vary over time in order to fit to historical information and to obtain the needed heterozygosity and F_{ST} between populations at the end of the domestication step. To justify the choice of the migration rates we tested alternative scenarios varying these values between no migration and twice more important migration rate. The size of the bottleneck at the beginning of the domestication was calibrated in the same way, using alternative scenarios without bottleneck and with a bottleneck twice more stringent than in the finally chosen scenario.

Using the same demographic parameters we elaborated two versions with different quantitative trait architectures: simple (quantitative trait controlled by 10 QTLs) and complex (quantitative trait controlled by 100 QTLs) following BERNARDO and YU (2007) and (DE ROOS *et al.* 2009). To simulate quantitative traits linked to population structure, we applied stabilizing selection for the first quantitative trait with both levels of complexity. Intensity and optima of selection varied among populations (to simulate different selection objectives) and over time (time since the selection bottleneck). The genetic architecture of a quantitative trait under selection affects genetic diversity evolution at the sub-population level. In order to maintain the same F_{ST} and to generate similar Q_{ST} (as a measure of phenotypic differentiation among population) for both complexity levels we adjusted the intensity and the optimum of the stabilizing selection in each domestication scenario. The heritability of quantitative traits was set by fixing the environmental variance to achieve a narrow-sense heritability of 0.8 in the first generation of the simulation.

Finally, we added a breeding step, simulating crosses between and within sub-populations, to mimic the effects of a breeding program. Founding individuals were chosen from each of the three sub-populations based on their phenotypic value for the trait under selection. For within sub-populations crosses, we chose the six individuals with the best phenotypic record compared to the selection optimum. For between sub-populations crosses we used the two individuals closest to the phenotypic mean of each sub-population of origin. In this way, we obtained four populations with six individuals in each, producing four times 200 descendants in the next generation via random mating. No selection and migration were used in this final step.

Core collection: MSTRAT software (v 4.1) developed by (GOUESNARD *et al.* 2001) used the M-method proposed by (SCHOEN and BROWN 1993) and allowed the construction of core collections that

maximize the number of observed alleles in the SSR data set. We defined a core-collection from the meta-population of 3,000 individuals using MStrat software and the 500 SSRs. This core-collection (Call) consisted in 1,000 individuals, including the founders of all breeding populations; it was built to represent the genetic diversity of the entire meta-population (all) with the minimum of redundancy (which is the aim concept of core-collection building). In each replicate of the domestication step, five core collections of 1,000 individuals were designed and ranked first by the number of SSR alleles captured; core-collections exhibiting the same allelic richness (determined by the total number of alleles represented) were then ranked using Shannon's index as second criterion. Finally, the core-collection presenting the most significant allelic richness with the highest Shannon's index was selected for further analysis.

Estimation of diversity indices: Diversity indices, such as genetic variance estimates, the level of differentiation in quantitative trait (Q_{ST}) following (SPITZE 1993), and F-statistics following (WEIR and COCKERHAM 1984) for each pair of populations and for all types of markers, were calculated with quantiNemo. To calculate unbiased heterozygosity and compare it to published data (LAUCOU *et al.* 2011) on highly polymorphic SSR markers, we selected all SSR with more than 10 alleles per locus at the end of the domestication step. Data analysis was performed using the "Excel Microsatellite Toolkit" (PARK 2001). We also calculated allele frequency for each SNP and QTL locus, in order to filter out rare SNPs with minor allele frequency (MAF) below 5% that would have biased association tests.

Population structure and relatedness: Population structure was calculated on the 3,000 individuals using 500 SSR with STRUCTURE software version 2.3.3 (PRITCHARD *et al.* 2000) accessed through Bioportal (KUMAR *et al.* 2009). We used an admixture model varying the ancestral number of population (K) from two to five, in order to identify the best K level of population subdivision. Within STRUTURE, we allowed an iterative process with a burn in phase of 15,000 iterations and a sampling phase of 15,000 replicates. Five replicates of each assumed K level subdivision were compared to estimate group assignation stability. Outputs were visualized and interpreted with Structure Harvester web v0.6.93 (EARL and VONHOLDT 2011). The optimal group number was chosen based on the estimated 'log probability of data'.

Realized relationship matrix (RRM; (EDING and MEUWISSEN 2001) was calculated using R (R CORE TEAM 2013) using all filtered SNPs on 3,000 individuals.

Linkage disequilibrium:

LD measures were performed with the R package LDcorSV (MANGIN *et al.* 2011) which corrects for the bias due to population structure and relatedness (r^2SV). LD was measured in two different positions:

in neutral genomic regions and around each QTL. In neutral positions, mean and median values of r^2 were calculated between each pair of SNP within five arbitrarily chosen windows of 600 kb. Around QTLs, r^2 was calculated between the QTL locus and all SNP located within 300 kb. We used the Hill and Weir formula (HILL and WEIR 1988) for describing the decay of r^2SV and we characterized LD by the distance corresponding to a r^2SV value of 0.2.

Genome-wide association:

GWAS were performed using the multi-locus mixed-model (mlmm) approach developed by (SEGURA *et al.* 2012), including the population structure as fixed covariant in the mixed model. This R script implements a forward-backward stepwise approach to include significant effects in the mixed model, while re-estimating the variance components of the model at each step. We run mlmm on the meta-population of 3,000 individuals and on the core-collection with a random polygenic term, with a variance proportional to the estimated RRM and a fixed population structure term (three groups) consisting in ancestry fractions estimated by Structure software. We also run mlmm on each sub-population with a random polygenic term only. Maximal number of forward steps was set to 25. For model selection we chose the multiple-Bonferroni (mBonf) criterion, selecting the largest model in which all cofactors have a P-value below a Bonferroni-corrected threshold (we used a threshold of 0.05). Cofactor effects were re-estimated at the end of the mlmm analysis and used to estimate the genetic value of descendent obtained in the breeding step in the simulation.

Genomic prediction:

We compared four prediction methods based on genome-wide high density SNP data: the sum of effects of markers previously detected in GWAS – using mlmm as described above – corresponding to classical MAS (cof), Ridge Regression BLUP (RR) (HOERL and KENNARD 1970), Bayesian LASSO (Least Absolute Shrinkage and Selection Operator) Regression (BLR) (PÉREZ *et al.* 2010) and a combination of MAS and RR-BLUP (cofRR). We also observed the evolution of prediction accuracy in different combinations of training and candidate populations. Training population always comprised 1,000 individuals, while candidate populations were composed of 200 or 800 individuals. We compared two levels of genetic architecture (10 or 100 underlying QTLs) and prediction accuracy of structured and non-structured quantitative traits (design summarized in Figure 2.S1).

For cof method, effects of significant markers and populations structure were first estimated with a mixed-model together with variances for genetic (polygenic) and residual random effects. In this

model the groups of population structure and the significant markers were declared as fixed effects. Then, in a second step the estimates of the associated markers were used for prediction.

Ridge Regression performs an extent of shrinkage that is homogenous across markers. For RR we defined the parameter lambda as $\lambda = \sigma_e^2/\sigma_g^2$, where environmental and genetic variances (σ_e^2 and σ_g^2) were estimated via REML in a mixed linear model using emma library (KANG *et al.* 2008).

The Bayesian LASSO (PARK and CASELLA 2008) method performs stronger shrinkage toward zero for the estimates of small-effect markers, and less for those with high effects. We performed BLR analysis with the R package BLR 1.3 (PÉREZ *et al.* 2010). The lambda parameter was set as random, sampled from a gamma distribution with rate=0.0001 and shape=0.53 according to (PARK and CASELLA 2008). The initial value of λ_0 was calculated using the heritability rules given by DE LOS CAMPOS *et al.* (2012): $\lambda_0 = 2 * n^{-1} * \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^2 * \frac{(1-h^2)}{h^2}$, where h^2 is the narrow-sense heritability, n is the number of individuals, m is the number of SNPs and X is the matrix of genotypes. σ_e^2 were chosen from the prior $\chi^{-2}(\nu_e, S_e^2)$, where $\nu_e = 4$ to ensure a finite a priori variance, and $S_e^2 = (\nu_e - 2) \times (1 - h^2) \times \sigma_p^2$, where σ_p^2 is the phenotypic variance. σ_g^2 were chosen from the prior $\sigma_g^2 \sim \chi^{-2}(\nu, S^2)$ where ν was set to 4 to ensure a finite a priori variance and $S^2 = (\nu - 2) \times \frac{\sigma_p^2}{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^2} \times h^2$. We allowed an iterative process with a burn in phase of 10,000 iterations and a sampling phase of 40,000 replicates.

In marker-assisted RR (cofRR) we combined RR-BLUP with the effects of markers previously detected with mlmm. Effects of significant markers and population structure were estimated as described for cof method and remaining SNPs were used in a RR model as described earlier. GEBVs were obtained summing the effects of all markers. The R script is available in File S1. Accuracy was calculated dividing the correlation coefficient (r^2) between GEBVs and true phenotypes, by the square root of the narrow-sense heritability.

Test validation on standard data:

cofRR method was tested on a standard data set – i.e. a real data set of loblolly pine described in RESENDE *et al.* (2012) – using a 10-fold cross-validation scheme. Data consisted of 926 individuals genotyped with 4853 SNPs and phenotyped for 17 traits. Information about population structure was not available.

For analysis, markers with more than 20% of missing data were removed in both training and validation sets. For the remaining loci, missing genotypes were imputed with the mean. In the

TABLE 2.1. Population statistics on simulated data for the five scenarios and reference values from published data.

		Domestication step	Published data	Alternative scenarios			
				No migration	Twice more migration	No bottleneck	Double intensity bottleneck
FST	WW-WE	0.04 (0.007)	0.05[a]	0.34 (0.018)	0.01 (0.001)	0.01 (0.001)	0.05 (0.012)
	WW-TE	0.07 (0.012)	0.07[a]	0.35 (0.001)	0.01 (0.003)	0.03 (0.003)	0.09 (0.015)
	WE-TE	0.04 (0.007)	0.05[a]	0.45 (0.014)	0.01 (0.001)	0.03 (0,001)	0.04 (0.008)
Heterozygosity		0.64 (0.026)	0.73[b]	0.46 (0.011)	0.64 (0.019)	0.72 (0.005)	0.60 (0.012)

The standard deviation is between brackets.

[a] LACOMBE (2012), [b] LAUCOU *et al.* (2011).

TABLE 2.2. Descriptive statistics on the simulated meta-population.

		simple trait	complex trait	Real
LD		11 kb		10[a]
SNP number	Total	111,004		-
	polymorphic	92,787.1 (309.5)		-
	MAF>0.05	81,555.0 (845.6)		-
QTL number	Total	2x10	2x100	-
	polymorphic	8.6 (1.03)	83.7 (3.94)	-
	MAF>0.05	7.2 (1.51)	72.2 (4.72)	-
heritability	structured trait	0.71 (0.080)	0.76 (0.037)	-
	Non-structured trait	0.78 (0.034)	0.77 (0.025)	-

[a] MYLES *et al.* (2010)

training set, we applied a 5% filtering on minor allele frequency (MAF>0.05). Kinship matrix (RRM) was calculated as described above. GWAS were performed using mlmm approach setting the maximal number of forward steps to 10. To limit the detection of falseassociated cofactors, we choose the extended Bayesian information criterion (EBIC (CHEN and CHEN 2008)) for model selection, which is more stringent than the multiple Bonferroni criterion (SEGURA *et al.* 2012). Predictions were performed using cof, RR and cofRR methods as described previously.

For the 10-fold cross-validation, individuals were randomly assigned to one of 10 equal folds. Each fold was dropped once from the training set and predicted. Accuracies were calculated as described above using the Mendelian segregation as heritability according to RESENDE *et al.* (2012), and the mean value was reported across all 10 folds.

RESULTS

Simulation:

We built the demographic scenario to simulate *Vitis vinifera* L. history in order to create three genetic pools as observed by (BACILIERI *et al.* 2013). Parameters (migration rate and bottleneck) of the domestication step were defined from bibliographic data. In order to validate the chosen migration rate and bottleneck intensity, we also tested four alternative scenarios i) without migration, ii) with a twice higher migration rate, iii) without bottleneck and iv) with a twice more stringent bottleneck. Ten replicates of each scenario were simulated. Diversity indices (F_{ST}, Q_{ST}, heterozygosity) were calculated for all five scenarios and compared to published data. The values obtained with the domestication step were closer to the expected level than for the alternative scenarios (Table 2.1). Heterozygosity was the only parameter with a value lower than expected (0.64 vs. 0.73), closer to the level observed in natural populations of *Vitis sylvestris* (LAUCOU *et al.* 2011). Changing bottleneck and migration ratio modified all diversity indices.

Descriptive statistics on simulated data: Because of genetic drift and selection, the number of polymorphic loci decreased over time. While, at the beginning of the burn in step (common to the 10 replicated simulations), 189,500 polymorphic SNP loci were defined, 111,004 polymorphic SNP loci only were observed at the end of this step (Table 2.2). After 505 generations, at the end of the domestication step, we observed on average 92,787 (sd = 309.5) polymorphic SNP loci for the entire meta-population of 3,000 individuals. After filtering on minor allele frequency (MAF>0.05) 81,555 SNPs (sd = 845.6) were retained. For both simple and complex quantitative traits (confounded or not with demographic structure) on average 85% of the QTLs were polymorphic and 73% passed the MAF>0.05 filter.

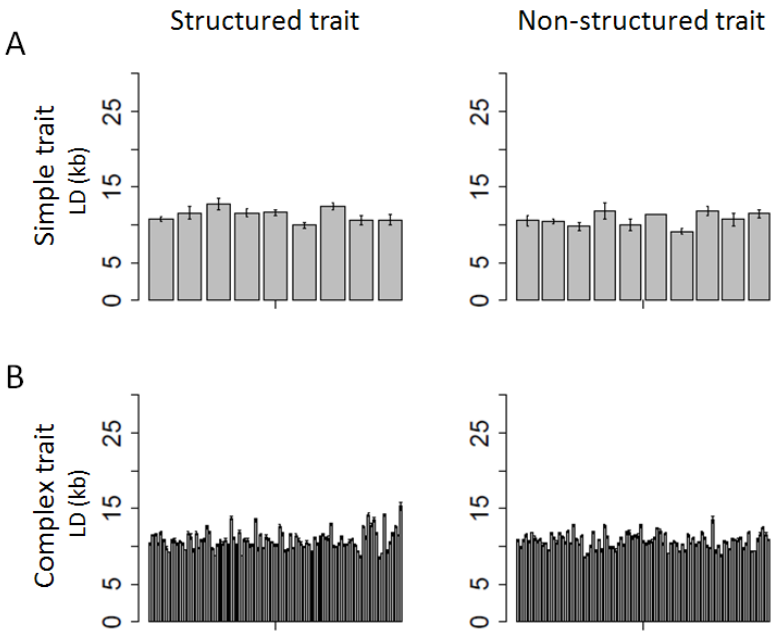


FIGURE 2.**2. Estimation of LD around QTLs.** Mean estimation of LD (in Kb) around the QTLs, calculated at r^2_{SV} =0.2 between all loci in the 600 Kb neighborhood of each QTL locus on 3,000 individuals, for simple traits (A) and complex traits (B) on the 10 replicates of the simulation. The two figures on the left side represent LD around structured trait's QTLs and the other two figures around non-structured traits QTLs. QTL loci were ranked as a function of theirs effects from negative to positive values. Error bars were calculated with 95% confidence intervals on the estimates of the means.

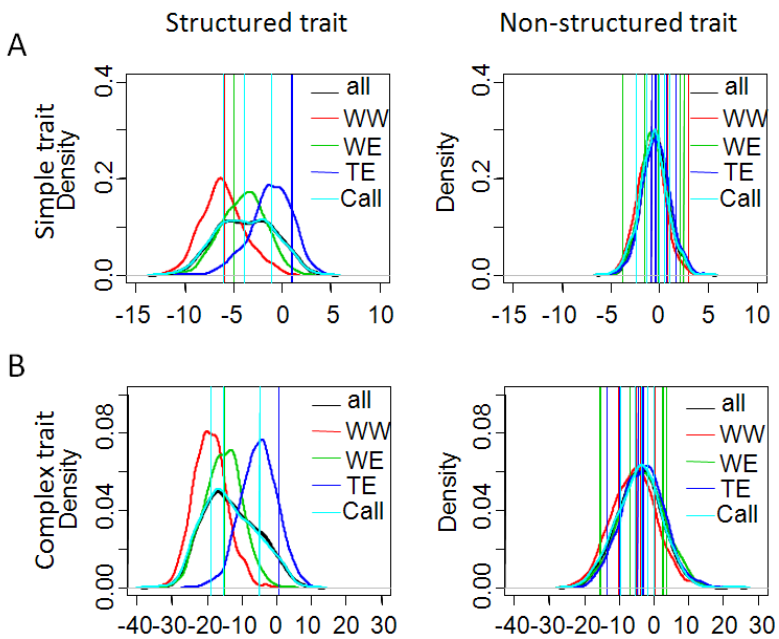


FIGURE 2.**3. Distribution of phenotypes in training (WW, WE, TE) populations.** Distributions are presented on one replicate of the simulation for the structured and non-structured simple (A) and complex (B) traits. The colored vertical lines show the phenotypes of the founder individuals of descendent populations. Call corresponds to the core-collection.

We measured LD decay in both neutral genomic regions and around QTLs. LD in neutral regions decreased rapidly (Figure 2.S2). An r^2_{SV} value of 0.2 was observed over a distance of nine to 13 kb depending on the replicate. This value is consistent with the LD observed over 10kb segments in a set of grape cultivars (MYLES *et al.* 2010). Around QTLs, we observed the same tendency except for structured traits, where LD extended further than 13 Kb in a few cases (Figure 2.2). Consequently, given the extent of LD, the number of SNPs present at the end of the domestication step allowed us to tag all the genome.

The F_{ST} statistics between simulated populations were measured with SSR markers. As expected from observed data (LACOMBE 2012) the historically more distant populations (WW-TE) showed the highest F_{ST} values of 0.07 while historically closer populations displayed lower (approx. 0.04) F_{ST} values (Table 2.1, Figure 2.S3).

The Structure analysis (L(K) method) over the entire meta-population (3,000 individuals) best supported clustering into three ancestral populations in all replicates of the simulation (data not shown) corresponding to the expected three simulated populations: WW, WE and TE.

The narrow-sense heritabilities for the simulated traits at the end of the domestication step were around 0.8 (0.72 to 0.78 for simple trait and 0.76 to 0.77 for complex) conform to initial settings. Q_{ST} was measured as an index of phenotypic distances between each pair of simulated sub-population. Q_{ST} values were always higher for selected traits than for neutral ones (Figure 2.S3). Overall Q_{ST} values reflected F_{ST} values with the TE population diverging more from the other two populations. However, since no published data on Q_{ST} are available yet, we were unable to compare our data with actual observations.

In conclusion, the simulated populations matched observed data reasonably well. We thus considered that the demographic scenario was able to generate pertinent genotypic and phenotypic data allowing further GWA studies and the building of GS models.

Descendent populations:

To simulate a breeding program, we realized breeding crosses using selected individuals from the three original gene pools (Figure 2.1). Three crosses were realized within populations leading to dWW, dWE, dTE, and one between populations leading to Mixed. In the original gene pools, traits distributions for non-structured traits were identical between sub-populations while they were different for the structured traits (Figure 2.3). Variance for simple traits was also smaller than for complex traits.

TABLE 2.3. Results of the GWAs analyses.

		Structured trait					Non-structured trait				
		through 10 replicates			mean per replicate		through 10 replicates			mean per replicate	
		never detected	always fixed	always detected	fixed	detected	never detected	always fixed	always detected	fixed	detected
Simple trait	WW	40%	10%	10%	14%	32%	10%	0%	10%	15%	59%
	WE	30%	10%	10%	16%	37%	20%	0%	10%	15%	57%
	TE	40%	10%	10%	16%	32%	10%	0%	10%	14%	57%
	Call	40%	10%	10%	14%	32%	10%	0%	10%	14%	55%
	all	40%	10%	10%	14%	39%	10%	0%	10%	14%	69%
Complex trait	WW	84%	5%	0%	17%	3%	84%	5%	1%	17%	5%
	WE	77%	5%	0%	17%	5%	86%	5%	1%	17%	5%
	TE	81%	5%	0%	17%	5%	84%	5%	0%	18%	5%
	Call	86%	5%	0%	16%	2%	88%	5%	0%	17%	4%
	all	62%	5%	0%	16%	13%	71%	5%	4%	17%	12%

This table presents the number of positive detection via associated markers of each simulated QTL using the mlmm method, out of the 10 replicates for both simple and complex traits and for structured and non-structured trait

The differences between mean phenotypic values of the breeding crosses and their respective original gene pools were smaller for simple traits than for complex ones (Figure 2.4). It was slightly higher between WW and dWW for non-structured traits compared to the other populations, but the highest difference was obtained between TE and dTE for structured traits.

Differences in phenotypic means were also measured between the breeding crosses and i) those original gene pools without direct parental link ii) the core-collection. We observed greater differences for structured traits than for non-structured ones and for simple traits than for complex ones (Figure 2.4). dTE was always more distant from the other sub-populations. Call behaved similarly to WE, and the Mixed population was closer to TE than to the other populations.

Genome-wide association study (GWAS):

The best mlmm model of each replicate realized on the whole meta-population explained 68 to 83% of the total variance. As expected, the composition of the variance differed between simulated traits (Figure 2.S4). Through the 10 replicates of the simulation of the four training sets (WW, WE, TE, Call, i.e. 1,000 individuals), significant associations were detected for 32 to 59% (on average) of the simulated QTLs in simple traits and 2 to 5% in the complex traits (Table 2.3). For simple traits, one to four QTL only were never detected through replicates, while for complex traits this number ranged from 77 to 88. The proportion of fixed QTLs was similar for all traits, on average 14 to 18% per replicate. Some QTLs were always fixed across the 10 replicates: one in the simple structured trait and five in complex traits. In the case of non-structured traits, one QTL was repeatedly detected across replicates for the simple trait and another QTL was detected in two subpopulations for the complex trait. As expected, more QTL could be identified for non-structured traits than in structured ones, especially with the simple trait (55 to 57%, while in non-structured trait only 32 to 37%). In the full meta-population of 3,000 individuals (all), more QTL were detected than in the training sets of 1,000 individuals, especially for complex traits. In the core-collection fewer QTL were identified than in sub-populations. Manhattan plots of the results in one replicate are shown as supplementary data (Figure 2.S5). In this example, SNPs linked to QTLs were detected for all types of traits with very high P-values (Table 2.S1).

LD measures between QTLs and the cofactors of mlmm showed that significant markers always presented higher LD with the closest QTL, than with other QTLs. However, some cofactors presented quite weak linkage ($r^2 < 0.05$) with the QTL, but strong linkage ($r^2 > 0.2$) with another cofactor, itself tightly linked to the QTL.

Prediction of phenotypes from genotypes:

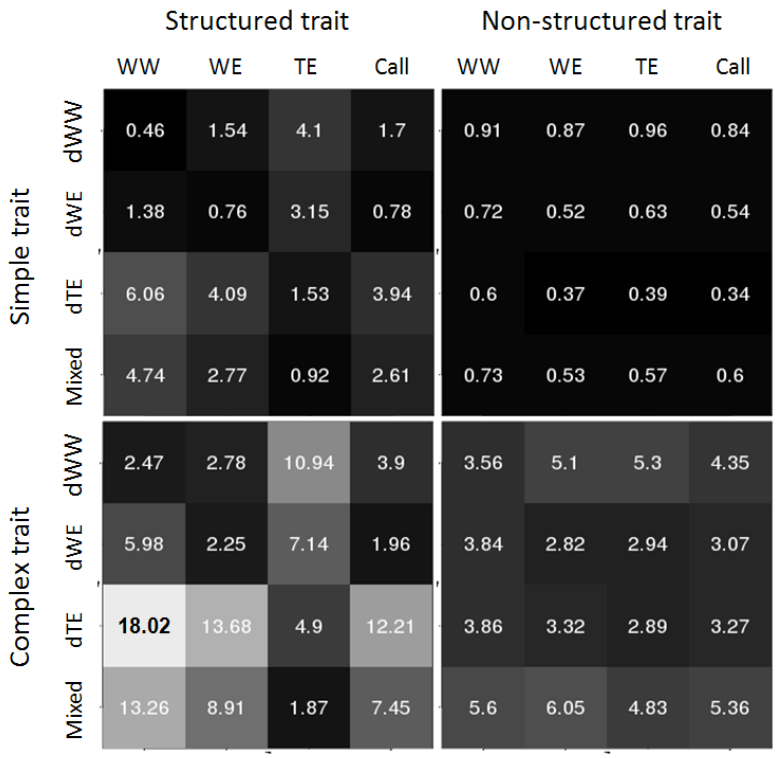


FIGURE 2.**4. Heat map presenting the** d**ifference between the phenotypic mean of training and candidate sets.** Mean values were calculated on the 10 replicates of the simulation.

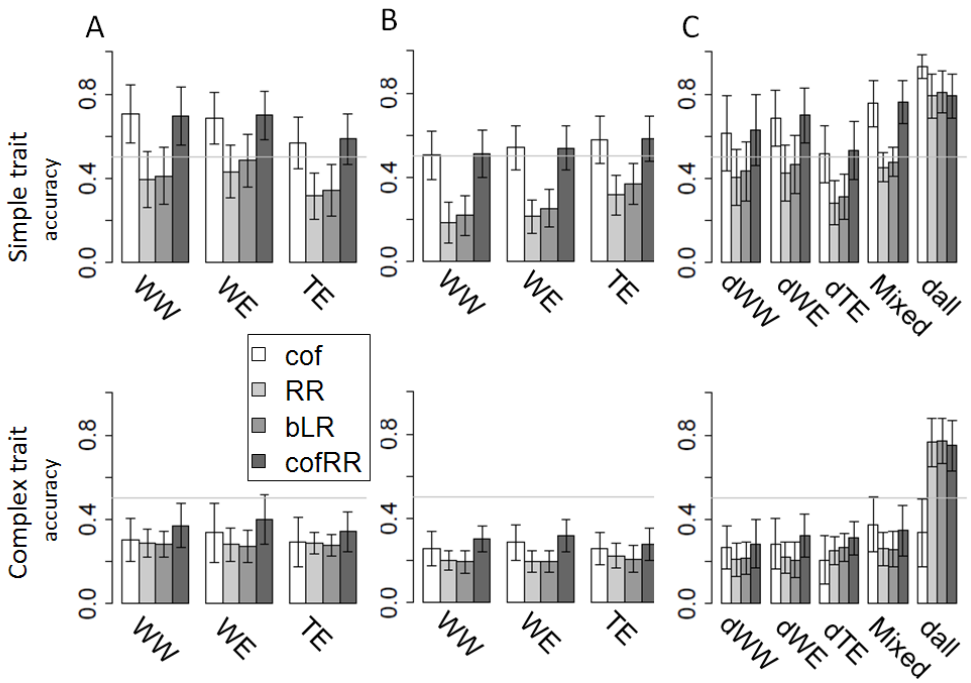


FIGURE 2.5. Mean prediction accuracy as a function of the training – candidate combination. Results are showed on simple and complex traits through the 10 replicates of the simulation. Figure A presents the prediction within sub-population **(**candidate set derived from the training set). Figure B shows the mean accuracy of prediction between sub-population (candidate sub-populations derived from a different training set). Training sets are indicated on the x axis, the four colors representing the four methods used (cof, RR, BLR, cofRR). Training and candidate sets comprised all individuals of the indicated sub-population (1,000 and 200 individuals respectively). In figure C the prediction models were built on the core-collection (Call) and applied to the four breeding sub-populations separately (dWW, dWE, dTE and Mixed, each composed of 200 individuals) and to the whole meta-population (dall, 800 individuals).

We used four methods (cof, RR, BLR, cofRR) to predict descendent populations phenotypes from their genotypes based on prediction models defined on the training populations (Figure 2.S6). We tested different combinations of training versus candidate populations in order to compare their prediction power in different situations of relationship and for different trait complexities and structures (Figure 2.5-2.6).

Model selection:

Auto-prediction (candidate set = training population) with high accuracy proved the relevance of all the models used (Figure 2.S7). Globally, the prediction models showed low (0.2) to high (0.9) accuracy depending on the methods, traits and combination of training and candidate populations. Simple traits were always better predicted than complex ones (accuracy of up to 0.9 versus accuracy of up to 0.5). Models built with cof and cofRR methods always performed better than models built with the other methods for simple traits (mean accuracy on the 10 replicates of 0.2 to 0.85 versus 0.1 to 0.5; Figure 2.S6). For complex traits, cof method was always as efficient as RR and BLR.

Relationship between training and candidate populations:

As expected, accuracies obtained from within sub-population predictions were always better than between sub-population predictions (+0.3 % to 400%; Figure 2.5A and 2.5B). Among within sub-populations predictions, accuracies for simple traits were better with WW and WE as training set than with TE, while no significant difference was observed for complex traits. Using the core-collection as training population, accuracies obtained on dWW, dWE and dTE were as good as for within sub-population prediction (Figure 2.5C). Accuracy was slightly better for the Mixed sub-population than for the others. The best accuracies were obtained predicting the totality of the descendant meta-population (800 individuals, dall). In this case cof method result showed a 15% better accuracy than other methods for simple traits, while it was 56% less accurate for complex traits.

The effect of trait structure:

Structured and non-structured traits were predicted within and between sub-population using cof and cofRR methods (Figure 2.6) and also with the core-collection as training set (Figure 2.7). We observed slightly higher values for non-structured traits than for structured traits, except in the case of WE for simple traits. All markers using models built on the core-collection predicted the structured

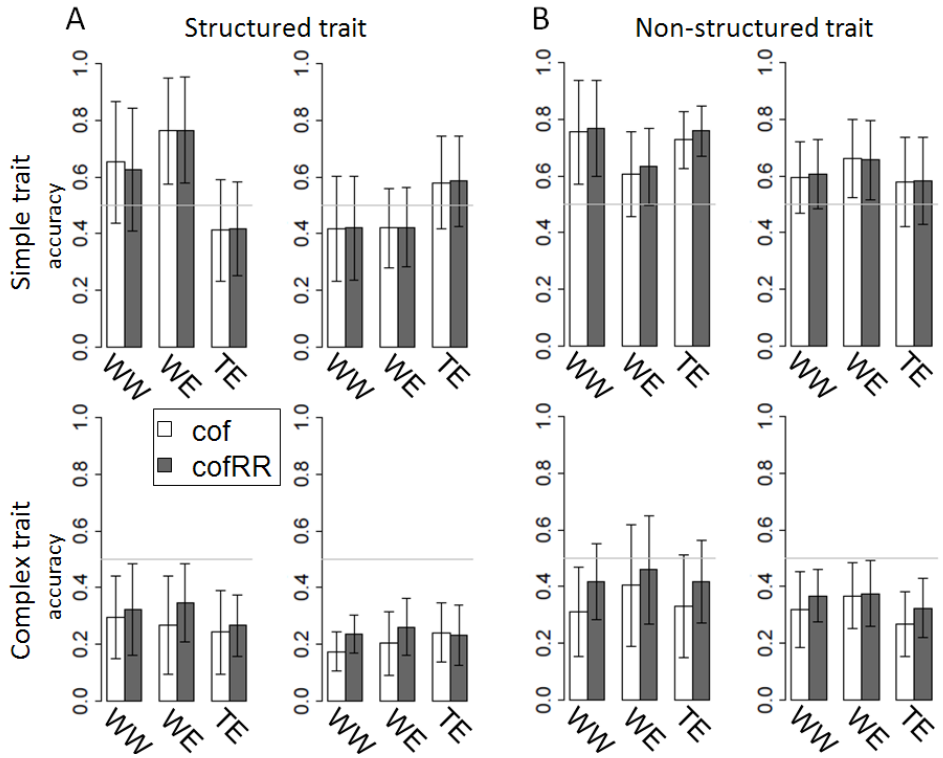


FIGURE 2.6. Mean accuracy of prediction in structured (**A**) and non-structured (**B**) trait. We also compared here two combinations of training – candidate sets (i.e. the two figures on the left present within sub-population predictions and the two figures on the right present between sub-population predictions) and simple and complex traits through 10 replicates of the simulation. Training sets are indicated on the x axis, the two colors are representing the methods used (cof, cofRR). Training and candidate sets comprised all individuals of the sub-population (1,000 and 200 individuals respectively), except for the model constructed on Call, which was tested on the entire breeding population (800 individuals).

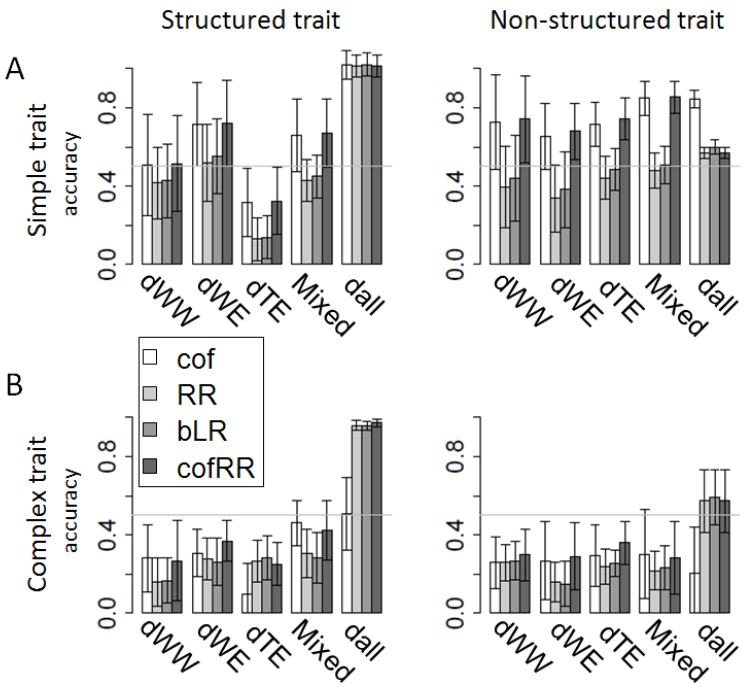


FIGURE 2.7. Prediction accuracy in structured (**A**) and non-structured (**B**) traits using the core-collection as training population. Mean prediction accuracy was calculated on all 10 replicates of the simulation using four methods (cof, RR, BLR, cofRR). Models were built on the core-collection (Call) and applied to the four breeding sub-populations separately (dWW, dWE, dTE and Mixed, each composed of 200 individuals) and on the whole breeding meta-population (dall, 800 individuals). The two figures on the left side represent accuracies observed on structured traits and the other two figures accuracies on non-structured traits.

traits better than the non-structured ones on dWE and on the entire meta-population. In these cases they over performed highly cof method for complex traits (200 to 300%).

Pine data:

After filtering on missing data and allele frequency, approx. 3047 (+/-5) SNPs were considered for the GWAS. One trait only, *fusiform rust susceptibility by presence or absence of rust: Rust_bin* – out of 17 – had cofactors which could always be identified through the 10 training sets of the cross-validation scheme. In this case, higher accuracies were obtained with cofRR method than with either RR or cof. For traits where no cofactors could be identified with mlmm, cof method accuracy was equal to zero, while RR and cofRR methods yielded exactly the same accuracies. Supplementary Figure 2.S8 presents the accuracy of these three methods on two traits having similar Mendelian segregation values (0.26 and 0.21 respectively). The first one is the *average branch diameter of six year-old trees (BD) considered as a complex architecture trait. As no cofactor could be detected for this trait, RR and cofRR yielded the same accuracy (0.50). The second trait is Rust_bin, an oligogenic trait, where one or two cofactors were detected depending on the training set*. Cof method showed poor prediction accuracy (0.24), while cofRR resulted in an accuracy of 0.77, thus outperforming RR method (0.67).

DISCUSSION

Simulated data:

Because high density SNP markers (over 20K) are still unavailable in grape, we have used simulations in order to test both GWAS and GS. Three populations of 1,000 individuals were simulated in order to reflect real data (BACILIERI *et al.* 2013): three genetic pools of high heterozygosity (He = 0.74) but with relatively low differentiation (F_{ST} values of up to 0.07).

The simulation of genomes and causative mechanisms (genetic architecture) in different species is complex. There are many different forms of genomic variability, a wide variety of plausible demographic and evolutionary histories, as well as considerable uncertainty about how mutation and recombination rates vary and about the mode and distribution of gene action (DAETWYLER *et al.* 2012). We chose forward simulation strategy and developed a complex demographic scenario based on historical information, which was implemented using quantiNemo software (NEUENSCHWANDER *et al.* 2008). We simulated natural (Hardy-Weinberg) populations with additional human selection and migration following historical data about grapevine's domestication. Despite the early domestication, human breeding in grape seems to have started late and was not very intensive compared to other crops (maize, rice). Instead of creating advanced lines from complex breeding schemes, a large

genetic diversity was maintained and is still cultivated today (LACOMBE *et al.* 2012). For unknown or hard to estimate parameters (bottleneck, migration rate, selection intensity, variation of parameters in the time, number of generations), we followed guidelines from grapevine's evolution history and defined alternative scenario to test the sensitivity of these parameters. The number of generations since grapevine's domestication was also difficult to estimate because of the combination of vegetative and generative propagation methods over time and across different geographical regions. Several sources suggested a very limited number of generative cycles. For wine cultivars (ARROYO-GARCIA *et al.* 2006) estimated 80 generations, (FOURNIER-LEVEL *et al.* 2010) expected 100. The values we used in our scenarios (505 generations for TE, 100 for WE and 50 for WW) were supported by these historical information, with a constraint to achieve desired population structure (F_{ST} and structure) and to create linkage disequilibrium (LD) between QTLs and surrounding neutral markers.

The simulation of the meta-population based on grape evolution's history led a large set of individuals forming highly polymorphic heterozygous structured populations close to the cultivated compartment of *Vitis vinifera* L. Heterozygosity level was however a little lower than observed, closer to the natural populations of *V. sylvestris*, the wild compartment of grape, which underwent little to no human selection. In this simulated data LD level around the QTLs was slightly higher than in neutral regions of the genome (nine to 16 kb and nine to 13 kb respectively). However, more extended LD can be observed in the region of QTLs controlling binary traits, such as berry color (FOURNIER-LEVEL *et al.* 2009) and muscat flavor (EMANUELLI *et al.* 2010). Indeed, (MYLES *et al.* 2011), using only 5,110 polymorphic SNPs on 289 individuals, were able to identify by GWAS several associations for berry color, which is a highly selected binary trait, indicating an extensive LD between loci located within a 43-kb region (FOURNIER-LEVEL *et al.* 2009). Nevertheless our study focused on quantitative traits, which are nowadays challenging breeding programs, and where genome-wide selection methods are needed.

In the simulations, a large number of parameters were declared (more than 50). These values were defined following the evolutionary history of grape and comparing multiple alternative scenarios. Finally we chose the model which fitted best real data based on four criteria: F_{ST}, LD, heterozygosity and population structure. The scenario we developed is just one possibility to create the target material. This model could be optimized using the Approximate Bayesian Computation (ABC) approach (BEAUMONT *et al.* 2002), but its implementation is very time-consuming and exceeds the scope of this study.

Feasibility of GWAS in grape:

One of the aims of this study was to test GWAS ability to detect simulated QTLs in highly heterozygous genomes in a structured meta-population with high level of genetic diversity, similar to grapevine. Genomes were covered by more than 80,000 well-distributed SNP markers and analyses realized with the mlmm method (SEGURA *et al.* 2012). We simulated four sets of 1,000 individuals (WW, WE, TE, Call) to investigate the genetic properties of four quantitative traits characterized by two levels of complexity (10 or 100 QTLs), linked or not to population structure.

GWAS was more efficient to detect a few QTLs with a large effect (characteristic of simple traits) than to identify multiple loci of too small additive effects, as showed in previous studies (ATWELL *et al.* 2010). In structured and complex traits, a number of underlying QTLs could never be perceived because of fixation. Due to the confounding effect of population structure in structured traits – using a model controlling for population structure – we detected slightly less associations explaining a smaller part of the total variance than in non-structured traits, as already mentioned (ZHAO *et al.* 2007; BUCKLER *et al.* 2009; HOUEL 2011; WANG *et al.* 2012). In this work, we fixed the number of SNPs to 111,000 (of which 92,787 remained polymorphic after running the simulation) so that at least one to two SNPs were present in every LD block of 10 kb. The cases where QTLs could not be detected were due to the small effect (percentage of the variance explained) of these loci (Figure 2.S9). Increasing the sample size of the studied panel can be a solution to detect these QTLs. Indeed, using 1,000 individuals instead of 3,000, only half of the QTLs could be identified in our data (Table 2.S1). Similarly, fewer QTL were identified, especially for the complex traits using the core-collection, meaning that as diversity increases, QTL detection power decreases.

In some cases we observed low LD (r^2<0.01) between a QTL and the significant associations indicated by the best model of mlmm. Some of these markers were found at the same time close to the target QTL and tightly linked to another more significant association. This phenomenon could result from an extremely large QTL effect; as, in addition, the causal loci were not included in the analysis, its variation was thus captured by multiple "complementary" SNPs not completely linked to the QTL. The other part of weakly linked associations was further from the QTL and can be the result of remaining kinship and population structure.

Prediction of phenotypes from genotypes by GEBV:

we will discuss here our GS results focusing on three points: i) the comparison of prediction methods ii) the definition of training and candidate sets in a structured population iii) the influence of trait structure on prediction accuracy.

Several studies identified parameters affecting prediction accuracy. The significance of marker density, size of the training population and trait heritability have already been well assessed (BERNARDO and YU 2007; MUIR 2007; CALUS *et al.* 2008). Therefore, we defined our parameters according to these previous findings, adjusting them to grapevine genome in order to reach optimal prediction accuracy: number of polymorphic SNPs (MAF>0.05 filtered) around 81,000 (one SNP in each 5.8 kb), training population size at 1,000, and heritability between 0.7 and 0.8.

Prediction methods:

We realized genomic predictions on simulated grapevine data using four methods, viz. a classical MAS approach with the cofactors identified in mlmm analysis (cof) and three "all genome" methods: Ridge-Regression BLUP (RR), Bayesian LASSO regression (BLR) and marker assisted Ridge-Regression (cofRR). For the cof and cofRR prediction models, we retained all significant cofactors identified by mlmm, and re-estimated their effects in a mixed model. Our results show that, by considering these effects, higher prediction accuracies can be obtained than by estimating all effects with RR or BLR methods (except for non-structured simple trait predicted with the core-collection on the totality of descendants, where RR, BLR and cofRR were on the same level and cof method outperformed them). The only cofactor-using method (cof) was also as efficient or more than RR and BLR methods in all cases, except for the prediction of the complex trait with the core-collection. A number of authors have shown that there are two major factors affecting prediction accuracy: LD between marker and QTL, and information on the genetic relationship captured by markers (HABIER *et al.* 2007, 2010; GODDARD 2008). The cofRR method uses two types of genomic information: i) the associated cofactors identified by GWA approach (mlmm) that capture the accuracy due to LD between marker and QTL, ii) the remaining markers of the polygenic term that capture the genetic background effect (such as population structure) of the training set. By contrast, cof method is using the first type of information only, while RR and BLR are principally capturing the genetic background effect (HABIER *et al.* 2007). The accuracy due to LD between marker and QTL supersedes the accuracy due to genetic relationship if SNP effect and/or LD are high (GODDARD 2008; ZHONG *et al.* 2009; HABIER *et al.* 2010). Our results on simple and complex traits are in agreement with this, i.e. prediction accuracy of cof method was higher in simple traits than in complex traits, where much fewer QTL could be detected by GWAS (in average 32-59% per replicates for simple trait and 2 to 5 % for complex trait). On the other hand, cof method was as efficient as RR and BLR even in complex traits that can likely be explained by the proportion of causal loci compared to neutral SNPs. The 100 QTLs of the complex traits represent 0.09% of the simulated loci, which is still far from the hypothesis of RR and BLR methods, that all or most of the markers have an effect different from zero. Moreover, (KIZILKAYA *et*

al. 2010) showed that, for a Bayesian prediction model, redundant and uninformative markers diminish prediction accuracy. Finally we can recommend the use of the cofRR method, which was able to predict a large part of the polygenic term, i.e. the variance not-captured by the cofactors, even in complex traits.

Tests on pine data confirmed that cofRR outperformed RR when cofactors could be identified in the training panel. However this advantage strongly relies on the quality and efficiency of GWA analysis with mlmm which provides the cofactors. The present results emphasize the importance of marker density – which is a limiting factor for real data – and information about population structure in the training material.

Combination of training and candidate sets:

We performed genomic predictions using four training sets and four candidate sets issued from crosses between selected training individuals, comparing four methods on four traits (simple/complex and structured/non-structured). Three of the four training sets (WW, WE, TE) comprised all individuals in each sub-population. The fourth training set (Call) was the core-collection defined from the entire meta-population, in order to maximize diversity using 1,000 individuals, including the founders of the four candidate populations. Predictions were developed either using models trained on the population from which the founders were chosen (within sub-population) or from the other populations (between sub-populations), or on a core-collection representing the diversity of the entire meta-population.

According to (DE ROOS *et al.* 2009), lower accuracies were obtained when the training set was not related to the candidate populations (between sub-populations) due to the lower genetic relationship between training and candidate sets. In fact, in our scenario, the three sub-populations diverged from each other due to genetic drift through 500 generations. Differentiation was accelerated by selection and slowed down by migration between sub-populations. However, Figure 2.S9 shows that the effect of QTLs did not vary much between sub-populations, maintaining the accuracy due to LD between marker and QTL. The highest accuracies (up to 0.9) were obtained either in within sub-population predictions or when using the core-collection as training population. Consistent with (HAYES *et al.* 2009b) and (DE ROOS *et al.* 2009), the combination of the individuals of all sub-population in the core-collection yielded as good an accuracy as in within sub-population situations. We have to specify here that the high marker density used in this study allowed capturing the effect of multiple polymorphic QTLs and a great part of the genetic relationship even if sub-populations diverged.

Influence of trait structure:

Our results show that population structure affects prediction accuracy in both simple and complex traits. Globally we observe that non-structured traits were predicted with higher accuracy (Figure 2.6). However, we observe higher accuracy for structured traits than for non-structured ones when predicting the entire breeding meta-population with all-genome using models (RR, BLR, cofRR) built on the core-collection (accuracy of 0.6 and 0.98 respectively; Figure 2.7). Therefore, if there is a significant population structure in the training population and in the candidate set, a trait following this structure is better predicted than a non-structured trait. A plausible explication for these results is that, in contrast to cof method, RR and BLR methods could capture the population structure in the core-collection. This becomes advantageous when the candidate set displays that same population structure (with all groups of structure), and leads to supplementary knowledge in the case of traits which co-segregate with this structure.

In conclusion, we can recommend the use of the cofRR method, which makes simultaneous use of information about QTLs (through cofactors obtained from GWAS), genetic relationship and population structure. Contrary to GWAS, GS using either RR, BLR and cofRR methods is able to take advantage of the population structure when predicting structured traits, if both training and candidate populations are following the same pattern.

This work is the first attempt to test both GWAS and GS in grape through simulations. On a large population of 3,000 individuals, up to 81,555 SNP makers with frequency above 5% and four traits (simple and complex, structured and non-structured) were simulated. Through GWAS, an average of 5.9 to 30% of the QTLs could be identified, the best results being obtained for simple non-structured traits. Genomic estimated breeding values (GEBV) were calculated using the same data set. Predictions for simple traits within population were always more accurate, with a very high accuracy of 0.9, while accuracy dropped to 0.2 for complex trait and between population predictions. Accuracy also depended on the pairs of populations in relation with the mean phenotypic differences between the training and candidate populations. The highest prediction accuracy (up to 0.9) was obtained using the combined GWAS-GS model (cofRR). Finally, for grapevine breeding or for other important economic crops with the same characteristics such as coffee or *Citrus* species, we recommend using the combined prediction model with a core-collection as training population.

ACKNOWLEDGEMENTS

We are grateful to Drs. L. Gay, J. Ronfort and J-.M. Boursiquot for discussion on simulation scenarios, and to Drs. L. Moreau and B. Courtois for discussion about genomic selection. S.N. was supported by

Swiss National Science Foundation grant 31003A_138180 to Dr. J. Goudet. We thank the IT team in the CIRAD cluster for informatics support. This work was funded in part by the French Ministry of Research and Higher Education and the French Ministry of Food, Agriculture and Fisheries (project CAS DAR n°10AAPIT n°1009) and a PhD grant from the French Grapevine and Wine Institute (IFV).

LITERATURE CITED

ARADHYA M. K., DANGL G. S., PRINS B. H., BOURSIQUOT J.-M., WALKER M. A., MEREDITH C. P., SIMON C. J., 2003 Genetic structure and differentiation in cultivated grape, Vitis vinifera L. Genet. Res. **81**: 179–192.

ARROYO-GARCIA R., RUIZ-GARCIA L., BOLLING L., OCETE R., LOPEZ M. A., ARNOLD C., ERGUL A., SÖYLEMEZO˝LU G., UZUN H. I., CABELLO F., IBAÑEZ J., ARADHYA M. K., ATANASSOV A., ATANASSOV I., BALINT S., CENIS J. L., COSTANTINI L., GORISLAVETS S., GRANDO M. S., KLEIN B. Y., MCGOVERN P. E., MERDINOGLU D., PEJIC I., PELSY F., PRIMIKIRIOS N., RISOVANNAYA V., ROUBELAKIS-ANGELAKIS K. A., SNOUSSI H., SOTIRI P., TAMHANKAR S., THIS P., TROSHIN L., MALPICA J. M., LEFORT F., MARTINEZ-ZAPATER J. M., 2006 Multiple origins of cultivated grapevine (Vitis vinifera L. ssp. sativa) based on chloroplast DNA polymorphisms. Mol. Ecol. **15**: 3707–3714.

ATWELL S., HUANG Y. S., VILHJALMSSON B. J., WILLEMS G., HORTON M., LI Y., MENG D., PLATT A., TARONE A. M., HU T. T., JIANG R., MULIYATI N. W., ZHANG X., AMER M. A., BAXTER I., BRACHI B., CHORY J., DEAN C., DEBIEU M., MEAUX J. DE, ECKER J. R., FAURE N., KNISKERN J. M., JONES J. D. G., MICHAEL T., NEMRI A., ROUX F., SALT D. E., TANG C., TODESCO M., TRAW M. B., WEIGEL D., MARJORAM P., BOREVITZ J. O., BERGELSON J., NORDBORG M., 2010 Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature **465**: 627–631.

BACILIERI R., LACOMBE T., CUNFF L. LE, VECCHI-STARAZ M. D., LAUCOU V., GENNA B., PEROS J.-P., THIS P., BOURSIQUOT J.-M., 2013 Genetic structure in cultivated grapevines is linked to geography and human selection. BMC Plant Biol. **13**: 25.

BEAUMONT M. A., ZHANG W., BALDING D. J., 2002 Approximate Bayesian Computation in Population Genetics. Genetics **162**: 2025–2035.

BERNARDO R., YU J., 2007 Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. **47**: 1082–1090.

BOURSIQUOT J.-M., LACOMBE T., LAUCOU V., JULLIARD S., PERRIN F.-X., LANIER N., LEGRAND D., MEREDITH C., THIS P., 2009 Parentage of Merlot and related winegrape cultivars of southwestern France: discovery of the missing link. Aust. J. Grape Wine Res. **15**: 144–155.

BOWERS, BOURSIQUOT, THIS, CHU, JOHANSSON, MEREDITH, 1999 Historical Genetics: The Parentage of Chardonnay, Gamay, and Other Wine Grapes of Northeastern France. Science **285**: 1562–1565.

BUCKLER E. S., HOLLAND J. B., BRADBURY P. J., ACHARYA C. B., BROWN P. J., BROWNE C., ERSOZ E., FLINT-GARCIA S., GARCIA A., GLAUBITZ J. C., GOODMAN M. M., HARJES C., GUILL K., KROON D. E., LARSSON S., LEPAK N. K., LI H., MITCHELL S. E., PRESSOIR G., PEIFFER J. A., ROSAS M. O., ROCHEFORD T. R., ROMAY M. C., ROMERO S., SALVO S., VILLEDA H. S., SILVA H. S. da, SUN Q., TIAN F., UPADYAYULA N., WARE D., YATES H., YU J., ZHANG Z., KRESOVICH S., MCMULLEN M. D., 2009 The Genetic Architecture of Maize Flowering Time. Science **325**: 714–718.

CALUS M. P. L., MEUWISSEN T. H. E., ROOS A. P. W. DE, VEERKAMP R. F., 2008 Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics **178**: 553–561.

CAMPOS G. DE LOS, HICKEY J. M., PONG-WONG R., DAETWYLER H. D., CALUS M. P. L., 2012 Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics **193**: 327–345.

CARDON L. R., PALMER L. J., 2003 Population stratification and spurious allelic association. Lancet **361**: 598–604.

CARRIER G., CUNFF L. LE, DEREEPER A., LEGRAND D., SABOT F., BOUCHEZ O., AUDEGUIN L., BOURSIQUOT J.-M., THIS P., 2012 Transposable elements are a major cause of somatic polymorphism in Vitis vinifera L. PloS One **7**: e32973.

CHEN J., CHEN Z., 2008 Extended Bayesian information criteria for model selection with large model spaces. Biometrika **95**: 759–771.

DAETWYLER H. D., CALUS M. P. L., PONG-WONG R., CAMPOS G. DE LOS, HICKEY J. M., 2012 Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics **193**: 347–365.

DOLIGEZ A., ADAM-BLONDON A. F., CIPRIANI G., GASPERO G. DI, LAUCOU V., MERDINOGLU D., MEREDITH C. P., RIAZ S., ROUX C., THIS P., 2006 An integrated SSR map of grapevine based on five mapping populations. TAG Theor. Appl. Genet. Theor. Angew. Genet. **113**: 369–382.

EARL D. A., VONHOLDT B. M., 2011 STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. **4**: 359–361.

EDING H., MEUWISSEN T. H. E., 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J. Anim. Breed. Genet. **118**: 141–159.

EMANUELLI F., BATTILANA J., COSTANTINI L., CUNFF L. L., BOURSIQUOT J.-M., THIS P., GRANDO M. S., 2010 A candidate gene association study on muscat flavor in grapevine (Vitis vinifera L.). BMC Plant Biol. **10**: 241.

EMANUELLI F., LORENZI S., GRZESKOWIAK L., CATALANO V., STEFANINI M., TROGGIO M., MYLES S., MARTINEZ-ZAPATER J. M., ZYPRIAN E., MOREIRA F. M., GRANDO M. S., 2013 Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biol. **13**: 39.

FOURNIER-LEVEL A., CUNFF L. L., GOMEZ C., DOLIGEZ A., AGEORGES A., ROUX C., BERTRAND Y., SOUQUET J.-M., CHEYNIER V., THIS P., 2009 Quantitative Genetic Bases of Anthocyanin Variation in Grape (Vitis vinifera L. ssp. sativa) Berry: A Quantitative Trait Locus to Quantitative Trait Nucleotide Integrated Study. Genetics **183**: 1127–1139.

FOURNIER-LEVEL A., LACOMBE T., CUNFF L. LE, BOURSIQUOT J.-M., THIS P., 2010 Evolution of the VvMybA gene family, the major determinant of berry colour in cultivated grapevine (Vitis vinifera L.). Heredity **104**: 351–362.

GODDARD M., 2008 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136**: 245–257.

GODDARD M. e., HAYES B. j., 2007 Genomic selection. J. Anim. Breed. Genet. **124**: 323–330.

GOUESNARD B., BATAILLON T. M., DECOUX G., ROZALE C., SCHOEN D. J., DAVID J. L., 2001 MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J. Hered. **92**: 93–94.

GRASSI F., LABRA M., IMAZIO S., SPADA A., SGORBATI S., SCIENZA A., SALA F., 2003 Evidence of a secondary grapevine domestication centre detected by SSR analysis. Theor. Appl. Genet. **107**: 1315–1320.

GRATTAPAGLIA D., RESENDE M. D. V., 2010 Genomic selection in forest tree breeding. Tree Genet. Genomes **7**: 241–255.

HABIER D., FERNANDO R. L., DEKKERS J. C. M., 2007 The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics **177**: 2389–2397.

HABIER D., FERNANDO R. L., DEKKERS J. C. M., 2008 The impact of genetic relationship information on genome-assisted breeding values. Genetics.

HABIER D., TETENS J., SEEFRIED F.-R., LICHTNER P., THALLER G., 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. **42**: 5.

HAMBLIN M. T., BUCKLER E. S., JANNINK J.-L., 2011 Population genetics of genomics-based crop improvement methods. Trends Genet. **27**: 98–106.

HANNAH L., ROEHRDANZ P. R., IKEGAMI M., SHEPARD A. V., SHAW M. R., TABOR G., ZHI L., MARQUET P. A., HIJMANS R. J., 2013 Climate change, wine, and conservation. Proc. Natl. Acad. Sci. **110**: 6907–6912.

HAYES B. J., BOWMAN P. J., CHAMBERLAIN A. J., GODDARD M. E., 2009a Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. **92**: 433–443.

HAYES B. J., BOWMAN P. J., CHAMBERLAIN A. C., VERBYLA K., GODDARD M. E., 2009b Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. **41**: 51.

HEFFNER E. L., LORENZ A. J., JANNINK J.-L., SORRELLS M. E., 2010 Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Sci. **50**: 1681–1690.

HILL W. G., WEIR B. S., 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. **33**: 54–78.

HOERL A. E., KENNARD R. W., 1970 Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics **12**: 55–67.

HOUEL C., 2011 Caractérisation de la variation phénotypique de la taille de la baie chez la vigne Vitis vinifera L. et approches de génétique d'association et de recherche de traces de sélection pour ce caractère.

HUANG X., WEI X., SANG T., ZHAO Q., FENG Q., ZHAO Y., LI C., ZHU C., LU T., ZHANG Z., LI M., FAN D., GUO Y., WANG A., WANG L., DENG L., LI W., LU Y., WENG Q., LIU K., HUANG T., ZHOU T., JING Y., LI W., LIN

Z., BUCKLER E. S., QIAN Q., ZHANG Q.-F., LI J., HAN B., 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. **42**: 961–967.

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.

JANNINK J.-L., LORENZ A. J., IWATA H., 2010 Genomic selection in plant breeding: from theory to practice. Briefings Funct. Genomics **9**: 166–177.

KANG H. M., ZAITLEN N. A., WADE C. M., KIRBY A., HECKERMAN D., DALY M. J., ESKIN E., 2008 Efficient Control of Population Structure in Model Organism Association Mapping. Genetics **178**: 1709–1723.

KIZILKAYA K., FERNANDO R. L., GARRICK D. J., 2010 Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. **88**: 544–551.

KUMAR S., SKJÆVELAND Å., ORR R. J., ENGER P., RUDEN T., MEVIK B.-H., BURKI F., BOTNEN A., SHALCHIAN-TABRIZI K., 2009 AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. BMC Bioinformatics **10**: 357.

LACOMBE T., 2012 Contribution à l'étude de l'histoire évolutive de la vigne cultivée (Vitis vinifera L.) par l'analyse de la diversité génétique neutre et de gènes d'intérêt.

LACOMBE T., BOURSIQUOT J.-M., LAUCOU V., VECCHI-STARAZ M. DI, PEROS J.-P., THIS P., 2012 Large-scale parentage analysis in an extended set of grapevine cultivars (<i>Vitis vinifera</i> L.). Theor. Appl. Genet.: 1–14.

LAUCOU V., LACOMBE T., DECHESNE F., SIRET R., BRUNO J.-P., DESSUP M., DESSUP T., ORTIGOSA P., PARRA P., ROUX C., SANTONI S., VARES D., PEROS J.-P., BOURSIQUOT J.-M., THIS P., 2011 High throughput

analysis of grape genetic diversity as a tool for germplasm collection management. TAG Theor. Appl. Genet. Theor. Angew. Genet. **122**: 1233–1245.

LEVADOUX L., 1956 Les populations sauvages et cultivées de Vitis vinifera L. Ann. Amélioration Plantes **1**: 59–118.

LIJAVETZKY D., CABEZAS J., IBAÑEZ A., RODRIGUEZ V., MARTINEZ-ZAPATER J. M., 2007 High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics **8**: 424.

MANGIN B., SIBERCHICOT A., NICOLAS S., DOLIGEZ A., THIS P., CIERCO-AYROLLES C., 2011 Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity **108**: 285–91.

MARCHINI J., CARDON L. R., PHILLIPS M. S., DONNELLY P., 2004 The effects of human population structure on large genetic association studies. Nat. Genet. **36**: 512–517.

MEUWISSEN T. H. E., HAYES B. J., GODDARD M. E., 2001 Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics **157**: 1819–1829.

MITA S. DE, THUILLET A.-C., GAY L., AHMADI N., MANEL S., RONFORT J., VIGOUROUX Y., 2013 Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol. Ecol. **22**: 1383–1399.

MORIONDO M., JONES G. V., BOIS B., DIBARI C., FERRISE R., TROMBI G., BINDI M., 2013 Projected shifts of wine regions in response to climate change. Clim. Change **119**: 825–839.

MOSER G., TIER B., CRUMP R. E., KHATKAR M. S., RAADSMA H. W., 2009 A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. **41**: 56.

MUIR W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. Z. Für Tierzüchtung Züchtungsbiologie **124**: 342–355.

MYLES S., BOYKO A. R., OWENS C. L., BROWN P. J., GRASSI F., ARADHYA M. K., PRINS B., REYNOLDS A., CHIA J.-M., WARE D., BUSTAMANTE C. D., BUCKLER E. S., 2011 Genetic Structure and Domestication History of the Grape. Proc. Natl. Acad. Sci. **108**: 3530–3535.

MYLES S., CHIA J.-M., HURWITZ B., SIMON C., ZHONG G. Y., BUCKLER E., WARE D., 2010 Rapid Genomic Characterization of the Genus Vitis. PLoS ONE **5**: e8219.

NAKAYA A., ISOBE S. N., 2012 Will genomic selection be a practical method for plant breeding? Ann. Bot. **110**: 1303–1316.

NEUENSCHWANDER S., HOSPITAL F., GUILLAUME F., GOUDET J., 2008 quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. Bioinformatics **24**: 1552–1553.

OLLAT N., FERNANDEZ L., ROMIEU C., DUCHENE E., LISSARAGUE J. R., LECOURIEUX D., AGEORGES A., KELLY M., CACHO J., RIVARS J., LAMUELA R., GOUTOULY J. P., LEEUWEN C. VAN, MARGUERIT E., PECCOUX A., BARRIEU F., LEBON E., THIS P., PELLEGRINO A., MARTINEZ-ZAPATER J. M., TORREGROSA L., 2011 Multidisciplinary research to select new cultivars adapted to climate changes. In: Asti and Alba, Italy.

PARK S. D. E., 2001 Trypanotolerance in West African Cattle and the Population Genetic Effects of Selection.

PARK T., CASELLA G., 2008 The Bayesian Lasso. J. Am. Stat. Assoc. **103**: 681–686.

PASLIER M.-C. LE, CHOISNE R., BACILIERI R., BOURSIQUOT J.-M., BRAS M., BRUNEL D., GASPERO G. DI, HAUSMANN L., LACOMBE T., LAUCOU V., LAUNAY A., MARTINEZ-ZAPATER J., MORGANTE M., RAJ P., PONNAIAH M., QUESNEVILLE H., SCALABRIN S., TORRES-PEREZ R., ADAM-BLONDON A.-F., 2013 The GrapeReSeq 18k Vitis genotyping chip. In: La Serena, Chile.

PEREZ P., CAMPOS G. DE LOS, CROSSA J., GIANOLA D., 2010 Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. Plant Genome J. **3**: 106–116.

PRITCHARD J. K., STEPHENS M., DONNELLY P., 2000 Inference of Population Structure Using Multilocus Genotype Data. Genetics **155**: 945–959.

R CORE TEAM, 2013 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

RESENDE M. F. R., MUNOZ P., RESENDE M. D. V., GARRICK D. J., FERNANDO R. L., DAVIS J. M., JOKELA E. J., MARTIN T. A., PETER G. F., KIRST M., 2012 Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics **190**: 1503–1510.

ROOS A. P. W. DE, HAYES B. J., GODDARD M. E., 2009 Reliability of Genomic Predictions Across Multiple Populations. Genetics **183**: 1545–1553.

SCHOEN D. J., BROWN A. H., 1993 Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. U. S. A. **90**: 10623–10627.

SEGURA V., VILHJALMSSON B. J., PLATT A., KORTE A., SEREN Ü., LONG Q., NORDBORG M., 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. **44**: 825–830.

SPITZE K., 1993 Population structure in Daphnia obtusa: quantitative genetic and allozymic variation. Genetics **135**: 367–374.

THIS P., LACOMBE T., THOMAS M. R., 2006 Historical origins and genetic diversity of wine grapes. Trends Genet. **22**: 511–519.

TIAN F., BRADBURY P. J., BROWN P. J., HUNG H., SUN Q., FLINT-GARCIA S., ROCHEFORD T. R., MCMULLEN M. D., HOLLAND J. B., BUCKLER E. S., 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. **43**: 159–162.

VIGOUROUX Y., JAQUETH J. S., MATSUOKA Y., SMITH O. S., BEAVIS W. D., SMITH J. S. C., DOEBLEY J., 2002 Rate and Pattern of Mutation at Microsatellite Loci in Maize. Mol. Biol. Evol. **19**: 1251–1260.

WANG M., JIANG N., JIA T., LEACH L., COCKRAM J., WAUGH R., RAMSAY L., THOMAS B., LUO Z., 2012 Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theor. Appl. Genet. **124**: 233–246.

WEIR B. S., COCKERHAM C. C., 1984 Estimating F-Statistics for the Analysis of Population Structure. Evolution **38**: 1358–1370.

WONG C. K., BERNARDO R., 2008 Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. **116**: 815–824.

ZHAO K., ARANZANA M. J., KIM S., LISTER C., SHINDO C., TANG C., TOOMAJIAN C., ZHENG H., DEAN C., MARJORAM P., NORDBORG M., 2007 An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet **3**: e4.

ZHONG S., DEKKERS J. C. M., FERNANDO R. L., JANNINK J.-L., 2009 Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics **182**: 355–364.

ZOHARY D., 1996 The domestication of the grapevine Vitis vinifera L. in the Near East. In: *The origins and ancient history of wine.*, McGovern PE, Fleming SJ, Katz SH, pp. 31–43.

3. Conclusion

Les résultats de l'article de simulation ont montré qu'avec suffisamment de marqueurs moléculaires (jusqu'à 90 000 SNPs), et avec des populations d'entrainement de taille importante (1000 individus par sous-population, 3000 individus en tout) et avec des populations candidates de type bi-parentales de 200 individus, il était possible de prédire les phénotypes d'une population candidate avec une bonne précision (jusqu'à 0,9 pour les caractères non structurés et simples, jusqu'à 0,5 pour des caractères complexes structurés) à partir des données génotypiques.

Comme pour la génétique d'association « genome-wide », la sélection génomique est affectée par la structure, et est plus efficace pour des caractères simples, alors que c'est pour des caractères complexes qu'elle serait la plus utile. Elle est cependant encore assez intéressante pour permettre – dans les conditions citées plus haut – une sélection d'individus dans des populations bi-parentales avec une précision suffisante, pour effectuer une première sélection des individus les plus prometteurs, qui seront ensuite plantés au vignoble pour validation.

Dans les conditions réelles, cependant, des effectifs aussi importants sont rarement atteints, pas plus que des couvertures en marqueurs aussi denses. Cette étude, représente donc certainement des conditions très favorables. Pour estimer les meilleures conditions possibles, nous aurions pu augmenter le nombre de marqueurs d'un facteur 10, puisque le reséquençage complet de centaines de variétés de vigne produirait sans doute plusieurs millions de SNPs, les 2 haplotypes de Pinot N, diffèrent déjà de quelques 2 millions de SNPs (VELASCO *et al.* 2007). Les capacités informatiques dont nous disposions, ne nous ont cependant pas permis de tester un tel nombre de marqueurs.

Les moyens actuels en termes de marqueurs moléculaires sont cependant plus modestes, le reséquençage massif d'accessions de vigne étant encore trop cher. Nous avons donc choisi dans un second temps, de tester des conditions plus limitées, obtenues sur du matériel végétal disponible dans l'UMT Géno-Vigne®.

CHAPITRE 3 : PREDICTION DES PHENOTYPES A PARTIR D'UN LARGE ECHANTILLON DE DIVERSITE

1. Introduction

Les analyses sur les données réelles ont été réalisées dans le cadre de 2 expérimentations se différenciant par la nature de la population d'entrainement. Dans le présent chapitre, afin de rester au plus près des conditions de l'étude de simulation, nous avons choisi de travailler avec un panel de diversité, représentatif de la diversité connue dans l'espèce *Vitis vinifera* L.

Ainsi, dans le cadre d'un projet ANR, le projet DL-Vitis, un échantillon diversifié représentatif de la structuration de la collection avait été défini. Cet échantillon de 279 variétés est composé de 3 sous-populations de 93 individus représentant les 3 pools génétiques : le pool des variétés de cuve de l'ouest (WW), le pool des variétés de cuve de l'est (WE) et le pool des variétés de table de l'est (TE).

Cet échantillon a été défini à partir des 2486 cultivars uniques de la collection de Vassal. Ceux-ci ont été analysés à l'aide de 20 marqueurs microsatellites répartis sur le génome de la vigne (LAUCOU *et al.* 2011). Les données moléculaires ont ensuite été analysées à l'aide du logiciel STRUCTURE (PRITCHARD *et al.* 2000) afin d'identifier les populations ancestrales. Nous avons retenu K=3, correspondant à l'analyse de cette population (BACILIERI *et al.* 2013). Seuls les individus appartenant à au moins 80% à l'une des trois populations ont été retenus. Parmi ces variétés, nous avons identifié les géniteurs les plus utilisés ou les géniteurs ancestraux, en utilisant les données d'apparentement (LACOMBE *et al.* 2012) ainsi que des données historiques. Les échantillons ont ensuite été complétés à 93, à l'aide de la procédure « max length subtree procedure » de DARwin (PERRIER and JACQUEMOUD-COLLET 2006), tout en retirant les individus directement apparentés par une recherche de paternité à l'aide du logiciel Famoz (GERBER *et al.* 2003).

Cet échantillon, a par ailleurs été planté au vignoble, en 2009, par surgreffage sur une parcelle de Marselan disponible au domaine Montpellier Supagro du Chapitre, directement adjacente à une parcelle expérimentale correspondant à une population issue d'un croisement entre les variétés Syrah et Grenache, dont 21 de ces descendants ont constitué la population candidate.

Ces 2 populations ont été phénotypées pour un certain nombre de caractères d'intérêt agronomique. Compte tenu des résultats de l'étude de simulation, nous avons décidé pour la présente étude d'analyser des caractères *a priori* liés à la structure de la population, que sont la taille de la baie, le poids et la longueur des grappes. Ce sont en effet des caractères différenciant très nettement les populations de raisins de table[3] de celle des raisins de cuve[4]. En plus de ces 3

[3] Variétés consommée en frais
[4] Variétés utilisées pour la production du vin

caractères, nous avons également étudié le poids de bois de taille[5], un caractère fortement associé à la vigueur des variétés.

Ces 4 caractères sont importants pour les viticulteurs et sont donc importants pour la création des nouvelles variétés.

[5] Il s'agit de l'ensemble des bois récoltés sur une variété lors de la taille d'hiver et pesés au champ.

2. Prédiction des phénotypes à partir d'un large échantillon de diveristé

Second article de la thèse, à soumettre pour publication dans la revue Theoretical and Applied Genetics.

Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) in grape for phenotype prediction using a large diversity panel

Agota Fodor[1,2], Jean-Pierre Péros[2], Amandine Launay[2], Agnès Doligez[2], Gilles Berger[2], Yves Bertrand[2], Maryline Roques[1], Isabelle Beccavin[1], Marie-Christine Le Paslier[3], Charles Romieu[2], Roberto Bacilieri[2], Valérie Laucou[2], Anne-Françoise Adam-Blondon[4], Jean-Michel Boursiquot[2], Patrice This[2], Loïc Le Cunff[1]

1. UMT Geno-Vigne®, IFV-INRA-Montpellier Supagro, Montpellier, 34060 France
2. UMR AGAP, INRA, Montpellier, 34060 France
3. UR EPGV, INRA, 2 rue Gaston Crémieux CP 5708, 91057 EVRY CEDEX, France
4. URGI, INRA, route de Saint-Cyr, RD 10, 78026 VERSAILLES CEDEX Versailles, France

Corresponding author: Loïc Le Cunff

Address: INRA, UMT Geno-Vigne, 2 place P. Viala, 34060 Montpellier, France

Tel: +33 4 99 61 30 97

E-mail: loic.lecunff@supagro.inra.fr

ABSTRACT

The availability of high throughput genotyping tools in grape, allows the development of GWAS and GS approaches in breeding programs. GS allows the estimation of genomic estimated breeding values (GEBV) of breeding candidates based on large sets of markers while GWAS using mlmm R-package, allows the selection of the more pertinent associations (selection of cofactors) to obtain a model explaining the maximum of phenotypic variability.

We estimated the efficiency of GWAS and GS, separately or combined, for the marker assisted selection in grape breeding programs using a large diversity panel on four different traits which covariate with the population structure (berry weight, pruning weight, cluster weight and cluster length). Genotypic informations were obtained with a 18.000 Illumina SNP genotyping chip (http://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K) which is currently the largest SNP genotyping tool available for *Vitis vinifera* L. Genotyping was performed on two different samples: Diversity panel (DP) of 279 individuals divided into three sub-populations used as training population for the phenotype prediction. 21 individuals of a Syrah X Grenache progeny were used as validation population for phenotype prediction.

GWAS was realized on each sub-population separately and on the entire DP. Associations were only identified in the sub-populations. Best mlmm models, contained cofactors only for berry weight and cluster length. We detected 0 to 9 cofactors according to the tested panel for a total of 24 significant SNPs. The percentage of variance explained varied between 2% and 28% per marker and from 27.5 to 91.7% per model.

Four different phenotype prediction models were performed based on i) effects of cofactors from mlmm analysis ii) Ridge regression iii) Bayesian LASSO and iv) combined model using cofactors effect from mlmm and Ridge regression. These phenotype prediction methods gave r values (correlation between GEBVs and phenotypes) ranged from -0.38 to +0.40. For each studied trait the best phenotype prediction gave r values superior to 0.3. In conclusion, we have thus demonstrated the capacity to predict with a good precision the phenotype of breeding candidates even if no association was found by GWAS.

INTRODUCTION

Genome-wide association studies (GWASs) and genomic selection (GS) are nowadays the more interesting approaches for marker assisted selection (MAS). GWAS methodology searches for molecular polymorphisms linked to the variation of selected traits. It is more frequently implemented, in a panel representing a large diversity of the species where new functional variations can be identified (HAMBLIN *et al.* 2011). GWAS was widely used in plant genetics and allowed to identify many common alleles of major effect (ATWELL *et al.* 2010; HUANG *et al.* 2010; TIAN *et al.* 2011). However it was less efficient for traits correlated with environmental gradients or human selection that generate a structure in the population (population structure), which introduce confounding effects leading to false positives at the association tests (CARDON and PALMER 2003; MARCHINI *et al.* 2004). Moreover, the efficiency of GWAS is also impacted by the genetic architecture of the studied trait: indeed, the detection of linked molecular markers in polygenic traits strongly depends both on the size of the sample and on the density of molecular marker (ZHAO *et al.* 2007; BUCKLER *et al.* 2009; WANG *et al.* 2012). Recently an advanced GWAS approach by multi-locus mixed-model – mlmm – was implemented by (SEGURA *et al.* 2012), which increase the power of GWAS and diminish false discovery rate by taking pertinent cofactors and family relationships into account.

The aim of GS is to predict breeding values – genomic estimated breeding values (GEBVs) – for the tested genotypes (MEUWISSEN *et al.* 2001). The statistical models developed for GS methodology are using all markers simultaneously. This approach allows capturing not only LD between markers and causal loci but also information about family relationships and population structure (HABIER *et al.* 2007, 2010). In the practice, GS is realized using two sets of individuals: a 'training set' (or estimation set) where the parameters of the prediction model are estimated and a set of breeding candidates ('candidate set') with genotypic data only (HEFFNER *et al.* 2009). However cross-validation tests are also widely used to investigate the feasibility of GS on a particular data set before setting up a breeding program (RESENDE *et al.* 2012a; KUMAR *et al.* 2012; STORLIE and CHARMET 2013; GOUY *et al.* 2013). Previous studies on animal and plant models, based on both simulated and real data demonstrated the interest of GS, especially for capturing small-effect quantitative trait loci (HAYES *et al.* 2009a; HEFFNER *et al.* 2010; JANNINK *et al.* 2010; GRATTAPAGLIA and RESENDE 2010; HAMBLIN *et al.* 2011; NAKAYA and ISOBE 2012). In recent year breeding schemes using GS were developed for several species (HEFFNER *et al.* 2010; IWATA and JANNINK 2011; SITZENSTOCK *et al.* 2013; HAYES *et al.* 2013) and are already used for dairy cattle (BOUQUET and JUGA 2013).

In grapevine, until now region specific association genetic studies were successful to identify candidate genes linked to berry color (FOURNIER-LEVEL *et al.* 2009) muscat flavor (EMANUELLI *et al.*

2013) and for pro-anthocyanidin composition (HUANG *et al.* 2012; CARRIER *et al.* 2013). The first study using genome-wide markers was realized by (MYLES *et al.* 2011), with 5,110 SNPs on 289 individuals. Since this marker density was weak compared to the LD in grapevine (10 kb at r^2=0.2 ; (MYLES *et al.* 2010), the analysis was only able to identify several associations for berry color, which is a highly selected binary trait, leading to an extensive LD between loci located within a 43-kb region (FOURNIER-LEVEL *et al.* 2009, 2010). These results suggest that much higher marker density will be needed to perform efficient GWAS on complex traits in grapevine. FODOR *et al*. (submitted) tested the feasibility of GWAS and GS in a large diversity and highly heterozygous panel via simulation study. A structured population with 3,000 individuals was generated with approx 90,000 polymorphic genome-wide SNP markers through a scenario following grapevine evolution history. These parameters enabled GWAS to detect approx. 50% and 12% of underlying QTLs for traits controlled by 10 and 100 QTLs respectively. Today, a 18K genotyping chip is available (LE PASLIER *et al.* 2013) but will only increase the number of markers available for *Vitis vinifera* L. up to 20K. We believe that difficulties linked to marker availability and costs of sequencing are progressively disappearing with the development of 'genotyping by sequencing' (GBS) technologies (ELSHIRE *et al.* 2011; POLAND and RIFE 2012).

In grapevine, no advanced breeding lines from complex schemes are available. Instead, breeders are handling a large parental panel with a high diversity both at morphological and molecular level. This material is highly heterozygous (He = 0.76(LAUCOU *et al.* 2011), characterized by a low level of linkage disequilibrium (LD) between marker loci (r^2 ~ 0.2 at 5-10 Kb, (LIJAVETZKY *et al.* 2007; MYLES *et al.* 2010). Most of the cultivars are interconnected by a series of first-degree relationships (for example, Pinot noir – Chardonnay – Gouais blanc, Cabernet franc – Merlot ; (BOWERS *et al.* 1999; BOURSIQUOT *et al.* 2009), but the number of connected generations is rather low (MYLES *et al.* 2011; LACOMBE *et al.* 2012) and the pedigree data is not available. Furthermore some major agricultural traits (for example berry size, bunch size) are linked to population structure, making association studies difficult (HOUEL 2011).

The more important traits in grapevine improvement are quality traits and resistance/tolerance traits for biotic and abiotic stresses (OLLAT *et al.* 2011; MORIONDO *et al.* 2013; HANNAH *et al.* 2013). Today, breeders are first selecting parents from the diversity and realize bi-parental crosses to create recombinant individuals. Then progenies containing interesting characteristics from both parents are selected, and eventually crossed with another cultivar or individual from another cross (multiple back-crosses are not recommended in grapevine because of the high level of inbreeding depression). MAS was already used to follow genes linked to resistances again pathogens (ADAM-BLONDON *et al.* 2001; PAUQUET *et al.* 2001; FISCHER *et al.* 2004; COLEMAN *et al.* 2009; MARGUERIT *et al.* 2009; DI GASPERO *et al.* 2012; SCHWANDER *et al.* 2012; RIAZ *et al.* 2013) or other mono- or oligogenic traits such as

muscat flavor (EMANUELLI *et al.* 2010), berry color (FOURNIER-LEVEL *et al.* 2009), seedless berry (ADAM-BLONDON *et al.* 2001; MEJÍA *et al.* 2011). But its implementation for complex traits need powerful tools to identify polymorphisms linked to the causal variation – GWAS – or a genomic selection approach.

Considering the breeding schemes developed for grape, the ideal way to implement GS would be to be able to estimate the breeding values of the progenies of several bi-parental crosses from prediction models developed on a single training set, representing the diversity of the potential progenitors. In a previous paper (FODOR *et al*. submitted), we have implemented the concept and methodology and demonstrated the potential of GWAS and GS for grape using abundant genotypic data from simulation.

In the present work, using newly developed tools i.e. a diversity panel (HOUEL *et al.* 2013); NICOLAS *et al.* in prep.) and the 18K genotyping chip (LE PASLIER *et al.* 2013), we evaluated the performance of GWAS on four structured traits (berry weight, cluster length and weight, pruning weight). We also investigated the feasibility and the efficiency of genomic prediction in a bi-parental cross using the diversity panel as training set. With the purpose to obtain the highest prediction accuracy possible, we tested four prediction methods: the sum of effects of markers detected in GWAS (cof) – using mlmm (SEGURA *et al.* 2012) – corresponding to classical MAS, Ridge Regression BLUP (RR) (HOERL and KENNARD 1970), Bayesian LASSO (Least Absolute Shrinkage and Selection Operator) Regression (BLR) (PÉREZ *et al.* 2010) and a combination of cof and RR-BLUP (cofRR; FODOR *et al*. submitted). In order to take into account the architecture of the traits, we also used different coding exploring additive and dominant effects of underlying QTLs.

MATERIALS AND METHODS

Plant material:

In this study we handled two samples of plant material. The first was a panel of 279 individuals (diversity panel: DP; Table 3.S1) that can be decomposed into three sub-populations corresponding to wine cultivars from Western Europe (WW), wine cultivars from Eastern Europe (WE) and table cultivars from Eastern Europe (TE) representing original grape genetic pools (BACILIERI *et al.* 2013). Each sub-population was represented through 93 individuals belonging in more than 80% to one of the three groups. Within each population individuals were selected in order be the more representative possible while reducing to a minimum any parental relationship (HOUEL *et al.* 2013); NICOLAS *et al.* in prep.).

The second sample was composed of *cv*. Syrah and *cv*. Grenache and 21 individuals of a progeny from a cross between them (SxG).

Field experiment:

Each plant of DP was over-grafted in 2009 on the cultivar Marselan in five randomized complete blocks with one repetition per block. SxG individuals were sampled from another trial described in (FOURNIER-LEVEL *et al.* 2009). Briefly, genotypes from a Syrah × Grenache cross were planted in 2003; each individual from the progeny and the parents were planted in one elementary plot comprising five plants. Both trials were localized at the INRA Chapitre experimental station (Hérault, France), maintained under classical local training system (3300 plants/ha plant density) and they were both irrigated during summer to avoid drought stress.

Phenotyping of the studied traits:

In this study, four traits were phenotyped: berry weight, cluster weight and length – in three years in the DP (2010, 2011, 2012) and one in SxG (2012)– and pruning weight – in two years in the DP (2010 and 2011) and one in SxG -2011). For each genotype, mean cluster weight and length were recorded on three clusters (representative of the genotype) harvested at maturity (20 Brix degrees). The same clusters were used to measure berry weight, i.e. the end parts of the clusters were discarded and 100 berries randomly sampled to estimate mean berry weight. At the end of the vegetation period, the number of shoots was recorded for each genotype (one individual per block in DP and five in SxG). Then the total wood weight recorded right after pruning was divided by this number to obtain a mean value for each individual in each block (pruning weight).

The 279 genotypes of the DP were controlled for the presence of five important viruses: grapevine fan leaf virus (CNa), grapevine leaf roll associated viruses type 1 to 3 (GLRaV1, GLRaV2, GLRaV3) and grapevine fleck virus (GFkV) (Table 3.S1). For the cluster weight, pruning weight, and cluster length, genotypes infected by grapevine fan leaf virus were automatically removed from the analysis.

Statistical tests and heritability:

Statistical analyses were performed on raw phenotypic data from each year on the DP using R statistical packages (R CORE TEAM 2013). Distribution normality was evaluated using the Shapiro-Wilk test (ROYSTON 1995). Since most data distributions significantly deviated from normality, we used non-parametric procedures to analyse the year effect (Kruskal-Wallis rank sum test) and to calculate and test phenotypic correlations (Spearman rank-order correlation coefficient).

When data distribution deviated from normality, we applied square root (sqrt) or natural logarithm (ln) transformations. These transformed values were used in a mixed model implemented with the function lmer of the R package lme4 (version 1.0-4; (BATES and MAECHLER 2009) to estimate the Best Linear Unbiased Predictors (BLUP) of genetic values across blocks and/or years, for use in GWAS and GS. Models selected were those with the lowest Bayesian Information Criterion (BIC), among several models always including a random genotypic effect, completed or not by fixed year and/or block effects, random effect of theirs interaction (i.e. genotype x block, genotype x year, block x year) and the fixed effect of one to five observed viruses (presence/absence).

For each studied traits the full model was:

$$P_{ijk} = \mu + G_i + b_j + y_k + m_1 + m_2 + m_3 + m_4 + m_5 + e_{ijk};$$

completed by two of the interactions:

$$(G \times b)_{ij}, (G \times y)_{ik}, (b \times y)_{jk}$$

where P_{ijk} was the phenotypic value of genotype *i* in block *j* and year *k*, μ the overall mean, G_i the random effect of genotype *i*, b_j the fixed effect of block *j*, y_k the fixed effect of year *k*, $m_1$to m_5 the fixed effects of viruses (CNa, GLRaV1, GLRaV2, GLRaV3 and GFkV) and e_{ijk} the residual error effect. $(G \times b)_{ij}$ was the interaction between genotype *i* and block *j*, $(G \times y)_{ik}$ the interaction between genotype *i* and year *k* and $(b \times y)_{jk}$ the interaction between block *j* and year *k*.

Variance estimates of the selected models were used to estimate broad-sense heritabilities on an inter-annual genotype mean basis, defined as $H^2 = \sigma_G^2 / (\sigma_G^2 + \sigma_e^2)$, where σ_G^2 and σ_e^2 were the genotypic and residual variances. When the effect of block and/or year was not included in the best model, σ_e^2 was divided by the number of observed years and/or blocks.

Genotyping:

DNA extraction:

For each genotype of the diversity panel, an accession of the Vassal collection (Domain de Vassal, Herault, France) was selected (Table 3.S1) and a batch of young leaves was collected. For Syrah, Grenache and the progenies the collect of the young leaves were performed at the INRA Chapitre experimental station (Hérault, France). Leaves were lyophilized for long-term conservation. One gram of lyophilized leaves was ground using nitrogen liquid. DNA was extracted using the Qiagen DNeasy Plant Maxi kit (Qiagen) following the manufacturer's instructions with minor modifications:

addition of 1% w/v of PVP-40 to the AP1 solution and second washing with absolute ethanol before elution.

Genotyping using the GrapeReSeq 18k Vitis genotyping chip:

The 279 individuals of the DP , Syrah, Grenache and the 21 individuals from the SxG cross were genotyped using the GrapeReSeq 18k Vitis genotyping chip http://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K presented by (LE PASLIER *et al.* 2013). The cluster file used to define the quality of the polymorphism included on the chip was performed at the CNG (Centre National de Génotypage; Evry, France) on a sample of more than two thousand genotypes. Nine different level of quality were identified and assigned at each SNP. In this study only SNP with a quality of 1 were used (perfect definition of the three class of genotype), this class contained 12 815 SNPs.

Genotypic data encoding:

Genotypic data (aa, aA, AA) were encoded in an additive (0, 0.5, 1; add) and a dominant (0, 1, 0; dom) way according to (FOURNIER-LEVEL *et al.* 2009). These sets were used separately or in combination (addom; resulting in twice more locus) to take into account the additive or the dominance effect of each locus and to test combined effects in GWA and GS analyses. Loci and individuals containing more than 20% of missing data were removed. In the remaining data, we imputed the missing genotypes with the corresponding average per SNP (according to (SEGURA *et al.* 2012). We also calculated allele frequency for each SNP locus, in order to filter out rare SNPs, with minor allele frequency (MAF) below 5% that would bias association tests. For GWAS and GS, we used the subset of common SNPs for which markers passed the above described filters in all sub-populations of the DP separately (Figure 3.S1).

Population structure and relatedness:

Population structure was estimated on the 279 individuals of the DP using 20 SSR with STRUCTURE software version 2.3.3 (PRITCHARD *et al.* 2000) accessed through Bioportal (KUMAR *et al.* 2009). We used an admixture model varying the ancestral number of populations (K) from two to five, in order to identify the best K level of population subdivision. We allowed an iterative process with a burn-in phase of 15,000 iterations and a sampling phase of 15,000 iterations. Five replicates of each assumed K level subdivision were compared to estimate group assignation stability. Outputs were visualized

and interpreted with Structure Harvester web v0.6.93 (EARL and VONHOLDT 2011). The optimal group number was chosen based on the estimated 'log probability of data'.

Genetic distance between each pair of individuals was calculated with the Euclidean method. From this matrix a principal coordinate decomposition was computed using the R package ape Version 3.0-11 (PARADIS *et al.* 2004). We visualized the population structure on the plane defined by the two first axes.

Realized relationship matrix (RRM; (EDING and MEUWISSEN 2001) was calculated with R using all filtered SNPs on the 279 individuals.

Genome-wide association:

First a marker by marker approach (emmax) was performed using mlmm script (SEGURA *et al.* 2012) with 5% Bonferroni and FDR thresholds. Then GWAS was performed following FODOR *et al*. (submitted) using the multi-locus mixed-model (mlmm) approach developed by (SEGURA *et al.* 2012), including the population structure as a fixed factor in the mixed model. This R script implements a forward-backward stepwise approach to include significant effects in the mixed model, while re-estimating the variance components of the model at each step. We ran mlmm on the whole DP with a random polygenic term with a variance proportional to the estimated RRM, and a fixed population structure term (three groups) consisting in ancestry fractions estimated by Structure software. We also ran mlmm on each sub-population with the random polygenic term only. The maximal number of forward steps was set to 10. For model selection, we chose the multiple-Bonferroni (mBonf) criterion, selecting the largest model in which all cofactors have a P-value below a Bonferroni-corrected threshold (we used a threshold of 0.05). Cofactor effects were re-estimated at the end of the mlmm analysis and used to predict the genetic value of SxG individuals (see below). For each cofactor identified, the genes were extracted from the annotation data of (GRIMPLET *et al.* 2012), in a window of 10kb.

Prediction methods:

We compared four prediction methods based on the genome-wide SNP data: the sum of effects of markers previously detected in GWAS – using mlmm as described above – corresponding to classical MAS (cof), Ridge Regression BLUP (RR) (HOERL and KENNARD 1970), Bayesian LASSO (Least Absolute Shrinkage and Selection Operator) Regression (BLR) (PÉREZ *et al.* 2010) and a combination of cof and RR-BLUP (cofRR; Fodor *et al.* submitted). We observed the precision of the prediction (Pearson correlation between phenotypes and GEBVs) also in different combinations of training and candidate

populations. The four training populations were WW, WE, TE and the entire DP, comprising between 63 and 242 individuals depending on the studied trait. The Syrah and Grenache with their 21 offspring formed the candidate set.

Cof and cofRR methods were performed as described in Fodor et al (submitted). These methods are based on the significant cofactors identified in mlmm. Their effect is at first re-estimated using a mixed-model, together with variances for genetic (polygenic) and residual random effects, considering the significant markers and the population structure (if exists) as fixed effects. In a second step, the estimates of the associated markers were used for prediction either alone – cof method – or combined with a RR model applied on remaining SNPs – cofRR –.

Ridge Regression performs an extent of shrinkage that is homogenous across markers. For RR we defined the parameter lambda as $\lambda = \sigma_e^2/\sigma_g^2$, where environmental and genetic variances (σ_e^2 and σ_g^2) were estimated via REML in a mixed linear model using emma R library (KANG *et al.* 2008).

The Bayesian LASSO (PARK and CASELLA 2008) method performs stronger shrinkage towards zero for the estimates of small-effect markers, and less for those with high effects. We performed BLR analysis with the R package BLR 1.3 (PÉREZ *et al.* 2010). The lambda parameter was set as random, sampled from a gamma distribution with rate=0.0001 and shape=0.53 according to (PARK and CASELLA 2008). The initial value of λ_0 was calculated using the heritability rules given by DE LOS CAMPOS *et al.* (2012): $\lambda_0 = 2 * n^{-1} * \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}^2 * \frac{(1-h^2)}{h^2}$, where h^2 is the narrow-sense heritability, n is the number of individuals, m is the number of SNPs and X is the matrix of genotypes. σ_e^2 were chosen from the prior $\chi^{-2}(\nu_e, S_e^2)$, where $\nu_e = 4$ to ensure a finite a priori variance, and $S_e^2 = (\nu_e - 2) \times (1 - h^2) \times \sigma_p^2$, where σ_p^2 is the phenotypic variance. σ_g^2 were chosen from the prior $\sigma_g^2 \sim \chi^{-2}(\nu, S^2)$ where ν was set to 4 to ensure a finite a priori variance and $S^2 = (\nu - 2) \times \frac{\sigma_p^2}{n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}^2} \times h^2$. We allowed an iterative process with a burn in phase of 10,000 iterations and a sampling phase of 20,000 replicates.

A ten-fold cross-validation was implemented in order to investigate the feasibility of GS in this kind of material using the available training set and marker density. For this test, individuals of DP were randomly assigned to one of 10 equal folds. Each fold was dropped once from the training set and predicted.

Accuracy was calculated dividing Pearson's correlation coefficient (r) between GEBVs and true phenotypes, by the square root of the broad-sense heritability, in order to compare results across

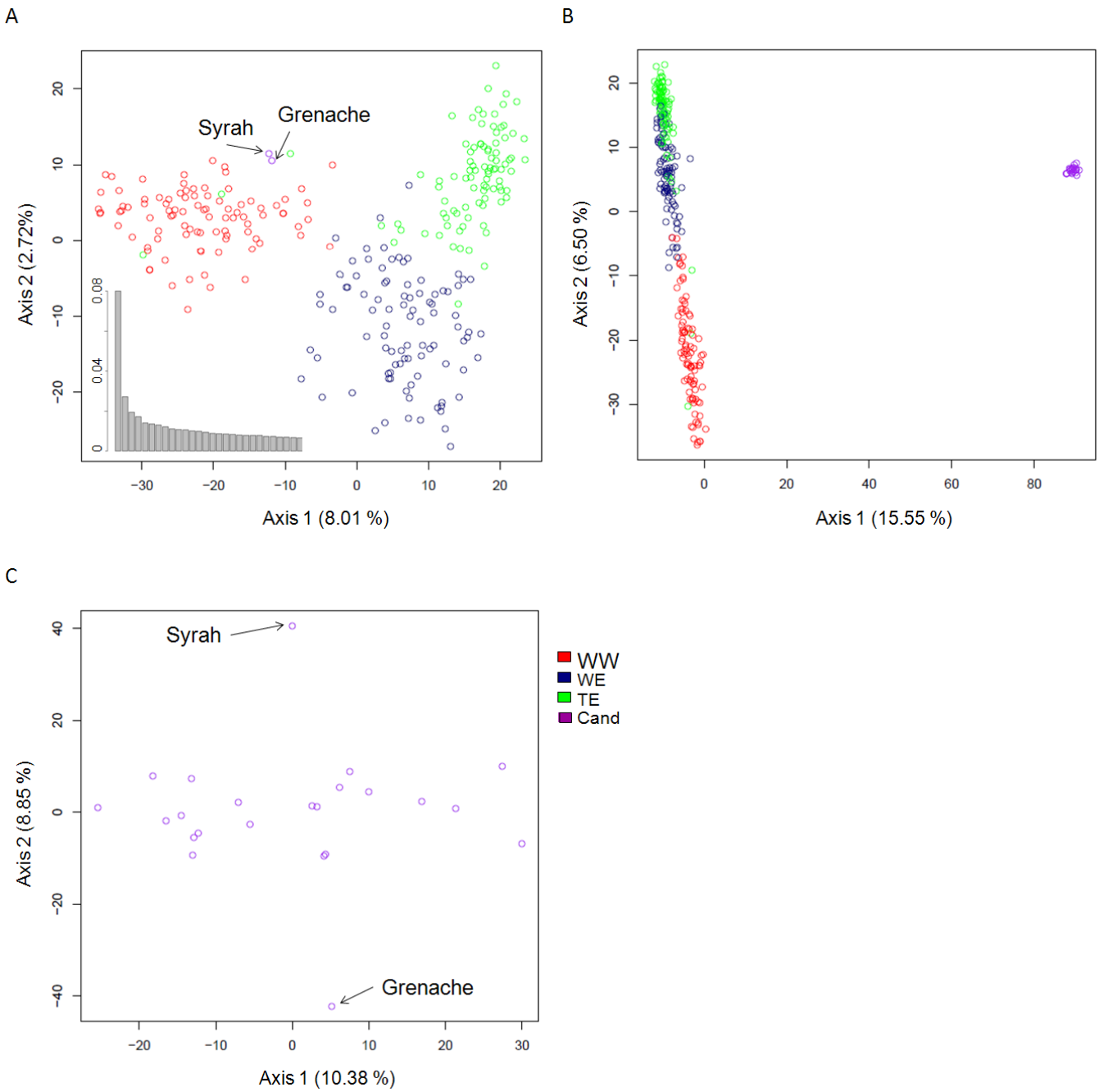


Figure 3.1. Representation of the relationship between individuals in the training and candidate panels through Principal Coordinates Analysis (PCA) performed on SNP markers. Figure A presents the plan of the first 2 axis of the PCA analysis including the diversity panel (DP) as well as Syrah and Grenache. Figure B shows the same plan for the PCA analysis when the SXG progenies were added. Figure C is the plan of the first 2 axis of the PCA calculated on syrah, Grenache and the 21 SXG progenies. For the analysis we used all polymorphic SNPs with up to 20% missing data in the DP (a total of 11,463 SNPs). Red, blue and green colors were attributed following the results of Structure software based on 20 SSR. The barplot on figure A shows the percentages explained by the different axes of the PCA.

traits in the cross-validation. The prediction performance on SxG family was expressed by the Pearson correlation between phenotypes and GEBVs.

RESULTS

Genotypic data:

Out of 18,071 SNPs on the genotyping array, 11,463 SNPs passed the all selection criteria in the DP (quality class 1, polymorphic with less than 20% missing data). In total, 10,058 SNPs were localized on the 19 grapevine chromosomes, 403 on random chromosomes and 709 could not be attributed to chromosomes (UN). The mean distance between SNPs was 42 kb for those localized on the chromosomes (Figure 3.S2). SNPs attributed to chromosomes covered 435.5 Mb corresponding to 90% of the grapevine genome (Figure 3.S3).

After filtering, we identified common SNPs (polymorphic in all sub-populations) and sub-population specific markers. The number of consensus SNPs ranged from 8,316 to 8,581 according to the trait and the encoding of genotypes (Table 3.1) representing 75% (+/-1%) of the total number of filtered markers for each trait. For each trait, the use of dominant encoding reduced the number of consensus markers. The highest number of sub-population specific marker was observed in WW (808 to 872), and the lowest in WE (76 to 113; Figure 3.S1).

Relationship between individuals:

As expected, the analysis with Structure software based on SSR markers, revealed three groups of population structure in the DP. On the SNP-based PCA, these groups were well separated along the first and the second axes of the PCA, explaining respectively 8.01% and 2.72% of the variability (Figure 3.1A). However some individuals attributed in TE by structure analysis were represented in WW on the basis of SNP information. Syrah and Grenache were localized on the periphery of WW group close to each other. When including SxG offspring in the PCA, the first axis clearly differentiated SxG family from the DP (Figure 3.1B). The PCA of SxG family, the offspring show intermediary position between Syrah and Grenache (Figure 3.1C).

Phenotypic data:

The distribution of the four traits was observed in each sub-population of the DP and on SxG (Figure 3.2). ANOVA test identified significant population structure for each trait explaining 10 to 28% of the variability (10% for pruning weight and cluster weight, 19% for cluster length and 28% for berry

Table 3.1. Genotypic and phenotypic data used in this study. This table presents the broad sense heritability for each trait, and the number of individuals with phenotypic data in the training panels (DP: Diversity Panel, WW: wine west, WE: wine east, TE: table east) and in the candidate set (Syrah, Grenache and 21 SxG offspring). The last three columns present the number of SNPs common for all population using three types of genotype encoding (additive, dominant and both).

Trait	Population	H^2	Effective	Observation	add	dom	addom
berry weight	DP	0.87	242*				
	WW		87*	3 years			
	WE		82*		8581	8422	17003
	TE		73*				
	SxG		23	1 year			
cluster weight	DP	0.5	198*				
	WW		70*	3 years			
	WE		65*		8476	8316	16792
	TE		63*				
	SxG		22	1 year			
wood weight	DP	0.45	198*				
	WW		77*	2 years			
	WE		65*		8532	8390	16922
	TE		56*				
	SxG		23	1 year			
cluster length	DP	0.93	201*				
	WW		72*	2 years			
	WE		66*		8510	8347	16857
	TE		63*				
	SxG		23	1 year			

* best linear unbiased predictor (BLUP) of phenotypes.

Table 3.2. Variance explained and correlation between phenotypes and GEBVs obtained using the cofactors identified by mlmm. The column "total" presents the total number of detected associations and column ">10%" contains the number of associations explaining more then 10% of the variance. Pearson correlation between GEBVs obtained using cof method and phenotypic values of the SxG family are presented in the last column.

Trait	H^2	TP*	TP* size	SNP number	Encoding	Nb of cofactors		Var. explained by best mlmm model	Pred. correl. in cof method
						total	>10%		
Berry weight	0.87	WW	87	8422	dom	9	3	81.7%	-0.03
Berry weight	0.87	TE	73	8422	dom	3	3	54.7%	0.45
Cluster length	0.48	TE	63	8347	dom	2	2	37.9%	-0.12
Cluster length	0.48	WE	66	16857	addom	3A + 6D	4	91.7%	-0.05
Clusterlength	0.48	WW	72	8510	add	1	1	27.5%	-0.30

*TP: training population

weight). For berry weight and pruning weight among subpopulations, TE presented the largest diversity including the highest values, while WW individuals were concentrated around lower values, and WE represented an intermediary group. For cluster weight and length, TE and WE presented similar distributions while WW were concentrated around lower values. Distribution of berry weight and pruning weight were similar in SxG and WW, while for cluster weight and cluster length SxG was between TE and WW. Among studied traits we observed two levels of broad-sense heritability: around 0.9 for cluster length and berry weight, and around 0.5 for cluster length and pruning weight (Table 3.1). Equations of the best mixed models for each trait are presented in Table 3.S2.

After phenotypic data treatment and filtering on missing genotypic data, the number of individuals with full data (BLUPs, SNPs) available in DP ranged from 198 to 242 individuals (out of 279) and from 56 to 87 individuals (out of 93) in the sub-populations, depending on the traits (Table 3.1).

GWAS:

Association studies were performed on each sub-population separately and on the entire DP, using three types of genotype encoding (additive, dominant and both together). With the classical emmax approach, only one SNP (for cluster length on chromosome 10 random) passed the 5% Bonferroni and FDR thresholds. Including cofactors in the models through mlmm analysis, we selected the best models explaining the phenotypic variation in the four traits for the four panels (16 combinations).

Best models contained no cofactors for pruning weight and cluster weight, while they contained 0 to 9 cofactors for berry weight and cluster length, according to the tested panel (Table 3.2) for a total of 24 significant SNPs; among them 13 SNPs explained more than 10%. Associations were identified in the sub-populations, but never in DP. The percentage of variance explained varied between 2% and 28% per marker and from 27.5 to 91.7% per model (Table 3.2).

For berry weight, a total of 12 associations were identified (Table 3.S3), all with dominant effects, on nine chromosomes and the chromosome 18 random. Nine associations in WW and three in TE explained respectively 81.7 and 54.7% of the variance observed in the two sub-populations. Between the two sub-populations, only two associations on chromosome 2 were observed close to each other (at approx. 400Kb). For cluster length, associations were detected in all sub-populations with additive and dominant effects but never with both, on seven chromosomes and two random chromosomes. Two of them were localized close to each other for WE and TE on the chromosome 19 at approx. 295Kb.

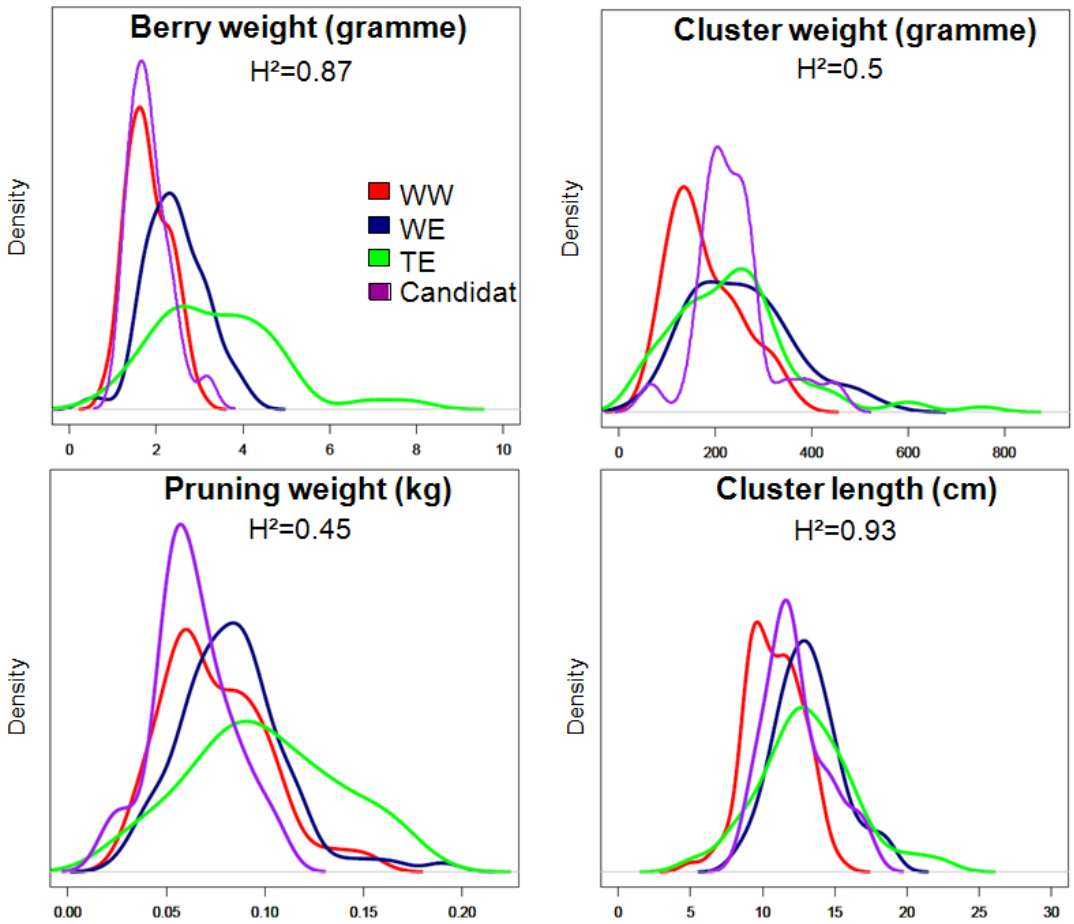


Figure 3.2. Distribution of phenotypes in the studied populations and broad sense heritabilities. WW represent the wine varieties of western origin, WE the wine varieties of eastern origin and TE the table varieties of eastern origin. The candidate set is composed of *cv.* Syrah, *cv.* Grenache and 21 of their progenies. Heritabilities were calculated on the three sub-populations together.

For a given trait in a given sub-population, only one type of encoding allowed to identify significant associations. For the analysis of berry weight, dominant encoding was the most adapted one in both WW and TE. For cluster length, the adapted encoding was different in each sub-population. In WW the associations were detected with the additive encoding, in TE with the dominant one and in WE with the additive and dominant ones together (Table 3.2).

In order to propose candidate genes associated to the studied traits, we identified the genes localized in a window of 10 kb around the SNP indicated by emmax or mlmm. For each SNP we found one or two genes (Table 3.S4) to a total of 35 genes. Among them 9 have no function identified. Close to the SNP identified with both mlmm and emmax (localized on the chromosome 10 random), we found a regulation factor involved in the reproductive development; its functional annotation correspond to the Ovate gene family.

Prediction of phenotypes using genotypes:

In order to test the usefulness of GS in grape using experimental data, we predicted the "genetic estimated breeding values" (GEBVs) of the 23 individuals of the SxG cross for the four traits (i.e. berry weight, cluster weight, pruning weight, cluster length) using the sets of common markers with the three types of genotype encoding (additive, dominant and both together) as presented in Table 3.1. The parameters of the prediction models were estimated using different training sets: the whole DP or each of the sub-populations. In addition, we used different prediction methods: cof, which estimates only the effects of the cofactors identified by mlmm, and three genome-wide prediction methods (RR, BLR and cofRR).

Cross-validation:

The 10-fold cross-validation led to accuracies from 0.45 to 0.7, what proves the pertinence of genome-wide prediction methods, using the set of SNP available and a training set of 180 to 220 individuals of the DP. Differences between genotype encodings and between methods were not significant (Figure 3.3). The cof method could be tested in only one case, where cofactors could be detected on the training set with mlmm (berry weight in additive + dominant genotype encoding).

Prediction of SxG family:

By definition, cof method could only be implemented when cofactors were identified by mlmm (i.e. in five cases; Table 3.2). However, the best model explained a large part of the variance in the studied sub-population, the cofactors identified showed poor precisions at the prediction of SxG.

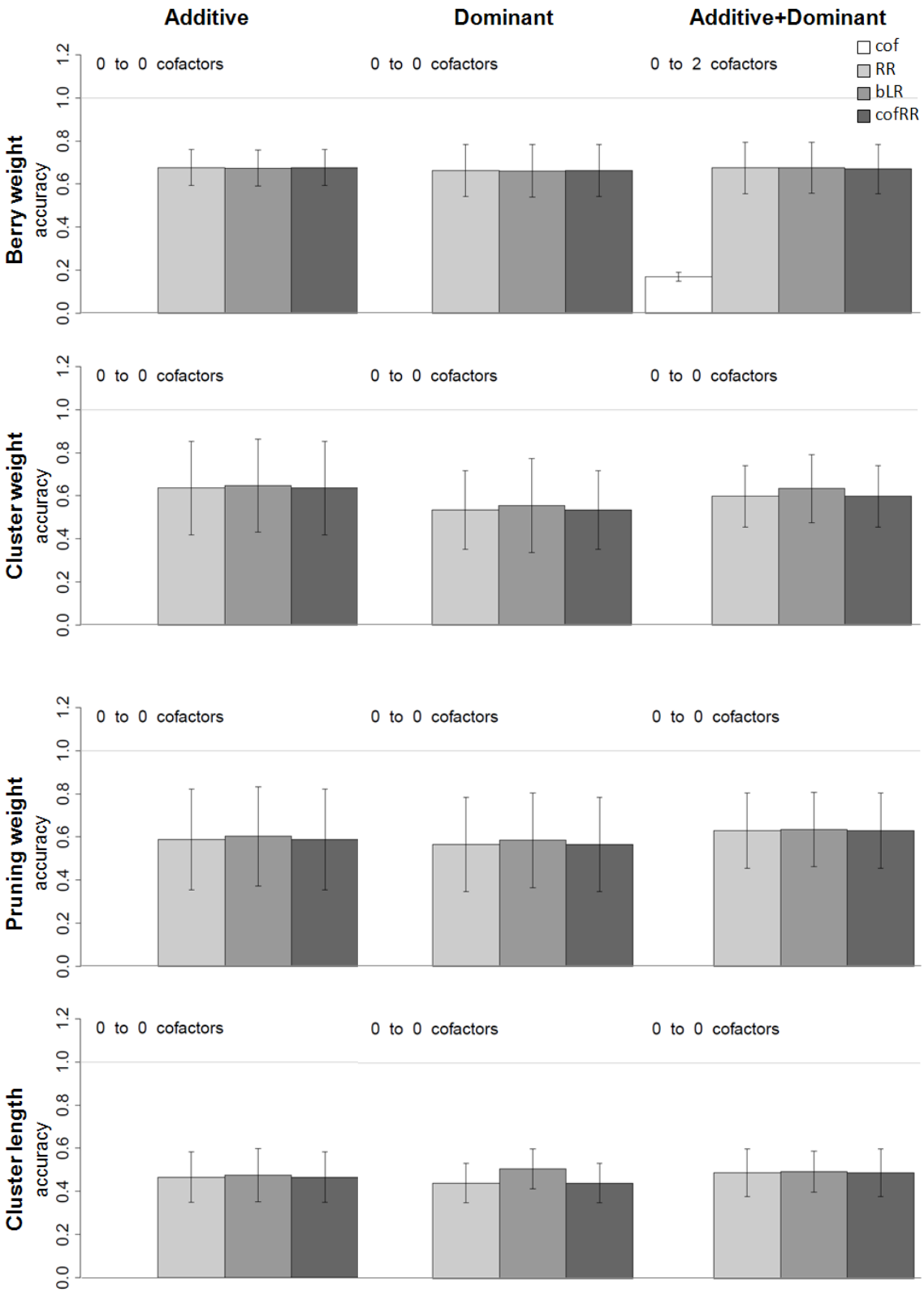


Figure 3.3. Ten fold Cross-validation accuracy in the diversity panel. Error bars were calculated with 95% confidence intervals on the means. On the figure, we indicated the number of cofactors available for cof and cofRR methods.

Except for berry weight – using TE sup-population with dominant encoding –, where cof method achieved a precision of 0.45.

For the genome-wide prediction methods the prediction precisions ranged from -0.38 to +0.40 depending on the trait, the training set and the genotype encoding (Figure 3.4). Globally, even if the size of the training set was higher using the whole DP, it led to less accurate predictions than the ones obtained with sub-population.

Not knowing the genetic architecture of the traits and in particular if the different genes and alleles behaved in an additive, or dominant manner or with mix effects, we tested all possibilities. Out of the 16 combination trait/training population, nine gave positive correlations: either all coding produced positive r (three cases), or two coding gave positive r (one case) or one of the three coding gave positive r, while the other coding gave negative correlations (five cases). Three other combinations gave r around zero, while the four last combinations lead to negative correlations. For each trait however, at least one combination yielded a positive precision.

The best genome-wide prediction over all was obtained for pruning weight using TE as training set and dominant coding (r=0.4). WW as training set with additive coding was only slightly less accurate than TE (r=0.32). For the other traits, prediction using WW was better for berry weight and cluster weight for addom coding, while WE was the best training set for cluster length with dominant coding.

Finally, among the tested parameters, prediction method had the smallest effect on the correlations in this analysis. The efficiency of RR, BLR and cofRR was very similar, except in few cases: berry weight where at least one method was different from the others in each sub-populations, or cluster length for WW as training set.

DISCUSSION

This study presents the first genomic predictions on grapevine and the most powerful genome-wide association study (GWAS) in term of marker density and observed diversity. Experiments were carried out on a diversity panel of 279 individuals (DP), representing the three historical and geographical genetic pools of cultivated *Vitis vinifera*, corresponding to three different morphotypes already proposed by (NEGRUL 1946) using approx. 8,500 polymorphic well distributed SNP markers on four structured traits.

Under these conditions, GWAS and GS lead to better results when working on the sub-populations level rather than with the entire diversity panel (DP). In GWAS by mlmm analysis, we detected 24 significant associations, up to 9 per sub-population, linked to berry weight and cluster length,

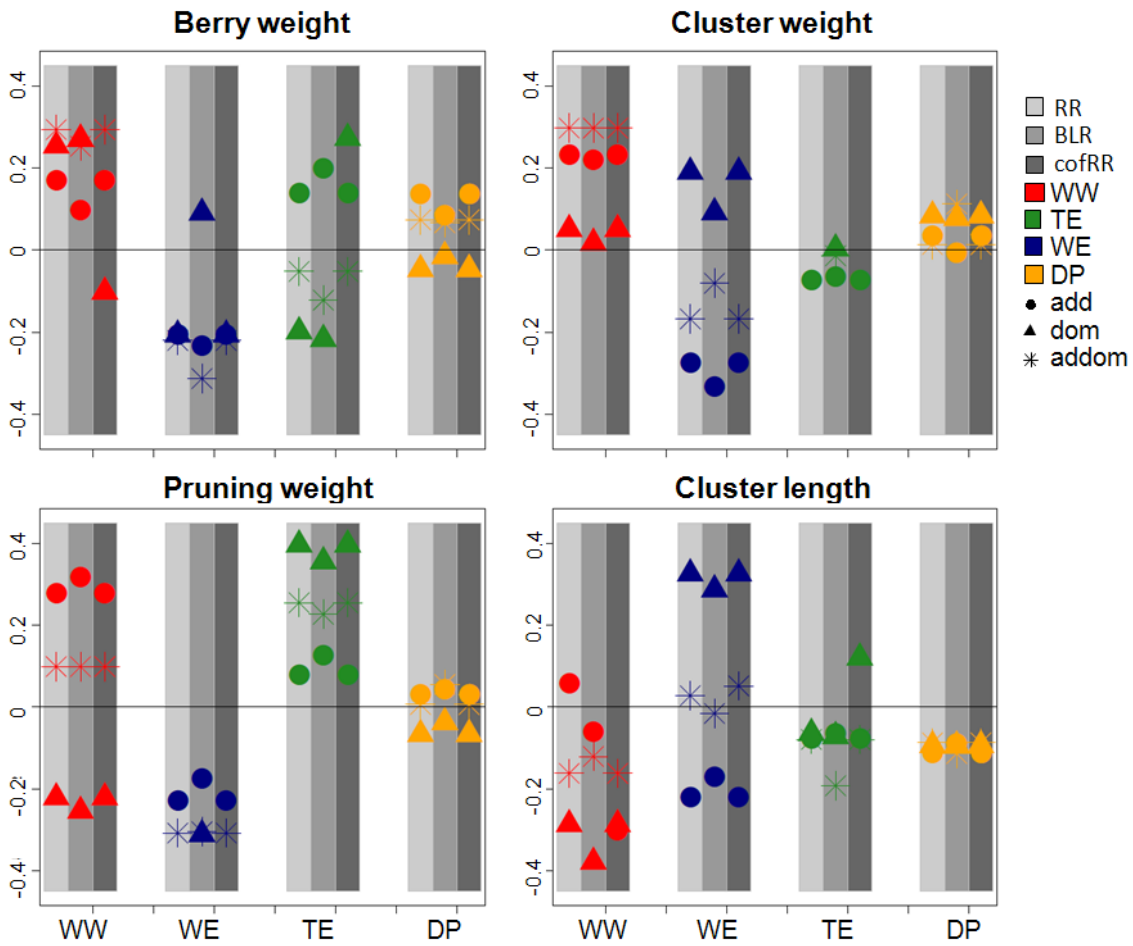


Figure 3.4. Correlation between BLUPs and genomic estimated breeding values (GEBVs) on the SxG cross. On this figure we compare four training populations (WW, WE, TE, all), three prediction model (RR, BLR, cofRR) and three type of genotype encoding (additive, dominant and both together).

explaining 27 to 91% of the variance phenotypic. We thus calculated genomic estimated breeding values (GEBVs) of Syrah, Grenache and 21 SxG progenies using marker effects estimated on various training populations corresponding to highly diverse panels. The best correlation (r) between phenotypes and genomic estimated breeding values (GEBVs) ranged from 0.3 to 0.4 for the four studied traits. Our results show the potential of GS to predict traits linked with population structure. Moreover we propose an approach to predict bi-parental crosses using structured diversity panels at reasonable marker density.

GWAS:

Through all traits, populations and coding tested in our study, we observed only one significant association (at 5% Bonferroni or FDR thresholds) when testing the effect of each marker one by one in a mixed linear model implemented in emmax algorithm (KANG *et al.* 2010). The mlmm approach increased the power of our GWAS thanks to the inclusion of pertinent cofactors in the model (SEGURA *et al.* 2012) enabling us to detect 24 significant associations.

Significant associations were observed in five cases, always using one of the three sub-populations but never on the whole diversity panel (DP). In fact, all four studied traits were linked with the population structure in DP and as described by several authors, the population structure limits the performance of GWAS (CARDON and PALMER 2003; MARCHINI *et al.* 2004). One of the possible solutions to avoid the effect of population structure is to perform GWAS within unstructured population as we did with WW, WE and TE sub-population of *Vitis vinifera* L.

The lack of confirmed co-localizations between sub-populations suggests that different genes are involved in the phenotypic variation of these traits. Taking the example of berry weight, the QTLs captured in WW but not in TE can be explained by the fixation of these loci in TE during the centuries of human selection for larger berries or inversely for those identified in TE and not WW or the introduction of different alleles with different effects in one or the other pool from wild individuals (ARROYO-GARCÍA *et al.* 2006). For cluster length, this suggestion can be completed by the observation, that associations in the different sub-populations were detected with different additive or dominant effects (using additive, dominant or additive+dominant encoding).

In our study, significant associations were detected only for two of the four traits: berry weight and cluster length. These traits, directly linked to the production are easy to select – contrary to pruning weight. Due to selection the LD may be higher around the QTLs underlying these traits than around non-selected traits QTLs. High LD allows the detection of significant associated even with lower marker density, while more markers are needed if the level of LD is weak (MYLES *et al.* 2010).

Cluster weight is a composite trait, determined by the number of berries, the weight of berries, the length and width of the cluster. Interactions between different genetic mechanisms and genes (epistatic effect) are leading to reduced detection power (EAVES 1994).

Finally a part of imprecision can be due to phenotyping process, through biotic or environmental factors that could not be controlled by the statistical model. In example, the vigor of the plants, that can affect biomass production, comprising pruning weight, and the number of representative clusters, may also be related to external factors such as the success of the graft.

In any case, the efficiency of GWAS for polygenic traits strongly depend both on the density of markers and on the size of the observed sample (ZHAO *et al.* 2007; BUCKLER *et al.* 2009; WANG *et al.* 2012). The SNP set used in this study is the largest SNP array actually available for grapevine (LE PASLIER *et al.* 2013) – 18K genotyping chip – and produced approx. 8,500 polymorphic SNPs in our material following the criteria described in the 'materials and methods' part. Nevertheless, because of the rapid decay of LD observed in this specie (10 Kb; (MYLES *et al.* 2010), more markers would be necessary to tag all genomic segments and give ideal conditions for GWAS. In a previous study based on simulation (FODOR *et al*. submitted) we demonstrated that even about 90,000 SNPs, well distributed along the genome with 1,000 individuals lead to association of half of the simulated QTLs in a simple trait or 10% in a complex trait. For better results, one will most certainly need a very high number of markers (over 100,000). Such number of molecular polymorphism could only be obtained by re-sequencing approaches such as GBS (Davey et al. 2011). Meanwhile, increasing the sample size at the sub-populations level can also help to improve the power of GWAS in such high diversity material.

Genomic selection:

Accuracy of the prediction revealed by 10-fold cross-validation ranged from 0.45 to 0.7. These values are similar to those obtained on 17 traits of loblolly pine using RR and BLR methods (0.37-0.73; RESENDE *et al.* 2012), superior than those observed on diverse traits of sugarcane (0.23-0.68; (GOUY *et al.* 2013) but inferior to the accuracies observed for fruit quality traits on the apple (0.67-0.89; (KUMAR *et al.* 2012). The accuracies could be improved increasing the number of SNPs.

Predicting the SxG family, best correlations (r) between BLUPs and GEBVs for each trait ranged from 0.3 to 0.4. For the moment very few papers with such comparison have been published. Higher level of precision (0.53) was observed in dairy cattle (VANRADEN *et al.* (2009), predicting net merit of a validation set using 38,416 SNPs and a training population of 3576 Holstein bulls. (GOUY *et al.* 2013) Obtained correlations ranging from 0.13 to 0.55, for10 argonomic traits in sugarcane, when using

1,499 Dart merkers and two independent panels (one for training set and the other for validation), each composed of 167 accessions representing sugarcane genetic diversity. These results are consistantes with the precisions obtained in our study.

In this study, we also only analyzed the prediction in the first generation of selection. Given the distance observed between the training panels and the progeny, this represents a good result. On subsequent breeding generation, as the distance increases and allele frequencies change, recombination, and inbreeding make prediction of the breeding values even weaker (MEUWISSEN 2009; JANNINK *et al.* 2010). In this case also increasing the number of marker in order to reduce the effect of recombination will be beneficial.

Structured training population:

Despite the larger sample size, DP used as training population lead to weaker estimations than those obtained on more restricted samples. As already described (HABIER *et al.* 2007, 2010) the diversity on which it is build as well as the relatedness between training and candidate populations drive the usefulness of a model. Compared to the DP, the diversity of each sub-population is much more restricted and to some extent closer to the diversity of the candidate population (both in term of phenotype and genotype). In addition, smaller training population more closely related to the candidate set yielded more pertinent information about family relationships than those offered by DP.

In addition, we observed different reliabilities according to the sub-population used in the training step: for berry weight and cluster weight the best prediction was obtained with WW while for pruning weight, it was better with TE (but also good using WW). The candidate set in this study is a bi-parental family, the result of a cross between Syrah and Grenache. Structure analysis realized using 20 SSR by (BACILIERI *et al.* 2013) revealed that Syrah belongs for 97% to WW group, while Grenache belongs for 57% to TE, 41% to WW and 2% to WE. This composition rather than the distribution of the phenotypes in the candidate population may explain that either WW or TE sub-populations lead to better predictions for the SxG family than WE. For cluster length a higher correlation between phenotypes and GEBVs was observed when using WE as training set. The phenotypic mean in TE and WE was very similar for this trait but distribution of the cluster length for the candidates is closer to WE, which may explain this point. In addition, in term of genetic pool, WE has a somewhat intermediate position between the gene pools of TE and WW that could explain its efficiency to predict some traits.

Prediction models:

Globally we did not detect important difference in the efficiency of RR and BLR models, which is in agreement with the results of several authors working with real data (LORENZANA and BERNARDO 2009; RESENDE *et al.* 2012a; KUMAR *et al.* 2012; GOUY *et al.* 2013). Prediction with cofRR differed from RR only when cofactors were identified in mlmm. The results show that fixing the estimated effect of cofactors identified in TE and WE improved the prediction, while the cofactors from WW are misleading it.

In recent years, many different methods were developed to realize predictions (reviewed and compared in MOSER *et al.* 2009; JANNINK *et al.* 2010; DE LOS CAMPOS *et al.* 2012). To take into account a large variety of genetic architectures, Bayesian approaches assume heterogeneity among SNP effects and consider different shapes of the prior distribution for marker effects. (LORENZANA and BERNARDO 2009) recommend simple genome-wide BLUP approach to Bayesian for the prediction of genotypic value in bi-parental plant populations. However, in the diversity panel where large effect alleles can be found for some traits – in ex. the association explaining 28% of the variance for cluster length –, their variation may be better characterized using different priors per markers.

Dominant effects:

The precision of the prediction show a large variation depending on the genotype encoding. These results suggest that the underlying genetic mechanisms involving more dominant effect in one case and more additive effect in another according to the trait and also to the training set. These results are in agreement with the lack of co-localization in GWAS. Since we do not necessarily know the genetic architecture of the traits, we advice to always try the different coding, even using other type of coding, closer to the real dominant situation.

GS or GWAS with the diversity panel?

An essential factor of success for both GS and GWAS methodologies is the ability to capture causal loci (QTLs) by neutral markers thanks to the LD between them. In general, the level of the LD between loci depend on their physical distance, the genomic region and the genetic diversity of the studied panel (GUPTA *et al.* 2005). The aim of a diversity panel is to represent the most genetic variation possible with the less individual (limiting family relationships) to identify the sources of phenotypic variation. This concept helps to obtain plenty of independent genomic fragments, giving a high resolution picture on the whole genome. But to scan it with efficiency, we need markers in LD with all of independent genomic fragments (HAMBLIN *et al.* 2011).

Similarity between related individuals (family relationships) and population structure are disturbing effects (CARDON and PALMER 2003; MARCHINI *et al.* 2004) that we want to control in GWAS by using advanced statistical models and genetic pools (PRITCHARD *et al.* 2000; YU *et al.* 2006), to obtain true associations which represent the causal loci in any genetic background. In the prediction context these components help to learn more from the observed material, for its better characterization through the effect attributed for each marker. This information can be very useful to describe a candidate set reflecting the same LD pattern and population structure (HABIER *et al.* 2007, 2010; WIENTJES *et al.* 2012), but in the other hand it is misleading and results in poor prediction accuracies if candidate set shows different characteristics due to genetic distance (MEUWISSEN 2009; DE ROOS *et al.* 2009) or sampling effect (PSZCZOLA *et al.* 2012; RINCENT *et al.* 2012).

In conclusion, applying GWAS or GS on a diversity panel we are searching for LD between markers and QTLs to capture causal genetic variation and its efficiency depends on the marker density. Limited representation of family relationships increases reliability of GWAS, while it gives only few, neutral information for GS. The population structure must be controlled to perform efficient GWAS, while in GS its role depends on the candidate set (FODOR *et al*. submitted).

Perspectives for GS in grapevine breeding:

As already discussed previously, DP represented too much diversity for the given marker density and population size and probably a too strong structure effect on the studied traits, therefore GWAS and GS on the entire DP was not efficient. Using the sub-populations allowed for both methods to give better results. Increasing the size of each sub-population would nevertheless have increased the statistical power of the population without increasing excessively its diversity and may have given better results. Along further breeding generations, the model developed on a diversity set, may however become much less useful. One solution can be to actualize the training set by including some selected individuals of the progeny (requiring their genotype and phenotype). Several studies showed that the reliability of the prediction is higher when selection candidates are more closely related to the training set. These conditions permit to decrease the needed marker density (MEUWISSEN 2009; HABIER *et al.* 2010).

One of the possibilities is to create a training set using a part of the progeny of the bi-parental cross to predict the other individuals from the same generation, and later the next generations. This solution could provide good prediction reliability –better then the use of DP in long term – but we needs to define new training set for new crosses if parents are not closely related what could diminish the coast-efficiency of the approach.

In conclusion, we have thus demonstrated the ability of GS to predict with a good correlation the GEBV of individuals from breeding population. We have however worked here only within *V. vinifera* background. In the objectives of breeding new resistant cultivars, introduction of other *Vitis* chromosomes segments into *V. vinifera* background may disrupt the prediction power of our models. As multiple back-crosses are not recommended in grapevine because of the high level of inbreeding depression we need to handle the mixed genetic background. Therefore, additional individuals from the resistant *Vitis* species and higher marker density may be needed in order to capture LD between marker and QTL and to better estimate the effect of markers in the new, mixed genetic background.

REFERENCES

ADAM-BLONDON A.-F., LAHOGUE-ESNAULT F., BOUQUET A., BOURSIQUOT J.-M., THIS P., 2001 Usefullness of two SCAR markers for marker-assisted selection of seedless grapevine cultivars. Vitis - Geilweilerhof- **40**: 147–155.

ARROYO-GARCIA R., RUIZ-GARCIA L., BOLLING L., OCETE R., LOPEZ M. A., ARNOLD C., ERGUL A., SÖYLEMEZO˝LU G., UZUN H. I., CABELLO F., IBAÑEZ J., ARADHYA M. K., ATANASSOV A., ATANASSOV I., BALINT S., CENIS J. L., COSTANTINI L., GORISLAVETS S., GRANDO M. S., KLEIN B. Y., MCGOVERN P. E., MERDINOGLU D., PEJIC I., PELSY F., PRIMIKIRIOS N., RISOVANNAYA V., ROUBELAKIS-ANGELAKIS K. A., SNOUSSI H., SOTIRI P., TAMHANKAR S., THIS P., TROSHIN L., MALPICA J. M., LEFORT F., MARTINEZ-ZAPATER J. M., 2006 Multiple origins of cultivated grapevine (Vitis vinifera L. ssp. sativa) based on chloroplast DNA polymorphisms. Mol. Ecol. **15**: 3707–3714.

ATWELL S., HUANG Y. S., VILHJALMSSON B. J., WILLEMS G., HORTON M., LI Y., MENG D., PLATT A., TARONE A. M., HU T. T., JIANG R., MULIYATI N. W., ZHANG X., AMER M. A., BAXTER I., BRACHI B., CHORY J., DEAN C., DEBIEU M., MEAUX J. DE, ECKER J. R., FAURE N., KNISKERN J. M., JONES J. D. G., MICHAEL T., NEMRI A., ROUX F., SALT D. E., TANG C., TODESCO M., TRAW M. B., WEIGEL D., MARJORAM P., BOREVITZ J. O., BERGELSON J., NORDBORG M., 2010 Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature **465**: 627–631.

BACILIERI R., LACOMBE T., CUNFF L. LE, VECCHI-STARAZ M. D., LAUCOU V., GENNA B., PEROS J.-P., THIS P., BOURSIQUOT J.-M., 2013 Genetic structure in cultivated grapevines is linked to geography and human selection. BMC Plant Biol. **13**: 25.

BATES D., MAECHLER M., 2009 lme4: Linear mixed-effects models using {S4} classes.{R} package version 0.999375-32.

BOUQUET A., JUGA J., 2013 Integrating genomic selection into dairy cattle breeding programmes: a review. Anim. Int. J. Anim. Biosci. **7**: 705–713.

BOURSIQUOT J.-M., LACOMBE T., LAUCOU V., JULLIARD S., PERRIN F.-X., LANIER N., LEGRAND D., MEREDITH C., THIS P., 2009 Parentage of Merlot and related winegrape cultivars of southwestern France: discovery of the missing link. Aust. J. Grape Wine Res. **15**: 144–155.

BOWERS, BOURSIQUOT, THIS, CHU, JOHANSSON, MEREDITH, 1999 Historical Genetics: The Parentage of Chardonnay, Gamay, and Other Wine Grapes of Northeastern France. Science **285**: 1562–1565.

BUCKLER E. S., HOLLAND J. B., BRADBURY P. J., ACHARYA C. B., BROWN P. J., BROWNE C., ERSOZ E., FLINT-GARCIA S., GARCIA A., GLAUBITZ J. C., GOODMAN M. M., HARJES C., GUILL K., KROON D. E., LARSSON S., LEPAK N. K., LI H., MITCHELL S. E., PRESSOIR G., PEIFFER J. A., ROSAS M. O., ROCHEFORD T. R., ROMAY M. C., ROMERO S., SALVO S., VILLEDA H. S., SILVA H. S. da, SUN Q., TIAN F., UPADYAYULA N., WARE D., YATES H., YU J., ZHANG Z., KRESOVICH S., MCMULLEN M. D., 2009 The Genetic Architecture of Maize Flowering Time. Science **325**: 714–718.

CAMPOS G. DE LOS, HICKEY J. M., PONG-WONG R., DAETWYLER H. D., CALUS M. P. L., 2012 Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics **193**: 327–345.

CARDON L. R., PALMER L. J., 2003 Population stratification and spurious allelic association. Lancet **361**: 598–604.

CARRIER G., HUANG Y.-F., CUNFF L. LE, FOURNIER-LEVEL A., VIALET S., SOUQUET J.-M., CHEYNIER V., TERRIER N., THIS P., 2013 Selection of candidate genes for grape proanthocyanidin pathway by an integrative approach. Plant Physiol. Biochem.

COLEMAN C., COPETTI D., CIPRIANI G., HOFFMANN S., KOZMA P., KOVACS L., MORGANTE M., TESTOLIN R., GASPERO G. DI, 2009 The powdery mildew resistance gene REN1 co-segregates with an NBS-LRR gene cluster in two Central Asian grapevines. BMC Genet. **10**: 89.

EARL D. A., VONHOLDT B. M., 2011 STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. **4**: 359–361.

EAVES L. J., 1994 Effect of genetic architecture on the power of human linkage studies to resolve the contribution of quantitative trait loci. Heredity (Edinb) **72 (Pt 2)**: 175–192.

EDING H., MEUWISSEN T. H. E., 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J. Anim. Breed. Genet. **118**: 141–159.

ELSHIRE R. J., GLAUBITZ J. C., SUN Q., POLAND J. A., KAWAMOTO K., BUCKLER E. S., MITCHELL S. E., 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L Orban, Ed.). PLoS ONE **6**: e19379.

EMANUELLI F., BATTILANA J., COSTANTINI L., CUNFF L. L., BOURSIQUOT J.-M., THIS P., GRANDO M. S., 2010 A candidate gene association study on muscat flavor in grapevine (Vitis vinifera L.). BMC Plant Biol. **10**: 241.

EMANUELLI F., LORENZI S., GRZESKOWIAK L., CATALANO V., STEFANINI M., TROGGIO M., MYLES S., MARTINEZ-ZAPATER J. M., ZYPRIAN E., MOREIRA F. M., GRANDO M. S., 2013 Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biol. **13**: 39.

FISCHER B. M., SALAKHUTDINOV I., AKKURT M., EIBACH R., EDWARDS K. J., TÖPFER R., ZYPRIAN E. M., 2004 Quantitative trait locus analysis of fungal disease resistance factors on a molecular map of grapevine. TAG Theor. Appl. Genet. Theor. Angew. Genet. **108**: 501–515.

FOURNIER-LEVEL A., CUNFF L. L., GOMEZ C., DOLIGEZ A., AGEORGES A., ROUX C., BERTRAND Y., SOUQUET J.-M., CHEYNIER V., THIS P., 2009 Quantitative Genetic Bases of Anthocyanin Variation in Grape (Vitis vinifera L. ssp. sativa) Berry: A Quantitative Trait Locus to Quantitative Trait Nucleotide Integrated Study. Genetics **183**: 1127–1139.

FOURNIER-LEVEL A., LACOMBE T., CUNFF L. LE, BOURSIQUOT J.-M., THIS P., 2010 Evolution of the VvMybA gene family, the major determinant of berry colour in cultivated grapevine (Vitis vinifera L.). Heredity **104**: 351–362.

GASPERO G. DI, COPETTI D., COLEMAN C., CASTELLARIN S. D., EIBACH R., KOZMA P., LACOMBE T., GAMBETTA G., ZVYAGIN A., CINDRIC P., KOVACS L., MORGANTE M., TESTOLIN R., 2012 Selective sweep at the Rpv3 locus during grapevine breeding for downy mildew resistance. Theor. Appl. Genet. **124**: 277–286.

GOUY M., ROUSSELLE Y., BASTIANELLI D., LECOMTE P., BONNAL L., ROQUES D., EFILE J.-C., ROCHER S., DAUGROIS J., TOUBI L., NABENEZA S., HERVOUET C., TELISMART H., DENIS M., THONG-CHANE A., GLASZMANN J. C., HOARAU J.-Y., NIBOUCHE S., COSTET L., 2013 Experimental assessment of the accuracy of genomic selection in sugarcane. Theor. Appl. Genet. **126**: 2575–2586.

GRATTAPAGLIA D., RESENDE M. D. V., 2010 Genomic selection in forest tree breeding. Tree Genet. Genomes **7**: 241–255.

GRIMPLET J., HEMERT J. VAN, CARBONELL-BEJERANO P., DIAZ-RIQUELME J., DICKERSON J., FENNELL A., PEZZOTTI M., MARTINEZ-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.

GUPTA P. K., RUSTGI S., KULWAL P. L., 2005 Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol. Biol. **57**: 461–485.

HABIER D., FERNANDO R. L., DEKKERS J. C. M., 2007 The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics **177**: 2389–2397.

HABIER D., TETENS J., SEEFRIED F.-R., LICHTNER P., THALLER G., 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. **42**: 5.

HAMBLIN M. T., BUCKLER E. S., JANNINK J.-L., 2011 Population genetics of genomics-based crop improvement methods. Trends Genet. **27**: 98–106.

HANNAH L., ROEHRDANZ P. R., IKEGAMI M., SHEPARD A. V., SHAW M. R., TABOR G., ZHI L., MARQUET P. A., HIJMANS R. J., 2013 Climate change, wine, and conservation. Proc. Natl. Acad. Sci. **110**: 6907–6912.

HAYES B. J., BOWMAN P. J., CHAMBERLAIN A. J., GODDARD M. E., 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. **92**: 433–443.

HAYES B. J., COGAN N. O. I., PEMBLETON L. W., GODDARD M. E., WANG J., SPANGENBERG G. C., FORSTER J. W., 2013 Prospects for genomic selection in forage plant species (OA Rognli, Ed.). Plant Breed. **132**: 133–143.

HEFFNER E. L., LORENZ A. J., JANNINK J.-L., SORRELLS M. E., 2010 Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Sci. **50**: 1681–1690.

HEFFNER E. L., SORRELLS M. E., JANNINK J.-L., 2009 Genomic Selection for Crop Improvement. Crop Sci. **49**: 1.

HOERL A. E., KENNARD R. W., 1970 Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics **12**: 55–67.

HOUEL C., 2011 Caractérisation de la variation phénotypique de la taille de la baie chez la vigne Vitis vinifera L. et approches de génétique d'association et de recherche de traces de sélection pour ce caractère.

HOUEL C., MARTIN-MAGNIETTE M.-L., NICOLAS S. d., LACOMBE T., CUNFF L. LE, FRANCK D., TORREGROSA L., CONEJERO G., LALET S., THIS P., ADAM-BLONDON A.-F., 2013 Genetic variability of berry size in the grapevine (Vitis vinifera L.). Aust. J. Grape Wine Res.: n/a–n/a.

HUANG Y.-F., DOLIGEZ A., FOURNIER-LEVEL A., CUNFF L. LE, BERTRAND Y., CANAGUIER A., MOREL C., MIRALLES V., VERAN F., SOUQUET J.-M., CHEYNIER V., TERRIER N., THIS P., 2012 Dissecting genetic architecture of grape proanthocyanidin composition through quantitative trait locus mapping. BMC Plant Biol. **12**: 30.

HUANG X., WEI X., SANG T., ZHAO Q., FENG Q., ZHAO Y., LI C., ZHU C., LU T., ZHANG Z., LI M., FAN D., GUO Y., WANG A., WANG L., DENG L., LI W., LU Y., WENG Q., LIU K., HUANG T., ZHOU T., JING Y., LI W., LIN Z., BUCKLER E. S., QIAN Q., ZHANG Q.-F., LI J., HAN B., 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. **42**: 961–967.

IWATA H., JANNINK J.-L., 2011 Accuracy of Genomic Selection Prediction in Barley Breeding Programs: A Simulation Study Based On the Real Single Nucleotide Polymorphism Data of Barley Breeding Lines. Crop Sci. **51**: 1915.

JANNINK J.-L., LORENZ A. J., IWATA H., 2010 Genomic selection in plant breeding: from theory to practice. Briefings Funct. Genomics **9**: 166–177.

KANG H. M., SUL J. H., SERVICE S. K., ZAITLEN N. A., KONG S., FREIMER N. B., SABATTI C., ESKIN E., 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42**: 348–354.

KANG H. M., ZAITLEN N. A., WADE C. M., KIRBY A., HECKERMAN D., DALY M. J., ESKIN E., 2008 Efficient Control of Population Structure in Model Organism Association Mapping. Genetics **178**: 1709–1723.

KUMAR S., CHAGNE D., BINK M. C. A. M., VOLZ R. K., WHITWORTH C., CARLISLE C., 2012 Genomic Selection for Fruit Quality Traits in Apple (Malus×domestica Borkh.). PLoS ONE **7**: e36674.

KUMAR S., SKJÆVELAND Å., ORR R. J., ENGER P., RUDEN T., MEVIK B.-H., BURKI F., BOTNEN A., SHALCHIAN-TABRIZI K., 2009 AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. BMC Bioinformatics **10**: 357.

LACOMBE T., BOURSIQUOT J.-M., LAUCOU V., VECCHI-STARAZ M. DI, PEROS J.-P., THIS P., 2012 Large-scale parentage analysis in an extended set of grapevine cultivars (<i>Vitis vinifera</i> L.). Theor. Appl. Genet.: 1–14.

LAUCOU V., LACOMBE T., DECHESNE F., SIRET R., BRUNO J.-P., DESSUP M., DESSUP T., ORTIGOSA P., PARRA P., ROUX C., SANTONI S., VARES D., PEROS J.-P., BOURSIQUOT J.-M., THIS P., 2011 High throughput analysis of grape genetic diversity as a tool for germplasm collection management. TAG Theor. Appl. Genet. Theor. Angew. Genet. **122**: 1233–1245.

LIJAVETZKY D., CABEZAS J., IBAÑEZ A., RODRIGUEZ V., MARTINEZ-ZAPATER J. M., 2007 High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics **8**: 424.

LORENZANA R. E., BERNARDO R., 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. **120**: 151–161.

MARCHINI J., CARDON L. R., PHILLIPS M. S., DONNELLY P., 2004 The effects of human population structure on large genetic association studies. Nat. Genet. **36**: 512–517.

MARGUERIT E., BOURY C., MANICKI A., DONNART M., BUTTERLIN G., NEMORIN A., WIEDEMANN-MERDINOGLU S., MERDINOGLU D., OLLAT N., DECROOCQ S., 2009 Genetic dissection of sex determinism, inflorescence morphology and downy mildew resistance in grapevine. Theor. Appl. Genet. **118**: 1261–1278.

MEJIA N., SOTO B., GUERRERO M., CASANUEVA X., HOUEL C., ÁNGELES MICCONO M. DE LOS, RAMOS R., CUNFF L. LE, BOURSIQUOT J.-M., HINRICHSEN P., ADAM-BLONDON A.-F., 2011 Molecular, genetic and transcriptional evidence for a role of VvAGL11 in stenospermocarpic seedlessness in grapevine. BMC Plant Biol. **11**: 57.

MEUWISSEN T. H., 2009 Accuracy of breeding values of “unrelated” individuals predicted by dense SNP genotyping. Genet. Sel. Evol. **41**: 35.

MEUWISSEN T. H. E., HAYES B. J., GODDARD M. E., 2001 Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics **157**: 1819–1829.

MORIONDO M., JONES G. V., BOIS B., DIBARI C., FERRISE R., TROMBI G., BINDI M., 2013 Projected shifts of wine regions in response to climate change. Clim. Change **119**: 825–839.

MOSER G., TIER B., CRUMP R. E., KHATKAR M. S., RAADSMA H. W., 2009 A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. **41**: 56.

MYLES S., BOYKO A. R., OWENS C. L., BROWN P. J., GRASSI F., ARADHYA M. K., PRINS B., REYNOLDS A., CHIA J.-M., WARE D., BUSTAMANTE C. D., BUCKLER E. S., 2011 Genetic Structure and Domestication History of the Grape. Proc. Natl. Acad. Sci. **108**: 3530–3535.

MYLES S., CHIA J.-M., HURWITZ B., SIMON C., ZHONG G. Y., BUCKLER E., WARE D., 2010 Rapid Genomic Characterization of the Genus Vitis. PLoS ONE **5**: e8219.

NAKAYA A., ISOBE S. N., 2012 Will genomic selection be a practical method for plant breeding? Ann. Bot. **110**: 1303–1316.

NEGRUL A. M., 1946 *Ampelography of USSR*. Frolov-Bagreev A., Moscow.

OLLAT N., FERNANDEZ L., ROMIEU C., DUCHENE E., LISSARAGUE J. R., LECOURIEUX D., AGEORGES A., KELLY M., CACHO J., RIVARS J., LAMUELA R., GOUTOULY J. P., LEEUWEN C. VAN, MARGUERIT E., PECCOUX A., BARRIEU F., LEBON E., THIS P., PELLEGRINO A., MARTINEZ-ZAPATER J. M., TORREGROSA L., 2011 Multidisciplinary research to select new cultivars adapted to climate changes. In: Asti and Alba, Italy.

PARADIS E., CLAUDE J., STRIMMER K., 2004 APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics **20**: 289–290.

PARK T., CASELLA G., 2008 The Bayesian Lasso. J. Am. Stat. Assoc. **103**: 681–686.

PASLIER M.-C. LE, CHOISNE R., BACILIERI R., BOURSIQUOT J.-M., BRAS M., BRUNEL D., GASPERO G. DI, HAUSMANN L., LACOMBE T., LAUCOU V., LAUNAY A., MARTINEZ-ZAPATER J., MORGANTE M., RAJ P., PONNAIAH M., QUESNEVILLE H., SCALABRIN S., TORRES-PEREZ R., ADAM-BLONDON A.-F., 2013 The GrapeReSeq 18k Vitis genotyping chip. In: La Serena, Chile.

PAUQUET J., BOUQUET A., THIS P., ADAM-BLONDON A.-F., 2001 Establishment of a local map of AFLP markers around the powdery mildew resistance gene Run1 in grapevine and assessment of their usefulness for marker assisted selection. Theor. Appl. Genet. **103**: 1201–1210.

PEREZ P., CAMPOS G. DE LOS, CROSSA J., GIANOLA D., 2010 Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. Plant Genome J. **3**: 106–116.

POLAND J. A., RIFE T. W., 2012 Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome J. **5**: 92.

PRITCHARD J. K., STEPHENS M., DONNELLY P., 2000 Inference of Population Structure Using Multilocus Genotype Data. Genetics **155**: 945–959.

PSZCZOLA M., STRABEL T., MULDER H. A., CALUS M. P. L., 2012 Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. **95**: 389–400.

R CORE TEAM, 2013 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

RESENDE M. F. R., MUNOZ P., RESENDE M. D. V., GARRICK D. J., FERNANDO R. L., DAVIS J. M., JOKELA E. J., MARTIN T. A., PETER G. F., KIRST M., 2012 Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics **190**: 1503–1510.

RIAZ S., BOURSIQUOT J.-M., DANGL G. S., LACOMBE T., LAUCOU V., TENSCHER A. C., WALKER M., 2013 Identification of mildew resistance in wild and cultivated Central Asian grape germplasm. BMC Plant Biol. **13**: 149.

RINCENT R., LALOË D., NICOLAS S., ALTMANN T., BRUNEL D., REVILLA P., RODRIGUEZ V. M., MORENO-GONZALEZ J., MELCHINGER A., BAUER E., SCHOEN C.-C., MEYER N., GIAUFFRET C., BAULAND C., JAMIN P., LABORDE J., MONOD H., FLAMENT P., CHARCOSSET A., MOREAU L., 2012 Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics **192**: 715–728.

ROOS A. P. W. DE, HAYES B. J., GODDARD M. E., 2009 Reliability of Genomic Predictions Across Multiple Populations. Genetics **183**: 1545–1553.

ROYSTON P., 1995 Calculation of unconditional and conditional reference intervals for foetal size and growth from longitudinal measurements. Stat. Med. **14**: 1417–1436.

SCHWANDER F., EIBACH R., FECHTER I., HAUSMANN L., ZYPRIAN E., TÖPFER R., 2012 Rpv10: a new locus from the Asian Vitis gene pool for pyramiding downy mildew resistance loci in grapevine. Theor. Appl. Genet. **124**: 163–176.

SEGURA V., VILHJALMSSON B. J., PLATT A., KORTE A., SEREN Ü., LONG Q., NORDBORG M., 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. **44**: 825–830.

SITZENSTOCK F., YTOURNEL F., SHARIFI A. R., CAVERO D., TÄUBERT H., PREISINGER R., SIMIANER H., 2013 Efficiency of genomic selection in an established commercial layer breeding program. Genet. Sel. Evol. **45**: 29.

STORLIE E., CHARMET G., 2013 Genomic Selection Accuracy using Historical Data Generated in a Wheat Breeding Program. Plant Genome **6**: 0.

TIAN F., BRADBURY P. J., BROWN P. J., HUNG H., SUN Q., FLINT-GARCIA S., ROCHEFORD T. R., MCMULLEN M. D., HOLLAND J. B., BUCKLER E. S., 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. **43**: 159–162.

VANRADEN P. M., TASSELL C. P. VAN, WIGGANS G. R., SONSTEGARD T. S., SCHNABEL R. D., TAYLOR J. F., SCHENKEL F. S., 2009 Invited Review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. **92**: 16–24.

WANG M., JIANG N., JIA T., LEACH L., COCKRAM J., WAUGH R., RAMSAY L., THOMAS B., LUO Z., 2012 Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theor. Appl. Genet. **124**: 233–246.

WIENTJES Y. C. J., VEERKAMP R. F., CALUS M. P. L., 2012 The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics **193**: 621–631.

YU J., PRESSOIR G., BRIGGS W. H., VROH BI I., YAMASAKI M., DOEBLEY J. F., MCMULLEN M. D., GAUT B. S., NIELSEN D. M., HOLLAND J. B., KRESOVICH S., BUCKLER E. S., 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38**: 203–208.

ZHAO K., ARANZANA M. J., KIM S., LISTER C., SHINDO C., TANG C., TOOMAJIAN C., ZHENG H., DEAN C., MARJORAM P., NORDBORG M., 2007 An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet **3**: e4.

CHAPITRE 4 : PREDICTION DES GENOTYPES SUR UNE POPULATION BIPARENTALE

1. Introduction

L'article présenté dans le chapitre 3 utilisait comme population de validation 21 descendants d'un croisement entre Syrah et Grenache. La population d'entrainement comprend un échantillon représentatif du compartiment cultivé de l'espèce *Vitis vinifera* L. (panel de diversité DP; (HOUEL *et al.* 2013). Les précisions observées dans cette étude varient entre 0,3 et 0,4 en fonction du caractère étudié mais aussi de la population d'entrainement. Nous avons observé que les précisions de prédiction obtenues avec le DP complet sont toujours inférieures à celles obtenues en utilisant une des sous-populations comme population d'entrainement. Nous avons formulé deux hypothèses pour décrire ce résultat : i. la diversité génétique à estimer par les marqueurs était différente dans les sous-populations (effet génétique) et dans DP (plus grand nombre d'effets génétiques + effet structure); ii. l'apparentement entre la population candidat et les sous-populations d'origine des parents est plus important qu'entre la population candidat et le DP complet. Ces résultats sont en accord avec ceux observés par des nombreux auteurs (HABIER *et al.* 2007, 2010; MEUWISSEN 2009) et nous suggèrent i. d'augmenter le nombre de marqueur utilisés, ou ii. de réduire la distance entre la population d'entrainement et les candidats à prédire pour obtenir des prédictions plus fiables. Pour assurer un bon niveau d'apparentement (réduire la distance génétique) entre les populations d'entrainement et candidate, la solution le plus répandue est d'intégrer dans la population d'entrainement des individus apparentés aux géniteurs utilisés pour créer la population candidate. En effet dans le cas de la plupart des programmes de sélection qui comptent implémenter – ou qui ont déjà implémenté – la sélection génomique, la population d'entrainement contient des individus apparentés avec les candidats de la sélection. (HEFFNER *et al.* 2010; IWATA and JANNINK 2011; SITZENSTOCK *et al.* 2013; HAYES *et al.* 2013; BOUQUET and JUGA 2013).

Dans ce chapitre, nous explorons le niveau de précision que l'on peut obtenir lorsque le niveau d'apparentement entre les populations d'entrainement et candidate est maximal. En effet nous utiliserons comme population d'entrainement et de validation des individus issus du même croisement, en réalisant des cross-validations au sein de la descendance Syrah x Grenache (SxG). Nous testons également différentes proportions entre population d'entrainement et candidate en faisant varier le ratio entre le nombre d'individus dans la population d'entrainement et le nombre d'individus dans la population de validation.

2. Matériels et méthodes

Le matériel végétal et les données phénotypiques utilisées pour cette analyse proviennent d'une étude QTL réalisée dans l'équipe (DOLIGEZ *et al*. soumis). Plus précisément, il s'agit d'une population composée de 189 descendants issus du croisement Syrah X Grenache, greffés et plantés en 2003 au

domaine INRA/Montpellier SupAgro du Chapitre (Hérault, France) en deux blocs complet randomisé. Chaque génotype a été planté en placette élémentaire comprenant cinq souches. Parmi les caractères étudiés dans DOLIGEZ *et al*. (soumis) nous avons choisi la taille de la baie – un caractère complexe et structuré – permettant une comparaison avec les résultats des chapitres précédents. Les phénotypes ont été observés pendant trois ans de la manière suivante : huit grappes ont été récoltées à maturité (20 degré de Brix). Les extrémités ont été éliminées, puis 100 baies ont été aléatoirement sélectionnées et pesées. Les phénotypes brutes ont été transformés (racine carrée) afin d'approcher une distribution normale. Des valeurs de BLUP (Best Linear Unbiased Predictor) ont été extraites pour chaque individu d'un modèle mixte comprenant l'effet aléatoire du génotype, l'effet fixe des blocs et des années et leurs interactions. Le meilleur modèle a été sélectionné selon le critère de BIC (Bayesian Information Criterion). L'héritabilité au sens large a été estimée à 0,79 (DOLIGEZ *et al*. soumis).

Les données génotypiques correspondent à 127 loci microsatellites génotypés sur tous les descendants SxG. Nous avons codé les loci séparément en fonction de leurs origines parentales : les loci homozygotes et hétérozygotes des parents ont été codés respectivement 0 et 1. De ce fait chaque marqueur est doublé avec un codage et une ségrégation des allèles Syrah et un codage et une ségrégation pour les allèles Grenache. Ce type de codage a été utilisé pour éviter un classement hiérarchique des génotypes (si nous les avions codés 1, 2, 3 et 4). Ainsi nous avons obtenu 2x127=254 loci pour nos analyses. Les données manquantes supérieures à 20% ont été retirées du jeu de données, ainsi que les loci pour lesquels la fréquence d'allèle minoritaire (MAF) était inférieure à 5%. La matrice de génotypes ainsi obtenue nous a permis d'estimer la matrice d'apparentement de type RRM (Realized relationship matrix; (EDING and MEUWISSEN 2001) par un script (File S1 de l'article N°1) réalisé avec le logiciel R (R CORE TEAM 2013).

La prédiction des phénotypes avec ces génotypes a été réalisée avec les méthodes et paramètres présentés dans le deuxième chapitre de la thèse (FODOR *et al*. in prep.). Nous avons utilisé les méthodes cof (FODOR *et al*. soumis), Ridge-Regression BLUP (HOERL and KENNARD 1970), Bayesian LASSO (Least Absolute Shrinkage and Selection Operator) Regression (BLR ; (PEREZ *et al.* 2010) et une combinaison de cof et de RR (cofRR ; FODOR *et al*. soumis). Brièvement, la « cof » méthode utilise des marqueurs identifiés dans une étude d'association « genome-wide » (Genome-Wide Association Study ; GWAS) réalisé par l'approche mlmm (SEGURA *et al.* 2012). Cette méthode utilise un algorithme de type « stepwise forward-backward », permettant d'intégrer des associations pertinentes dans le modèle d'association comme cofacteur fixe. Pour la « cof » méthode, l'effet fixe des cofacteurs a été estimé à partir du modèle mixte le plus pertinent (selon le critère de mutiple Bonferroni) intégrant la matrice d'apparentement et l'erreur résiduelle comme effet aléatoire (File S1 de l'article N°1). Les

Tableau 4.1. Résultats de l'étude d'association en comparaison avec les résultats de la détection de QTL (DOLIGEZ *et al*. soumis).

	étude QTL				GWAS			
GL	Carte	Intervalle de confiance	Max LOD score	effet détecté chez le parent	Marqueur	Parent	position (cM)	nbr d'identification (sur 30)
8	SxG	21,2-56,3	8,5	femelle	-	-	-	-
17	SxG	9,3-20,4	17,1	male	M_111	male	13,7	29
13	S	0,0-13,3	5,9	femelle	F_80	femelle	21,9	2
18	SxG	32,6-43,3	6,6	male	M_118	male	41,1	1

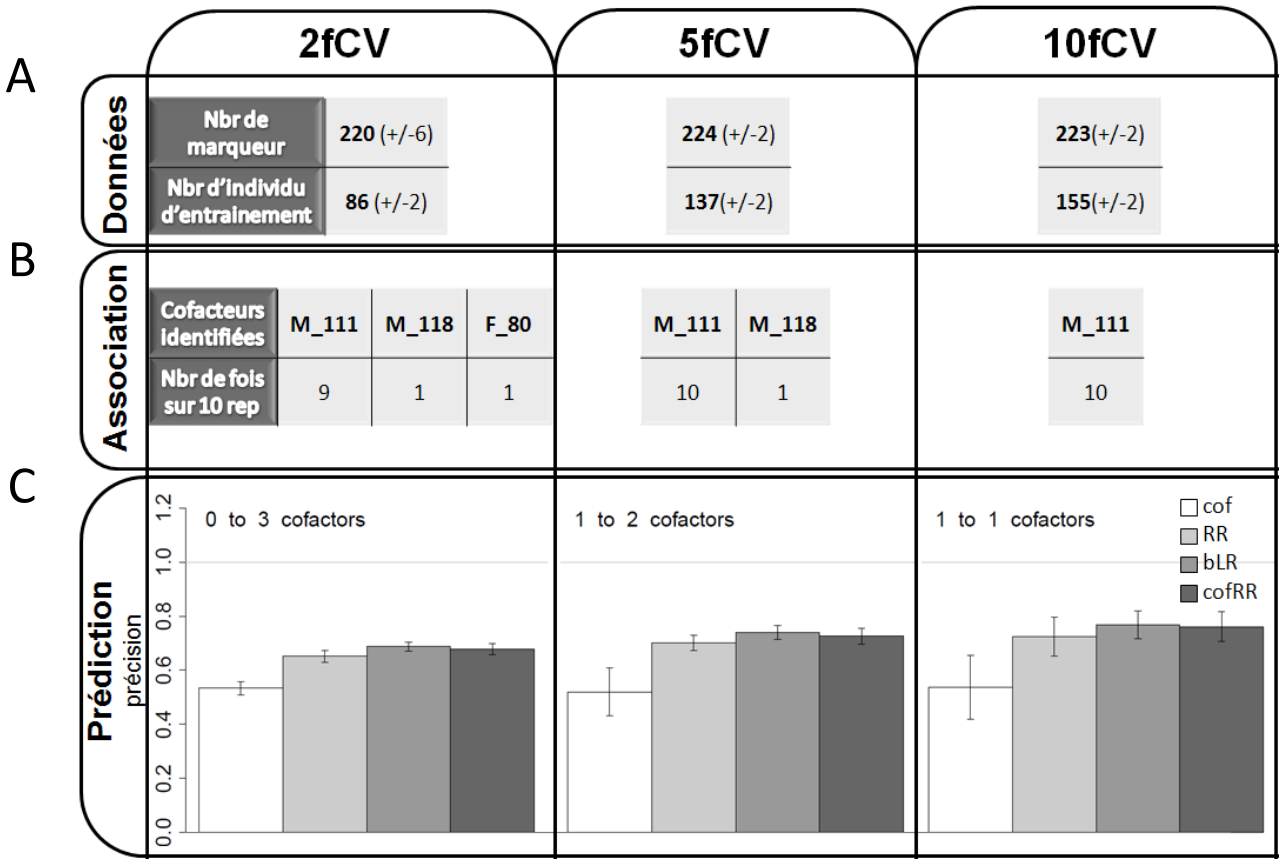


Figure 4.1 : Prédictions obtenues en cross-validations sur la descendance Syrah x Grenache. Cette figure présente les données et les résultats obtenus dans les trois types de cross-validation : 2fCV, 5fCV et 10fCV. La figure A présente le nombre d'individus dans la population d'entrainement et le nombre de marqueurs utilisés dans l'analyses ; la figure B présente les marqueurs identifiés dans l'étude de génétique d'association réalisé par mlmm (SEGURA *et al.* 2012) et le nombre de fois où ils étaient significativement identifiés sur les 10 répétitions de l'analyse. La figure C présente la précision de la prédiction moyenne et avec les quatre méthodes testées : cof, Ridge-Regression BLUP, Bayesien LASSO et cofRR (les barres d'erreurs ont été capsulés avec un intervalle de confiance de 95% autour de la moyenne).

autres méthodes de prédiction de cette étude utilisent l'information génétique issue de la totalité des marqueurs. Le dernier modèle utilisé appelé « cofRR » combine les deux types d'approches : pour les marqueurs sélectionnés en mlmm, il utilise les valeurs estimé en cof, et pour le reste des marqueurs il estime les effets par « Ridge-Regression BLUP ».

La cross-validation a été réalisée avec trois nombres d'individus différents dans la population d'entrainement. i. Pour 10fCV (10-fold Cross-Validation), la descendance SxG a été découpée en 10 parts équivalentes de façon aléatoire. Sur chacune des dix parties une prédiction a été réalisée avec comme population d'entrainement les 9 autres parties. ii. Pour 5fCV, 1/5 des individus de la population de validation pour 4/5 individus dans la population d'entrainement et iii. pour 2fCV, la moitié des individus de la population de validation pour une moitié des individus dans la population d'entrainement. Afin d'éviter un biais d'échantillonnage les niveaux 5fCV et 2fCV ont été répétés respectivement 2 et 5 fois.

Pour comparer les prédictions entre nos conditions nous avons calculé la coefficient de corrélation de Pearson (r) entre les valeurs estimées et prédites.

3. Résultats et discussion

Sur les 254 loci observés en moyenne 220 à 224 ont passé les filtres définis plus haut. Dans les trois types de cross-validations les populations d'entrainements sont composées de 86 (pour 2fCV), 137 (pour 5fCV) et de 155 individus (pour 10fCV) en moyenne (Figure 4.1A). Les populations candidates contiennent donc en moyenne 86 (2fCV), 35 (5fCV) et 17 (10fCV) individus. Nous avons effectué 10 prédictions pour chaque condition, ce qui a pour conséquence que certaines conditions ont été répétées. Nous avons effectué une seule répétition de 10fCV, deux de 5fCV et 5 de 2fCV.

3.1. Génétique d'association

Les tests d'associations réalisés avec la méthode « cof », ont identifiés de 0 à 3 cofacteurs avec les différentes populations d'entraînements (Figure 4.1B, Tableau 4.1) : le marqueur M_111, cartographié sur le groupe de liaison (GL) 17 en position 13,7 cM sur la carte Grenache qui a été identifié pour chaque analyse, le marqueur M_118, détecté pour une répétition de 2fCV et 5fCV, cartographié sur le GL 18 en position 41,1 cM sur la carte Grenache et le marqueur F_80 détecté une seul fois sur une répétition de2fCV et qui se trouve sur le GL 13 de la carte Syrah. Par ailleurs, nous avons constaté que le nombre d'associations identifiées diminuait lorsque la taille des échantillons augmentait : en utilisant le maximum d'individus (155) un seul marqueur (M_111) était identifié à chaque répétition (avec un seuil de Bonferroni à 5%).

DOLIGEZ *et al*. (soumis) en utilisant le même jeux de données avaient cartographiés trois QTL pour la taille de la baie sur la carte consensus sur les groupes de liaisons 8, 17 et 18 ainsi qu'un QTL supplémentaire sur le GL 13 (Tableau 4.1). Les marqueurs M_111 et M_118 co-localisent avec un des QTLs cartographié par DOLIGEZ *et al*. (soumis) (Tableau 4.1). Par contre le marqueur F_80 ne semble pas co-localiser avec le QTL sur le GL 13. De plus, nous n'avons pas confirmé la détection du QTL sur le GL 8.

3.2. Prédiction des phénotypes

Lorsque les marqueurs identifiés dans l'analyse mlmm ont été utilisés pour prédire le phénotype de la population candidate par la methode « cof », la précision moyenne de la prédiction phénotypique varie entre 0,52 et 0,78 (Figure 4.1C). Les méthodes utilisant tous les marqueurs donnent des précisions très similaires les unes des autres mais sont toujours plus performantes que celle passant par une sélection de cofacteurs (cof méthode)

En utilisant les autres modèles, la meilleure précision de prédiction obtenue est de 0,93 dans le contexte 10fCV avec le modèle issu de la méthode « cofRR». Alors que dans le contexte 5fCV la meilleure précision était de 0,8, et obtenue avec la méthode « BLR ». Finalement dans le contexte 2fCV, la meilleure précision était de 0,75 et obtenue avec la méthode « cofRR ». Ces résultats montrent que la taille de la population d'entrainement a un impact sur la prédiction, une plus grande population d'entrainement permet des prédiction plus précises, ce qui est en accord avec la bibliographie (BERNARDO and YU 2007; WONG and BERNARDO 2008; LORENZANA and BERNARDO 2009). Le nombre idéal d'individus d'entrainement, ainsi que le nombre de marqueurs nécessaires pour capturer la variance génétique présente dans le matériel observé, dépendent aussi du nombre de fragments chromosomiques indépendants (GODDARD 2008; HAYES *et al.* 2009c). Il y a moins de segments indépendants quand la taille efficace du matériel observé est faible. Dans ce cas, l'étude nécessite moins de marqueurs et aussi moins d'individu dans la population d'entrainement. La population Syrah x Grenache représente une diversité génétique beaucoup plus restreinte que le panel de diversité étudié dans le chapitre 3. Vu que cette population est i. issue de seulement deux parents donc présentant un maximum de 4 allèles par locus, ii. contient uniquement des pleins frères donc les individus de la population d'entrainement et de la population de validation sont de la même génération, et donc présentent un nombre de recombinaison, très faible entre eux. Dans ces conditions 220 marqueurs ont été suffisants pour obtenir des précisions plus élevées que ceux obtenus avec les panels plus diversifiés, même avec seulement 86 individus dans la population d'entrainement. En utilisant la méthode « cof », le nombre de marqueurs nécessaire pour prédire les phénotypes avec une précision de 0,5 est même seulement de 1 à 3 marqueurs.

La variation entre les précisions obtenues dans les 10 répétitions était la plus forte en 10fCV et la plus faible en 2fCV, alors que 5fCV avait une position intermédiaire. Ces résultats nous rappellent que pour obtenir des prédictions précises, la population d'entrainement doit bien représenter la diversité génétique de la population candidate (DAETWYLER *et al.* 2012). C'est-à-dire la taille de la population d'entrainement n'est pas un critère suffisant, sa composition est aussi très importante. En effet, plusieurs études ont été réalisées pour comprendre quels étaient les facteurs qui influençaient le plus la précision de la prédiction ; un des facteurs les plus importants semble être l'apparentement entre les individus des deux populations (HABIER *et al.* 2007, 2010; DE ROOS *et al.* 2009; CLARK *et al.* 2012). Des concepts et des méthodes ont été proposées pour optimiser la composition de la population d'entrainement en prenant en compte ces informations (ALBRECHT *et al.* 2011; PSZCZOLA *et al.* 2012; RINCENT *et al.* 2012).

En conclusion, dans ce chapitre nous avons montré, que dans une population biparentale de vigne la prédiction du phénotype peut être réalisée avec une précision élevée (corrélation moyenne entre 0,65 et 0,78) en utilisant un ensemble de 110 SSR marqueurs à l'aide des méthodes de sélection génomique. Cependant, en utilisant la méthode « cof » seule nous obtenons des précisions proches de 0,5 en n'utilisant qu'un à trois marqueurs. L'efficacité des méthodes utilisant tous les marqueurs pourrait être améliorée en optimisant la population d'entrainement sur les informations de l'apparentement entre individus (RINCENT *et al.* 2012). Cette optimisation pourrait également déterminer le nombre d'individu d'entrainement recommandé. Une autre sélection possible pourrait être une approche de « BIN mapping » (VISION *et al.* 2000) permettant de maximiser le nombre de recombinaisons parmi les individus sélectionnés pour la population d'entrainement.

L'amélioration de la prédiction par rapport à celle obtenue en utilisant des échantillons diversifiés (panel de diversité) est substantielle puisque le r obtenu dans les meilleures conditions avec l'échantillon diversifié était de 0.4, mais un tel schéma va nécessiter la réalisation de formules de prédictions pour chaque nouveau croisement.

CHAPITRE 5 : DISCUSSION ET PERSPECTIVES

1. Discussion

1.1. Contexte de l'étude

L'ambition de cette thèse a été de proposer une méthodologie, incluant les dernières connaissances et les derniers outils de la recherche, pour développer la création variétale chez la vigne. Comme nous l'avons déjà évoqué la viticulture française comme d'autres filières agricoles françaises doit faire face à 3 grands défis:

- ✓ la réduction des intrants phytosanitaires (plan Ecophyto 2018),
- ✓ les changements climatiques,
- ✓ et de nouveaux concurrents sur le marché, notamment les pays du nouveau monde.

L'organisation actuelle de la viticulture ne laisse que peu de place à la création variétale ; en effet les AOCs imposent un unique lieu de production mais aussi une liste de cépages pour produire un vin d'appellation. Enfin, le consommateur averti est sensibilisé aux cépages dans les vins. Il n'en demeure pas moins que la création variétale peut-être une des solutions pour répondre aux différents défis, ce qui rend les conditions dans un futur très proche particulièrement favorables. Toutes les régions, en AOC comme en non AOC sont de plus en plus réceptives.

Par ailleurs, la vigne dispose de nombreux outils puissants permettant d'envisager la mise en place des méthodologies d'amélioration les plus innovantes :

- ✓ Des ressources génétiques très importantes sont disponibles pour la vigne (*Vitis vinifera*) liées à la forte diversité de l'espèce. Dans le cadre de ce projet nous avons eu la possibilité de travailler avec la plus grande collection au monde de ressources génétiques disponible de l'espèce *Vitis vinifera* L. Cette collection est localisée au domaine INRA de Vassal. Un sous ensemble de cette collection comprenant 279 individus représentatifs de la diversité des trois pools génétiques de l'espèce (NEGRUL 1946; LEVADOUX 1956; BACILIERI *et al.* 2013; EMANUELLI *et al.* 2013) avait par ailleurs été défini (Nicolas et al, *in prep*).
- ✓ La séquence complète du génome (JAILLON *et al.* 2007; VELASCO *et al.* 2007) ; version actuelle avec une couverture de 12X) qui facilite grandement la mise en place de méthodologies haut débit d'identification de marqueurs. Notamment grâce à l'utilisation des nouvelles technologies de séquençage et d'approche du type « Genotyping by sequencing » (GBS) qui permettent de découvrir de nombreux polymorphismes et de diminuer le prix du génotypage (MYLES *et al.* 2010; DAVEY *et al.* 2011). Cette technique a été utilisée sur un

échantillon de 32 cépages diversifiés dont une grande partie appartient à la core-collection développée par LE CUNFF *et al.* (2008) afin de développer une puce de génotypage de 18 K SNP (LE PASLIER *et al.* 2013) ;
http://urgi.versailles.inra.fr/Species/Vitis/GrapeReSeq_Illumina_20K)

Cependant, l'amélioration de la vigne présente aussi des difficultés, liée aux caractéristiques de l'espèce.

- ✓ Plusieurs études menées sur le déséquilibre de liaisons chez la vigne ont détecté un niveau de DL très faible (BARNAUD *et al.* 2006; LIJAVETZKY *et al.* 2007; THIS *et al.* 2007; MYLES *et al.* 2010). Selon la littérature le DL chute lorsque la taille du fragment est supérieure à 10Kb (le seuil de r^2 choisi est généralement 0,2). Des données récentes (NICOLAS *et al., in prep*) obtenues dans le cadre du projet ANR DL-Vitis, montrent que le DL dans 4 régions de 1Mb chacune varie fortement. Un r^2 de 0,2 est maintenu sur une distance variant de 7 à 70 kb selon les régions, le DL se maintenant sur une distance plus forte dans une zone liée à la taille de la baie, donc une zone qui aurait subi plus fortement la sélection.
- ✓ Les cycles de reproduction chez la vigne sont relativement long (> 3 ans). Afin d'accélérer les croisements et de raccourcir le temps d'un programme de sélection notamment pour le pyramidage de gènes de résistance, nous avons initié l'utilisation d'un génotype particulier appelé « Dwarf ». Ce génotype mutant « semi-naturel » est le résultat d'une mutation spontanée apparue dans une des couches cellulaires du Pinot Meunier le rendant insensible aux gibbérellines (BOSS and THOMAS 2002). Cette insensibilité a pour conséquence de réduire le temps de mise à fruit des génotypes porteurs de cette mutation à 9-10 mois et de permettre une floraison continue, non inféodée aux saisons (CHAÏB *et al.* 2010). Ce génotype présente une difficulté majeure notamment pour le phénotypage de caractères de production et de qualité, dans la mesure où les grappes et dans un moindre mesure les baies ne correspondent en rien à celles des vignes au vignoble.

  Enfin d'autres caractéristiques de l'espèce impactent directement les programmes de création variétale : C'est une espèce très hétérozygote, sans population de pré-breeding, ni de lignées élites

Les outils de génomique disponibles sur la vigne ainsi que les difficultés pour la création variétale des espèces pérennes sont des facteurs favorables pour la mise en place d'un programme de sélection génomique.

1.2. Simulation

Dans le cadre de la thèse, nous avons ainsi cherché à estimer l'intérêt et l'efficacité de la sélection génomique (ou prédiction phénotypique) chez la vigne. A notre connaissance, aucune étude de prédiction phénotypique sur cette espèce n'a encore été publiée et encore peu chez des espèces présentant des caractéristiques proches (KUMAR *et al.* 2012) .

Compte tenu de l'absence de données chez la vigne et des difficultés liées aux caractéristiques de la plante, la première étape indispensable a été d'estimer la faisabilité de la méthode chez la vigne en estimant notamment l'effet du nombre de marqueurs, du nombre d'individus dans la population d'entrainement, de son éloignement avec la population de validation pour obtenir un bon niveau de corrélation entre les prédictions et les phénotypes réels. Le principal objectif était d'estimer si un panel de diversité défini pour des études de génétique d'association peut aussi être utilisé comme population d'entrainement pour définir les paramètres des modèles de sélection génomique. Cette question peut aussi être formulée comme suit : peut-on envisager qu'un panel de diversité puisse servir de population d'entrainement « universelle » à l'ensemble des croisements réalisables entre deux cépages de *Vitis vinifera* L. quelque soit leur origine génétique?

Pour cela nous avons choisi, dans un premier temps, de travailler sur des données simulées qui traduisent « l'évolution de l'espèce *Vitis vinifera* L. » ou la construction démographique des trois pools génétiques de cette espèce (NEGRUL 1946; LEVADOUX 1956; BACILIERI *et al.* 2013; EMANUELLI *et al.* 2013). Afin de simuler ce type de données, nous avons utilisé, en collaboration avec Samuel Neuenschwander, une nouvelle version du logiciel quantiNEMO (NEUENSCHWANDER *et al.* 2008) qui permet, via des approches forward, de simuler des scénarii démographiques et d'obtenir à chaque génération des fichiers contenant le génotype et le phénotype de chaque individu. Ce logiciel très bien classé dans une étude comparant plusieurs logiciels équivalents (HOBAN *et al.* 2012) donne accès à un grand choix de paramètres. Il permet ainsi de :

- ✓ choisir le nombre et le type de marqueurs (bi-allèliques ou multi-allèliques),
- ✓ choisir le nombre de QTLs et leurs effets,
- ✓ choisir la variance de la résiduelle et donc indirectement l'héritabilité des caractères étudiés,
- ✓ appliquer des sélections sur les caractères phénotypiques en choisissant soit un optimum à atteindre, soit un sens (vers de grandes ou de petites valeurs pour ce caractère),
- ✓ travailler avec plusieurs populations pouvant échanger des migrants.

Dans cette étude nous avons choisi de ne faire varier que quelques paramètres. En effet le nombre de marqueurs choisis et l'héritabilité des caractères sont quasiment les mêmes pour chaque

condition. Le nombre de marqueurs choisis (~100 000 SNPs) correspond en moyenne à deux marqueurs tous les blocks de DL d'après les estimations de MYLES *et al.* (2010) et NICOLAS *et al.* (*in prep*.).

L'utilisation de ce logiciel nous a permis d'obtenir un échantillon dont les paramètres de diversité sont proches de ceux connus pour la vigne. La possibilité de réaliser une sélection différentielle entre les populations simulées est très intéressante car cela permet de créer un échantillon structuré et donne accès à des caractères phénotypiques liés à la structure des populations. Les caractères structurés sont difficiles à travailler en génétique d'association parce que la variabilité du caractère est expliquée en grande partie (ou complètement dans le pire des cas) par le cofacteur « Structure ». Pour ce type de caractères, l'identification de marqueurs causaux ne peut donc se faire que dans des sous populations non structurés et ces marqueurs peuvent ne pas être pertinent dans une autre sous population.

Avec les données simulées, sans doute parce que les échantillons étaient très importants (1000 individus pour chaque sous-population) nous n'avons pas observé cette tendance, plus de marqueurs ayant été identifiés dans la population globale que dans chaque sous population. Par contre, cela a bien été le cas de l'analyse du chapitre 3 : aucune association n'ayant été identifiée avec l'échantillon global.

Les résultats de la simulation montrent que, lorsque le nombre de marqueurs et d'individus peuvent assurer une bonne puissance statistique à la GWAS, il est préférable de prendre en compte les associations identifié en mlmm avec leurs effet estimé par une modèle mixte (avec l'apparentement en effet aléatoire), et estimer l'effet des marqueurs restant dans un modèle de GS (dans nos études la Ridge Regression).

Les résultats montrent aussi que les modèles de sélection génomique captent l'effet de la structure présent dans la population d'entrainement. En effet dans le cas d'un caractère structuré, il n'y a pas que les polymorphismes causaux qui montrent une co-variation avec le phénotype, mais également les marqueurs historiquement liés. Le même type d'observation a été fait sur l'apparentement par HABIER *et al.* (2007, 2010). Intuitivement, il semble que les marqueurs expliquant la différenciation entre individus et non liés fonctionnellement aux caractères soient nombreux et avec des effets faibles sur la variation de ces derniers. En absence de gène majeur, ces conditions correspondent parfaitement à l'hypothèse de la « Ridge Regression » qui ne permet pas d'effet extrême, mais des effets faibles et moyens.

Pratiquement ce résultat indique qu'il est donc possible de prédire des caractères pour lesquels l'identification de gènes fonctionnellement impliqués dans la variation observée dans l'ensemble de la diversité est très difficile voir impossible. Ces caractères sont importants pour la vigne, nous pouvons notamment citer la taille de la baie, l'architecture des grappes ou la forme de la feuille, mais vraisemblablement aussi des caractères adaptatifs aux différents environnements. Ces derniers sont des cibles très importantes pour répondre à un des défis de la création variétale que sont les changements climatiques.

Avec des échantillons de grande taille, un génotypage relativement dense, et des phénotypes simulés donc par nature de bonne qualité, nous avons donc montré qu'il était possible de prédire le phénotypes d'individus issue d'une descendance utilisant un modèles de prédictions entrainé sur une collection de diversité large, avec des qualités de prédiction (accuracy) très forte de l'ordre de 0.9. Dans la « vrai vie », nous n'avons cependant pas accès ni à un nombre d'individus aussi important, ni à un nombre de marqueurs permettant une couverture aussi dense du génome et nous disposons des données phénotypiques au vignoble, certes avec des répétitions mais tout de même liées aux aléas de l'environnement au sens large. Il nous a donc semblé important de valider les données de simulation par un test avec des données réelles.

1.3. Données réelles

Ces travaux sur les données réelles ont suivies deux axes différents :

Le premier tentait également de répondre à la question : peut-on envisager qu'un panel de diversité puisse servir de population d'entrainement « universelle » à l'ensemble des croisements qui peuvent être envisagés entre deux cépages de *Vitis vinifera* L. quelque soit leur origine génétique? Pour cet axe nous avons travaillé avec des données issues du vignoble, obtenues sur un panel de diversité de 279 individus composés de trois sous populations de 93 individus chacune, choisis pour représenter la diversité de chacune de ces sous populations (NICOLAS *et al*. *in prep*). Ce panel nous a permis de définir quatre populations d'entrainement (l'ensemble du panel et les trois sous population utilisées séparément). De plus des individus issus d'un croisement Syrah X Grenache plantés sur la même parcelle ont été utilisés comme population de validation. Cette étude utilise des données de génotypage, obtenues dans le cadre du projet GrapeReSeq, issue de l'exploitation d'une puce Illumina 18 K SNP. Cet axe montre très clairement que cette option est assez bonne puisqu'elle conduit à l'identification de marqueurs pour 2 des 4 caractères par GWAs et à des coefficent de corrélation de Pearson entre les valeurs prédites et les phénotypes de l'ordre de 0.4 dans les meilleures conditions. Le fait que nous obtenions des résultats en association avec un nombre assez faible de marqueurs (moins de 9 000) alors que les estimations sur le DL montrent que 10 fois plus de

marqueurs seraient nécessaires, s'expliquent par la nature des caractères et par le fait qu'ils aient probablement été sélectionnées fortement par l'homme conduisant à des zones où le DL est plus important (FOURNIER-LEVEL *et al.* 2009; MYLES *et al.* 2011). L'efficacité de la sélection génomique est par contre identique entre les caractères, démontrant que le DL n'est pas le seul en cause. Il n'en demeure pas moins qu'il a fallu tester toutes les hypothèses possibles concernant l'effet (additifs ou dominants) des marqueurs ainsi que toutes les populations d'entrainement pour identifier les meilleurs combinaisons et éviter d'obtenir des corrélations négatives. Il est également fort probable que le nombre de marqueurs doive être augmenté pour les deux méthodes (GWAS et GS) pour pouvoir utiliser l'ensemble des 279 individus. En effet des associations ont été trouvées uniquement pour deux caractères en étudiant les sous populations mais rien avec les 279 individus.

En GS l'augmentation du nombre d'individus d'entrainement dans cette étude était directement liée à une augmentation de la diversité observée. Dans ce contexte, l'entrainement du modèle de prédiction sur plus d'individu implique la caractérisation d'une plus large diversité d'effets, au lieu de fournir des échantillons génétiquement plus similaires qui pourraient renforcer la puissance de l'estimation des effets déjà représentées. En revanche, en augmentant le nombre de marqueurs il semble que les effets de chaque fragment chromosomique « tagé » par un marqueur pourraient être mieux cernés, permettant la prise en compte plus précise des effets de polymorphismes causaux et aussi la structure et l'apparentement, comme observé sur notre étude de simulation. Dans une étude mené sur le dispositif US Holsteins, (VAZQUEZ *et al.* 2010) démontrent qu'augmenter le nombre de SNP de 5 000 à 10 000 n'améliore pas de façon significative la prédiction. Cependant, des études de prédiction génomique sur la taille des humaines (MAKOWSKY *et al.* 2011) ont donné plus d'importance à l'augmentation du nombre de marqueurs de 10 000 à 100 000 SNP, ce qui correspond plus à nos observations.

Cependant les prédictions obtenues en utilisant les 3 sous populations donnent des résultats différents et ce avec des codages différents (effet additif ou effet dominant). Ceci semble indiquer que les architectures génétiques subjacentes sont différentes d'un caractère à l'autre mais aussi entre les différentes sous populations.

Contrairement aux résultats des études de simulations, qui démontraient l'intérêt d'utiliser des méthodes de prédictions différentes pour des caractères aux architectures génétiques différents, la majorité des études faites sur les données réelles montrent que le choix de méthode a peu d'influence sur la précision de la prédiction, comme cela a été le cas dans notre étude. Les possibles explications de cette observation – discutées en détail dans DE LOS CAMPOS *et al.* (2012) – pourraient être i. l'architecture du caractère qui n'est pas si extrême en réalité que dans des jeux simulées ; ii.

les conditions des études réelles en terme de nombre de marqueurs et nombre d'individus dans la population d'entrainement qui ne permettent pas aux méthodes de réaliser leur performance théorique. Ce deuxième point affecte surtout les méthodes qui prennent en compte des effets forts comme Bayes B et C. En effet, attribuer un effet fort a un impact fort sur la prédiction et engendre un grand risque de mauvaise prédiction. Si l'effet est pertinent, la prédiction peut être significativement améliorée, mais dans le cas inverse le biais introduit peut faire chuter la précision. Pour bien profiter d'une telle méthode, il faut un DL fort entre le marqueur et le polymorphisme causal, et que le fragment chromosomique « tagé » soit assez court, pour éviter d'avoir d'autres effets à estimer par dessus. Sur notre matériel, certains marqueurs identifiés en GWAS explique une grande part de la variation (jusqu'à 28% pour la longueur de la grappe) et plusieurs peuvent expliquer des parts de variation importantes (13 marqueur expliquant plus de 10% de variations pour la taille de la baie et la longueur de la grappe). Dans le panel de diversité la taille du DL est faible, donc les fragments chromosomiques sont courts. Si nous arrivons à augmenter la densité de marquage, nous pourrons obtenir suffisamment de marqueurs en DL fort avec les polymorphismes causaux. Ces conditions semblent satisfaire les critères pour une utilisation efficiente des méthodes qui prennent en compte des effets extrêmes (comme Bayes A, B, C). Chez des plantes comme la vigne – qui se caractérise par une forte hétérozygotie, l'absence de lignées élites, une large diversité et l'existence de pools génétiques différenciés, structurés – en fonction du matériel observé et des architectures génétiques pressenties du caractère à prédire, ces modèles seraient à envisager.

Le second axe vise à réduire au maximum l'investissement. Compte tenu du nombre important de marqueurs nécessaires avec un échantillon de diversité comme population d'entrainement, nous avons imaginé un contexte différent. Le sélectionneur choisi deux géniteurs pour des caractéristiques simples mais cherche à l'évaluer aussi sur des phénotypes très couteux qui ne peuvent être mesurés que sur une partie des descendants. Il plante donc une partie des génotypes au vignoble, définit le modèle de prédiction sur ces individus et l'applique sur le reste des pépins issus de ce croisements (une population de plusieurs centaines d'individus) afin de ne mettre au champ que les meilleurs descendants pour une étude plus fine. Ici les résultats sont très satisfaisants en utilisant seulement quelques centaines de marqueur et pour 100 à 150 individus. Mais ces résultats ne sont valables que pour cette population et à cette génération car en s'éloignant dans le temps (futures générations) ou du pool génétique original les effets estimés des marqueurs vont perdre leur pertinence dû à la recombinaison entre marqueur et QTL causaux, ou dû aux effets nouveaux ou aux différentes caractéristiques du nouveau matériel qui fera chuter la précision de la prédiction (GODDARD 2008; MEUWISSEN 2009; DE ROOS *et al.* 2009). Une solution proposé dans plusieurs schémas impliquant la sélection génomique, pourrait être l'actualisation de la population d'entrainement incluant des

individus représentant la nouvelle configuration du DL ou la diversité ajoutée (HEFFNER *et al.* 2010; JANNINK *et al.* 2010). Mais cela nécessite le phénotypage des nouveaux individus ce qui demande du temps et des dépenses supplémentaires.

1.4. Différence entre simulation et données réelles

La grande différence entre des données issues de simulation et les données réelles est l'existence de facteurs limitant. L'expérimentation présentée au chapitre 3 en présente quelques uns. Tout d'abord le nombre de marqueurs. Même s'il n'existe pas de données plus importantes actuellement chez la vigne, le nombre de marqueurs disponibles est inférieur au besoin théorique (~100 000 pour couvrir avec au moins quelques marqueurs tous les blocs de DL). A terme, une possibilité pour augmenter le nombre de marqueurs serait de réaliser des imputations (LI *et al.* 2009). Il faudrait pour cela avoir des références plus fournies. Afin d'obtenir ce type de références nous pourrions envisager de séquencer entièrement ou partiellement plusieurs individus couvrant la diversité présente chez *Vitis vinifera* L., comme ceux proposé par LE CUNFF *et al.* (2008).

Le nombre d'individus de la population d'entrainement est lui aussi limitant. Cependant nos résultats issus de l'analyse des simulations ou des données réelles montrent que le nombre d'individus n'est pas un des critères les plus importants dans notre contexte. A nombre d'individus équivalent nous obtenons de meilleurs résultats quand les populations d'entrainement et la population de validation sont proches. Ce résultat montre que l'éloignement génétique à un impact plus fort que le nombre d'individus à nombre de marqueurs équivalent. Cependant par la simulation, lorsque le nombre d'individus et le nombre de marqueurs sont importants (1000 et 100 000), nous avons montré que l'utilisation d'une population très diversifiée (core-collection) permet l'obtention de prédictions très intéressantes sur l'ensemble des populations de validation testées et ce quelque soit l'origine génétique des parents.

1.5. Utilisation du Genotype « Dwarf »

Une des limites de la création variétale chez la vigne est son cycle de vie, en effet un cépage, en général, fleuri au plus tôt deux à trois ans après l'obtention du pépin. Cette caractéristique, commune à beaucoup d'espèces pérennes, limite le nombre de générations réalisables lors d'un schéma de sélection et donc indirectement le potentiel de la création variétale. Cependant chez la vigne a été identifié un mutant issu de culture in vitro (somaclone) de cellules de feuille de Pinot Meunier (BOSS and THOMAS 2002) (Boss et al. 2002). Ce mutant qui est insensible aux gibbérellines permet d'obtenir un cycle du pépin au pépin de l'ordre de 8 à 10 mois, il est nain et fleurit toute l'année en serre. Une mutation dominante est responsable de ce phénotype (BOSS and THOMAS 2002). Ce mutant devrait nous permettre, en théorie, de réaliser des schémas de sélection plus

complexes et de créer des géniteurs élites homozygotes pour des caractères monogéniques d'intérêt comme les résistances aux maladies. Ce génotype peut donc, en théorie, nous permettre un gain de plusieurs années sur les schémas de sélection. Cependant son phénotype empêche de sectionner visuellement un grand nombre de caractères, en dehors des baies qui suivent un développement normal (RIENTH *et al.* 2013), l'ensemble des autres caractères comme la phénologie, le rendement ou l'architecture des grappes sont affectés par cette mutation.

Dans un contexte d'utilisation du Dwarf pour accélérer la sortie de nouvelles variétés chez la vigne, nous pourrions chercher à cumuler plusieurs gènes de résistance aux maladies cryptogamique (PAUQUET *et al.* 2001; FISCHER *et al.* 2004; MARGUERIT *et al.* 2009; DI GASPERO *et al.* 2012; SCHWANDER *et al.* 2012; VENUTI *et al.* 2013) et d'autres caractères de type monogénique comme le goût muscat (EMANUELLI *et al.* 2010), ou l'apyrénie (MEJIA *et al.* 2011) ; dans ces cas la sélection assisté par marqueurs « classique » pourrait être suffisante. Cependant il convient aussi de sélectionner pour des caractères quantitatifs plus complexes comme quantité et qualité en polyphénols (HUANG *et al.* 2012) qui sont des analyses longues et couteuses ou comme le rendement (FANIZZA *et al.* 2005) qui n'est pas observable dans un fond génétique « Dwarf ».

Ce contexte a plusieurs conséquences, un grand nombre de cycles seront nécessaires pour cumuler ces gènes chez un même individu, et seules des prédictions des « Genetic Estimated Breeding Values » permettront de garder les génotypes « Dwarf » contenant les caractères ciblés.

2. Perspectives

Ces travaux ont été financés dans le cadre d'une collaboration entre l'INRA, Montpellier SupAgro et l'IFV (UMT Geno-Vigne® ; bourse CIFRE) et d'un projet CASDAR Innovation réservé au développement de la recherche finalisée et de l'innovation dans les instituts techniques agricoles. Une des sorties de ces travaux est une réflexion sur l'utilisation de méthodes et d'outils innovants pour développer la création variétale chez la vigne.

Les résultats acquis lors de cette thèse montrent qu'il est possible de sélectionner chez la vigne pour des caractères complexes en utilisant des marqueurs moléculaires alors que ces caractères n'étaient auparavant sélectionné que « classiquement » par observation des phénotypes. Même si le coût du génotypage est en constante baisse, alors que le phénotypage s'affine et de ce fait son coût augmente, l'investissement à réaliser est-il intéressant ? Afin d'apporter des éléments de réponse à cette question et suite aux résultats obtenus dans cette étude trois scénarii sont proposés. Nous tenterons de les présenter en présentant leurs avantages et leurs inconvénients.

2.1. Le développement d'une population d'entrainement universelle

Cette proposition a déjà été largement présentée dans cette thèse, elle en fut même la colonne vertébrale. Cette proposition est très séduisante. Elle permettrait de réaliser un seul investissement important rentable pour plusieurs années, voir décennies et ce quelques soit l'idéotype envisagé. Cet investissement comprendrait le génotypage avec 100 000 SNPs et le phénotypage multi-sites d'un grand nombre d'individus (jusqu'à un millier). Le cout de génotypage des populations de validation pourrait être moindre grâce à l'utilisation d'algorithmes d'imputations. Au vue de la densité du marquage moléculaire envisagée, les modèles devraient être peu affectés par le nombre de génération, qui resterait de toute façon faible en dehors du pyramidage avec les génotypes nains.

2.2. La prédiction en ségrégation

Nous avons aussi montré dans cette thèse que la prédiction de phénotype sur une population en ségrégation donnait des résultats très intéressants. Cette étude montre qu'avec un marquage simple (uniquement les cofacteurs identifiés) les prédictions étaient déjà intéressantes. Cependant les modèles développés ne sont pas adaptés à d'autres croisements et le lapse de temps entre le génotypage et le phénotypage de la population d'entrainement et l'utilisation des modèles sur une population candidate dépend du phénotype envisagée. Pour la vigne si l'on veut travailler sur des critères relatifs aux baies, il faudra attendre au moins 2 à 3 ans pour l'obtention des premières baies ; et plus longtemps si l'on veut un phénotypage au champ.

2.3. Set de géniteurs élites

Une proposition intermédiaire moins ambitieuse que la population « universelle » mais plus généralisable que « la prédiction en ségrégation » serait de sélectionner un set d'individus qui pourraient être les futurs géniteurs les plus importants pour répondre aux questions de la filière viticole. Ce set sera comme dans le cas de la population universelle, génotypé et phénotypé en multi-sites. Ici la représentation de la diversité de l'espèce n'est pas un critère de choix et de ce fait limite l'accès à de la diversité extérieure au set sélectionné. Le nombre de marqueurs à utiliser sera vraisemblablement moindre puisse le nombre efficace et la taille du déséquilibre de liaison pour la nouvelle diversité utilisé seront moins des contraintes (taille efficace plus faible et porté de DL plus long).

Une étude économique approfondie devrait cependant être réalisée, mais la nécessité de mettre au vignoble une partie de chaque descendance pour une espèce comme la vigne qui demande de la place (3000 souches à l'hectare) et pour plusieurs années va certainement excéder le coût

additionnel de génotypage de la collection diversifiée. Un avantage pourrait tout de même être trouvé si la sélection génomique permet de remplacer des phénotypages lourds et complexes comme la vinification, qui pour être réalisés en conditions favorables demandent un nombre important de répétitions.

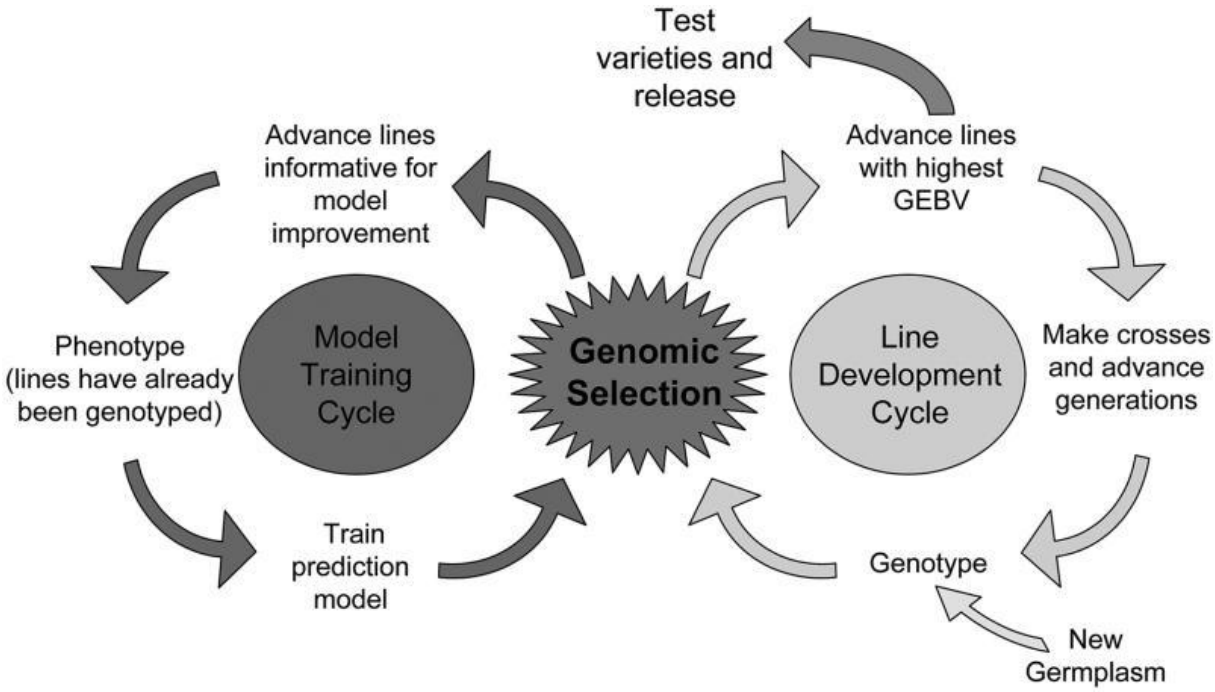


Figure 1.7. Un schéma de sélection génomique avec actualisation du modèle de prédiction (issu de HEFFNER *et al.* 2009). La partie gauche présente l'étape d'entrainement (création puis actualisation) du modèle de prédiction sur des individus génotypés et phénotypés. La partie droite présente la sélection dans le matériel végétal à améliorer, dont nous ne connaissons que les génotypes. Après l'application du modèle de prédiction nous obtenons des GEBVs. Certains génotypes vont être phénotypés (vers la partie gauche) pour actualiser le modèle statistique. Les individus avec un GEBV satisfaisant peuvent ensuite participer à des croisements ou être testés comme candidats de nouvelles variétés.

3. Conclusion

Compte tenu des outils de génotypage actuellement disponibles (puce 18 KSNP), l'option d'une « population universelle » nécessiterait un effort important pour atteindre 100 000 marqueurs moléculaires bien répartis sur le génome, ce qui permettrait d'obtenir des précisions de prédictions suffisantes. Alors que l'option « population en ségrégation » pourrait être envisagée pour un projet long, cette option couplée à l'utilisation de génotypes « Dwarf » pour un objectif précis doit tout de même être testée en augmentant le nombre de marqueurs disponibles. La profession viticole est aujourd'hui en très forte attente et le gain du temps à l'hypothèse « population universelle » est un facteur à prendre également en compte. L'hypothèse « set de géniteurs » élite nécessitera quant à elle des discussions avec la profession, pour identifier les idéotypes attendus. Il est fort probable que nous soyons capable d'identifier les génotypes de demain, mais que l'identification des génotypes « d'après demain » soit à l'heure actuelle encore trop difficile à réaliser. Des modèles de prédictions de l'évolution du climat et de l'évolution des risques sanitaires sont d'autant plus nécessaires pour réussir ce pari.

Dans le cadre de notre étude, nous n'avons pas estimé la puissance de nos modèles après plusieurs générations de croisements. Vraisemblablement d'après les études de simulation et des données réelles de maïs et de blé (HEFFNER *et al.* 2009, 2010; MEUWISSEN 2009; JANNINK *et al.* 2010), si le nombre de générations est important après le développement du modèle, il est préférable d'ajouter au set d'entrainement des individus issus de niveaux ultérieurs pour réactualiser les modèles (Figure 1.7).

Dans cette étude, nous nous sommes volontairement restreint à l'étude dans l'espèce *V. vinifera*. Même si les résistances aux maladies fongiques et des tolérances à certains stress abiotiques ne sont présentes que dans les espèces apparentées du genre *Vitis*, le standard en termes de qualité à atteindre est celui de l'espèce *V. vinifera*. Il n'en demeure pas moins, que des études avec les espèces du genre *Vitis* permettraient d'identifier des marqueurs pour tracer les zones du génome des espèces, impliquées notamment dans des caractères négatifs apportés par ces portions de génome. Le DL encore plus faible dans ce type de matériel (Nicolas et al. *in prep*) impliquerait un besoin encore plus fort en marqueurs, pour des résultats qui sont aujourd'hui encore à valider, ce type d'études interspécifiques étant très rare.

RÉFÉRENCES BIBLIOGRAPHIQUES

ADAM-BLONDON A.-F., LAHOGUE-ESNAULT F., BOUQUET A., BOURSIQUOT J.-M., THIS P., 2001 Usefullness of two SCAR markers for marker-assisted selection of seedless grapevine cultivars. Vitis - Geilweilerhof- **40**: 147–155.

ALBRECHT T., WIMMER V., AUINGER H.-J., ERBE M., KNAAK C., OUZUNOVA M., SIMIANER H., SCHÖN C.-C., 2011 Genome-based prediction of testcross values in maize. Theor. Appl. Genet. **123**: 339–350.

ARADHYA M. K., DANGL G. S., PRINS B. H., BOURSIQUOT J.-M., WALKER M. A., MEREDITH C. P., SIMON C. J., 2003 Genetic structure and differentiation in cultivated grape, Vitis vinifera L. Genet. Res. **81**: 179–192.

ARROYO-GARCIA R., RUIZ-GARCIA L., BOLLING L., OCETE R., LOPEZ M. A., ARNOLD C., ERGUL A., SÖYLEMEZO˝LU G., UZUN H. I., CABELLO F., IBAÑEZ J., ARADHYA M. K., ATANASSOV A., ATANASSOV I., BALINT S., CENIS J. L., COSTANTINI L., GORISLAVETS S., GRANDO M. S., KLEIN B. Y., MCGOVERN P. E., MERDINOGLU D., PEJIC I., PELSY F., PRIMIKIRIOS N., RISOVANNAYA V., ROUBELAKIS-ANGELAKIS K. A., SNOUSSI H., SOTIRI P., TAMHANKAR S., THIS P., TROSHIN L., MALPICA J. M., LEFORT F., MARTINEZ-ZAPATER J. M., 2006 Multiple origins of cultivated grapevine (Vitis vinifera L. ssp. sativa) based on chloroplast DNA polymorphisms. Mol. Ecol. **15**: 3707–3714.

ATWELL S., HUANG Y. S., VILHJALMSSON B. J., WILLEMS G., HORTON M., LI Y., MENG D., PLATT A., TARONE A. M., HU T. T., JIANG R., MULIYATI N. W., ZHANG X., AMER M. A., BAXTER I., BRACHI B., CHORY J., DEAN C., DEBIEU M., MEAUX J. DE, ECKER J. R., FAURE N., KNISKERN J. M., JONES J. D. G., MICHAEL T., NEMRI A., ROUX F., SALT D. E., TANG C., TODESCO M., TRAW M. B., WEIGEL D., MARJORAM P., BOREVITZ J. O., BERGELSON J., NORDBORG M., 2010 Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature **465**: 627–631.

BACILIERI R., LACOMBE T., CUNFF L. LE, VECCHI-STARAZ M. D., LAUCOU V., GENNA B., PEROS J.-P., THIS P., BOURSIQUOT J.-M., 2013 Genetic structure in cultivated grapevines is linked to geography and human selection. BMC Plant Biol. **13**: 25.

BARNAUD A., LACOMBE T., DOLIGEZ A., 2006 Linkage disequilibrium in cultivated grapevine, Vitis vinifera L. Theor. Appl. Genet. **112**: 708–716.

BATES D., MAECHLER M., 2009 lme4: Linear mixed-effects models using {S4} classes.{R} package version 0.999375-32.

BEAUMONT M. A., ZHANG W., BALDING D. J., 2002 Approximate Bayesian Computation in Population Genetics. Genetics **162**: 2025–2035.

BELLIN D., PERESSOTTI E., MERDINOGLU D., WIEDEMANN-MERDINOGLU S., ADAM-BLONDON A.-F., CIPRIANI G., MORGANTE M., TESTOLIN R., GASPERO G. D., 2009 Resistance to Plasmopara viticola in grapevine “Bianca” is controlled by a major dominant gene causing localised necrosis at the infection site. Theor. Appl. Genet. **120**: 163–176.

BERNARDO R., 2009 Genomewide Selection for Rapid Introgression of Exotic Germplasm in Maize. Crop Sci. **49**: 419.

BERNARDO R., YU J., 2007 Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. **47**: 1082–1090.

BOSS P. K., THOMAS M. R., 2002 Association of dwarfism and floral induction with a grape “green revolution” mutation. Nature **416**: 847–850.

BOUQUET A., 1982 Origine et évolution de l’encépagement francais à travers les siècles. Progrès Agric. Vitic. **5**: 110–121.

BOUQUET A., JUGA J., 2013 Integrating genomic selection into dairy cattle breeding programmes: a review. Anim. Int. J. Anim. Biosci. **7**: 705–713.

BOUQUET A., PAUQUET J., ADAM-BLONDON A.-F., ET AL., 2000 *Vers l'obtention de variétés de vigne résistantes à l'oïdium et au mildiou par les méthodes conventionnelles et biotechnologiques*.

BOURSIQUOT J.-M., LACOMBE T., LAUCOU V., JULLIARD S., PERRIN F.-X., LANIER N., LEGRAND D., MEREDITH C., THIS P., 2009 Parentage of Merlot and related winegrape cultivars of southwestern France: discovery of the missing link. Aust. J. Grape Wine Res. **15**: 144–155.

BOWERS, BOURSIQUOT, THIS, CHU, JOHANSSON, MEREDITH, 1999 Historical Genetics: The Parentage of Chardonnay, Gamay, and Other Wine Grapes of Northeastern France. Science **285**: 1562–1565.

BOWERS J. E., MEREDITH C. P., 1996 Genetic Similarities among Wine Grape Cultivars Revealed by Restriction Fragment-length Polymorphism (RFLP) Analysis. J. Am. Soc. Hortic. Sci. **121**: 620–624.

BUCKLER E. S., HOLLAND J. B., BRADBURY P. J., ACHARYA C. B., BROWN P. J., BROWNE C., ERSOZ E., FLINT-GARCIA S., GARCIA A., GLAUBITZ J. C., GOODMAN M. M., HARJES C., GUILL K., KROON D. E., LARSSON S., LEPAK N. K., LI H., MITCHELL S. E., PRESSOIR G., PEIFFER J. A., ROSAS M. O., ROCHEFORD T. R., ROMAY M. C., ROMERO S., SALVO S., VILLEDA H. S., SILVA H. S. da, SUN Q., TIAN F., UPADYAYULA N., WARE D., YATES H., YU J., ZHANG Z., KRESOVICH S., MCMULLEN M. D., 2009 The Genetic Architecture of Maize Flowering Time. Science **325**: 714–718.

CALUS M. P. L., MEUWISSEN T. H. E., ROOS A. P. W. DE, VEERKAMP R. F., 2008 Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics **178**: 553–561.

CAMPOS G. DE LOS, HICKEY J. M., PONG-WONG R., DAETWYLER H. D., CALUS M. P. L., 2012a Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics **193**: 327–345.

CAMPOS G. DE LOS, HICKEY J. M., PONG-WONG R., DAETWYLER H. D., CALUS M. P. L., 2012b Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics **193**: 327–345.

CARDON L. R., PALMER L. J., 2003 Population stratification and spurious allelic association. Lancet **361**: 598–604.

CARRIER G., CUNFF L. LE, DEREEPER A., LEGRAND D., SABOT F., BOUCHEZ O., AUDEGUIN L., BOURSIQUOT J.-M., THIS P., 2012 Transposable elements are a major cause of somatic polymorphism in Vitis vinifera L. PloS One **7**: e32973.

CARRIER G., HUANG Y.-F., CUNFF L. LE, FOURNIER-LEVEL A., VIALET S., SOUQUET J.-M., CHEYNIER V., TERRIER N., THIS P., 2013 Selection of candidate genes for grape proanthocyanidin pathway by an integrative approach. Plant Physiol. Biochem.

CATALOGUE DES VARIÉTÉS ET CLONES DE VIGNE, MINISTÈRE DE L'AGRICULTURE de la pêche et de l'alimentation, COMITÉ TECHNIQUE PERMANENT DE LA SÉLECTION, INSTITUT FRANÇAIS DE LA VIGNE ET DU VIN, 2007 *Catalogue des variétés et clones de vigne cultivés en France*. Institut français de la vigne et du Vin, Montpelier.

CAVANAGH C., MORELL M., MACKAY I., POWELL W., 2008 From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr. Opin. Plant Biol. **11**: 215–221.

CHAÏB J., TORREGROSA L., MACKENZIE D., CORENA P., BOUQUET A., THOMAS M. R., 2010 The grape microvine - a model system for rapid forward and reverse genetics of grapevines. Plant J. Cell Mol. Biol. **62**: 1083–1092.

CHANFREAU S., ROUSSEAU J., 2013 De nouvelles sources génétiques de résistance, pour plus de durabilité. In: *Les cépages résistants aux maladies chyptogamiques: Panorama européen*, eds. J. Rousseau et S. Chanfreau, Lattes, France, p. Guide Technique: 17–26.

CHEN J., CHEN Z., 2008 Extended Bayesian information criteria for model selection with large model spaces. Biometrika **95**: 759–771.

CLARK S. A., HICKEY J. M., DAETWYLER H. D., WERF J. H. VAN DER, 2012 The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. **44**: 4.

COLEMAN C., COPETTI D., CIPRIANI G., HOFFMANN S., KOZMA P., KOVACS L., MORGANTE M., TESTOLIN R., GASPERO G. DI, 2009 The powdery mildew resistance gene REN1 co-segregates with an NBS-LRR gene cluster in two Central Asian grapevines. BMC Genet. **10**: 89.

COLLARD B. C. ., MACKILL D. J., 2008 Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. B Biol. Sci. **363**: 557–572.

CROSSA J., CAMPOS G. DE LOS, PEREZ P., GIANOLA D., BURGUEÑO J., ARAUS J. L., MAKUMBI D., SINGH R. P., DREISIGACKER S., YAN J., OTHERS, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics **186**: 713–724.

CUNFF L. LE, FOURNIER-LEVEL A., LAUCOU V., VEZZULLI S., LACOMBE T., ADAM-BLONDON A.-F., BOURSIQUOT J.-M., THIS P., 2008 Construction of nested genetic core collections to optimize the exploitation of natural diversity in Vitis vinifera L. subsp sativa. BMC Plant Biol. **8**: 31.

CUNFF L. LE, LACOMBE T., FODOR A., FARNOS M., AUDEGUIN L., THIS P., 2013 Créer les cépages de demain avec les outils d'aujourd'hui. In: *Les cépages résistants aux maladies: Panorama européen*, Rousseau J., Chanfreau S., Lattes, France:Groupe ICV., p. Guide Technique: 34–40.

DAETWYLER H. D., CALUS M. P. L., PONG-WONG R., CAMPOS G. DE LOS, HICKEY J. M., 2012 Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics **193**: 347–365.

DAETWYLER H. D., PONG-WONG R., VILLANUEVA B., WOOLLIAMS J. A., 2010 The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics **185**: 1021–1031.

DAVEY J. W., HOHENLOHE P. A., ETTER P. D., BOONE J. Q., CATCHEN J. M., BLAXTER M. L., 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. **12**: 499–510.

DENIS M., BOUVET J.-M., 2012 Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genet. Genomes **9**: 37–51.

DOLIGEZ A., ADAM-BLONDON A. F., CIPRIANI G., GASPERO G. DI, LAUCOU V., MERDINOGLU D., MEREDITH C. P., RIAZ S., ROUX C., THIS P., 2006 An integrated SSR map of grapevine based on five mapping populations. TAG Theor. Appl. Genet. Theor. Angew. Genet. **113**: 369–382.

EARL D. A., VONHOLDT B. M., 2011 STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. **4**: 359–361.

EAVES L. J., 1994 Effect of genetic architecture on the power of human linkage studies to resolve the contribution of quantitative trait loci. Heredity (Edinb) **72 (Pt 2)**: 175–192.

EDING H., MEUWISSEN T. H. E., 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J. Anim. Breed. Genet. **118**: 141–159.

ELSHIRE R. J., GLAUBITZ J. C., SUN Q., POLAND J. A., KAWAMOTO K., BUCKLER E. S., MITCHELL S. E., 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L Orban, Ed.). PLoS ONE **6**: e19379.

EMANUELLI F., BATTILANA J., COSTANTINI L., CUNFF L. L., BOURSIQUOT J.-M., THIS P., GRANDO M. S., 2010 A candidate gene association study on muscat flavor in grapevine (Vitis vinifera L.). BMC Plant Biol. **10**: 241.

EMANUELLI F., LORENZI S., GRZESKOWIAK L., CATALANO V., STEFANINI M., TROGGIO M., MYLES S., MARTINEZ-ZAPATER J. M., ZYPRIAN E., MOREIRA F. M., GRANDO M. S., 2013 Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biol. **13**: 39.

FALCONER D. S., MACKAY T. F. C., 1996 *Introduction to quantitative genetics*. Longman, Essex, England.

FANIZZA G., LAMAJ F., COSTANTINI L., CHAABANE R., GRANDO M. S., 2005 QTL analysis for fruit yield components in table grapes (Vitis vinifera). Theor. Appl. Genet. **111**: 658–664.

FISCHER B. M., SALAKHUTDINOV I., AKKURT M., EIBACH R., EDWARDS K. J., TÖPFER R., ZYPRIAN E. M., 2004 Quantitative trait locus analysis of fungal disease resistance factors on a molecular map of grapevine. TAG Theor. Appl. Genet. Theor. Angew. Genet. **108**: 501–515.

FOURNIER-LEVEL A., CUNFF L. L., GOMEZ C., DOLIGEZ A., AGEORGES A., ROUX C., BERTRAND Y., SOUQUET J.-M., CHEYNIER V., THIS P., 2009 Quantitative Genetic Bases of Anthocyanin Variation in

Grape (Vitis vinifera L. ssp. sativa) Berry: A Quantitative Trait Locus to Quantitative Trait Nucleotide Integrated Study. Genetics **183**: 1127–1139.

FOURNIER-LEVEL A., LACOMBE T., CUNFF L. LE, BOURSIQUOT J.-M., THIS P., 2010 Evolution of the VvMybA gene family, the major determinant of berry colour in cultivated grapevine (Vitis vinifera L.). Heredity **104**: 351–362.

GALET P., 1988 *Cépages et vignobles de France*. C. Déhan, Montpellier [France].

GALET P., 2000 *Dictionnaire encyclopédique des cépages*. Hachette Pratique, Paris.

GASPERO G. DI, COPETTI D., COLEMAN C., CASTELLARIN S. D., EIBACH R., KOZMA P., LACOMBE T., GAMBETTA G., ZVYAGIN A., CINDRIC P., KOVACS L., MORGANTE M., TESTOLIN R., 2012 Selective sweep at the Rpv3 locus during grapevine breeding for downy mildew resistance. Theor. Appl. Genet. **124**: 277–286.

GERBER S., CHABRIER P., KREMER A., 2003 famoz: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. Mol. Ecol. Notes **3**: 479–481.

GIANOLA D., WU X.-L., MANFREDI E., SIMIANER H., 2010 A non-parametric mixture model for genome-enabled prediction of genetic value for a quantitative trait. Genetica **138**: 959–977.

GODDARD M., 2008 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136**: 245–257.

GODDARD M. e., HAYES B. j., 2007 Genomic selection. J. Anim. Breed. Genet. **124**: 323–330.

GOUESNARD B., BATAILLON T. M., DECOUX G., ROZALE C., SCHOEN D. J., DAVID J. L., 2001 MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J. Hered. **92**: 93–94.

GOUY M., ROUSSELLE Y., BASTIANELLI D., LECOMTE P., BONNAL L., ROQUES D., EFILE J.-C., ROCHER S., DAUGROIS J., TOUBI L., NABENEZA S., HERVOUET C., TELISMART H., DENIS M., THONG-CHANE A., GLASZMANN J. C., HOARAU J.-Y., NIBOUCHE S., COSTET L., 2013 Experimental assessment of the accuracy of genomic selection in sugarcane. Theor. Appl. Genet. **126**: 2575–2586.

GRASSI F., LABRA M., IMAZIO S., SPADA A., SGORBATI S., SCIENZA A., SALA F., 2003 Evidence of a secondary grapevine domestication centre detected by SSR analysis. Theor. Appl. Genet. **107**: 1315–1320.

GRATTAPAGLIA D., RESENDE M. D. V., 2010 Genomic selection in forest tree breeding. Tree Genet. Genomes **7**: 241–255.

GRIMPLET J., HEMERT J. VAN, CARBONELL-BEJERANO P., DIAZ-RIQUELME J., DICKERSON J., FENNELL A., PEZZOTTI M., MARTINEZ-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.

GROUPE ICV, 2013 *Les cépages résistants aux maladies cryptogamiques : Panorama européen*. eds. J Rousseau et S Chanfreau, Lattes, France.

GUPTA P. K., RUSTGI S., KULWAL P. L., 2005 Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol. Biol. **57**: 461–485.

HABIER D., FERNANDO R. L., DEKKERS J. C. M., 2007 The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics **177**: 2389–2397.

HABIER D., FERNANDO R. L., DEKKERS J. C. M., 2008 The impact of genetic relationship information on genome-assisted breeding values. Genetics.

HABIER D., FERNANDO R. L., KIZILKAYA K., GARRICK D. J., 2011 Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics **12**: 186.

HABIER D., TETENS J., SEEFRIED F.-R., LICHTNER P., THALLER G., 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. **42**: 5.

HAMBLIN M. T., BUCKLER E. S., JANNINK J.-L., 2011 Population genetics of genomics-based crop improvement methods. Trends Genet. **27**: 98–106.

HANNAH L., ROEHRDANZ P. R., IKEGAMI M., SHEPARD A. V., SHAW M. R., TABOR G., ZHI L., MARQUET P. A., HIJMANS R. J., 2013 Climate change, wine, and conservation. Proc. Natl. Acad. Sci. **110**: 6907–6912.

HAYES B. J., BOWMAN P. J., CHAMBERLAIN A. J., GODDARD M. E., 2009a Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. **92**: 433–443.

HAYES B. J., BOWMAN P. J., CHAMBERLAIN A. C., VERBYLA K., GODDARD M. E., 2009b Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. **41**: 51.

HAYES B. J., COGAN N. O. I., PEMBLETON L. W., GODDARD M. E., WANG J., SPANGENBERG G. C., FORSTER J. W., 2013 Prospects for genomic selection in forage plant species (OA Rognli, Ed.). Plant Breed. **132**: 133–143.

HAYES B. J., DAETWYLER H. D., BOWMAN P., MOSER G., TIER B., CRUMP R., KHATKAR M., RAADSMA H. W., GODDARD M. E., 2009c Accuracy of genomic selection: comparing theory and results. http://www.aaabg.org/livestocklibrary/2009/hayes034.pdf.

HAYES B. J., VISSCHER P. M., GODDARD M. E., 2009d Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. **91**: 47.

HEFFNER E. L., JANNINK J.-L., IWATA H., SOUZA E., SORRELLS M. E., 2011 Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. Crop Sci. **51**: 2597.

HEFFNER E. L., LORENZ A. J., JANNINK J.-L., SORRELLS M. E., 2010 Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Sci. **50**: 1681–1690.

HEFFNER E. L., SORRELLS M. E., JANNINK J.-L., 2009 Genomic Selection for Crop Improvement. Crop Sci. **49**: 1.

HESLOT N., YANG H.-P., SORRELLS M. E., JANNINK J.-L., 2012 Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. **52**: 146.

HILL W. G., WEIR B. S., 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. **33**: 54–78.

HOBAN S., BERTORELLE G., GAGGIOTTI O. E., 2012 Computer simulations: tools for population and evolutionary genetics. Nat. Rev. Genet.

HOERL A. E., KENNARD R. W., 1970 Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics **12**: 55–67.

HOUEL C., 2011 Caractérisation de la variation phénotypique de la taille de la baie chez la vigne Vitis vinifera L. et approches de génétique d'association et de recherche de traces de sélection pour ce caractère.

HOUEL C., MARTIN-MAGNIETTE M.-L., NICOLAS S. d., LACOMBE T., CUNFF L. LE, FRANCK D., TORREGROSA L., CONEJERO G., LALET S., THIS P., ADAM-BLONDON A.-F., 2013 Genetic variability of berry size in the grapevine (Vitis vinifera L.). Aust. J. Grape Wine Res.: n/a–n/a.

HUANG Y.-F., DOLIGEZ A., FOURNIER-LEVEL A., CUNFF L. LE, BERTRAND Y., CANAGUIER A., MOREL C., MIRALLES V., VERAN F., SOUQUET J.-M., CHEYNIER V., TERRIER N., THIS P., 2012 Dissecting genetic architecture of grape proanthocyanidin composition through quantitative trait locus mapping. BMC Plant Biol. **12**: 30.

HUANG X., WEI X., SANG T., ZHAO Q., FENG Q., ZHAO Y., LI C., ZHU C., LU T., ZHANG Z., LI M., FAN D., GUO Y., WANG A., WANG L., DENG L., LI W., LU Y., WENG Q., LIU K., HUANG T., ZHOU T., JING Y., LI W., LIN Z., BUCKLER E. S., QIAN Q., ZHANG Q.-F., LI J., HAN B., 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. **42**: 961–967.

HUSMANN G., 1880 *American Grape Growing and Wine Making*. Orange Judd Company.

IBANEZ-ESCRICHE N., FERNANDO R. L., TOOSI A., DEKKERS J. C. M., 2009 Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. GSE **41**: 12.

IWATA H., JANNINK J.-L., 2011 Accuracy of Genomic Selection Prediction in Barley Breeding Programs: A Simulation Study Based On the Real Single Nucleotide Polymorphism Data of Barley Breeding Lines. Crop Sci. **51**: 1915.

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.

JANNINK J.-L., LORENZ A. J., IWATA H., 2010 Genomic selection in plant breeding: from theory to practice. Briefings Funct. Genomics **9**: 166–177.

JI W., LI Z.-Q., ZHOU Q., YAO W.-K., WANG Y.-J., 2013 Breeding new seedless grape by means of in vitro embryo rescue. Genet. Mol. Res. **12**: 859–869.

KANG H. M., SUL J. H., SERVICE S. K., ZAITLEN N. A., KONG S., FREIMER N. B., SABATTI C., ESKIN E., 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42**: 348–354.

KANG H. M., ZAITLEN N. A., WADE C. M., KIRBY A., HECKERMAN D., DALY M. J., ESKIN E., 2008 Efficient Control of Population Structure in Model Organism Association Mapping. Genetics **178**: 1709–1723.

KIZILKAYA K., FERNANDO R. L., GARRICK D. J., 2010 Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. **88**: 544–551.

KOLEDA J., 1975 Ergebnisse von Kreuzungen zw. Vitis amurensis und Vitis vinifera in der Züchtung frostwiderstandsfähiger Reben. Vitis **14** : 1-5.

KUMAR S., CHAGNÉ D., BINK M. C. A. M., VOLZ R. K., WHITWORTH C., CARLISLE C., 2012 Genomic Selection for Fruit Quality Traits in Apple (Malus×domestica Borkh.). PLoS ONE **7**: e36674.

KUMAR S., SKJÆVELAND Å., ORR R. J., ENGER P., RUDEN T., MEVIK B.-H., BURKI F., BOTNEN A., SHALCHIAN-TABRIZI K., 2009 AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. BMC Bioinformatics **10**: 357.

LACOMBE T., 2012 Contribution à l'étude de l'histoire évolutive de la vigne cultivée (Vitis vinifera L.) par l'analyse de la diversité génétique neutre et de gènes d'intérêt.

LACOMBE T., BOURSIQUOT J.-M., LAUCOU V., VECCHI-STARAZ M. DI, PEROS J.-P., THIS P., 2012 Large-scale parentage analysis in an extended set of grapevine cultivars (Vitis vinifera L.). Theor. Appl. Genet.: 1–14.

LAUCOU V., LACOMBE T., DECHESNE F., SIRET R., BRUNO J.-P., DESSUP M., DESSUP T., ORTIGOSA P., PARRA P., ROUX C., SANTONI S., VARES D., PEROS J.-P., BOURSIQUOT J.-M., THIS P., 2011 High throughput analysis of grape genetic diversity as a tool for germplasm collection management. TAG Theor. Appl. Genet. Theor. Angew. Genet. **122**: 1233–1245.

LEVADOUX L., 1956 Les populations sauvages et cultivées de Vitis vinifera L. Ann. Amélioration Plantes **1**: 59–118.

LI Y., WILLER C., SANNA S., ABECASIS G., 2009 Genotype Imputation. Annu. Rev. Genomics Hum. Genet. **10**: 387–406.

LIJAVETZKY D., CABEZAS J., IBAÑEZ A., RODRIGUEZ V., MARTINEZ-ZAPATER J. M., 2007 High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics **8**: 424.

LORENZANA R. E., BERNARDO R., 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. **120**: 151–161.

LUND M. S., SAHANA G., KONING D.-J. DE, SU G., CARLBORG Ö., 2009 Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. BMC Proc. **3**: S1.

MAKOWSKY R., PAJEWSKI N. M., KLIMENTIDIS Y. C., VAZQUEZ A. I., DUARTE C. W., ALLISON D. B., CAMPOS G. DE LOS, 2011 Beyond Missing Heritability: Prediction of Complex Traits (G Gibson, Ed.). PLoS Genet. **7**: e1002051.

MANGIN B., SIBERCHICOT A., NICOLAS S., DOLIGEZ A., THIS P., CIERCO-AYROLLES C., 2011 Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity **108**: 285–91.

MARCHINI J., CARDON L. R., PHILLIPS M. S., DONNELLY P., 2004 The effects of human population structure on large genetic association studies. Nat. Genet. **36**: 512–517.

MARGUERIT E., BOURY C., MANICKI A., DONNART M., BUTTERLIN G., NEMORIN A., WIEDEMANN-MERDINOGLU S., MERDINOGLU D., OLLAT N., DECROOCQ S., 2009 Genetic dissection of sex determinism, inflorescence morphology and downy mildew resistance in grapevine. Theor. Appl. Genet. **118**: 1261–1278.

MEJIA N., SOTO B., GUERRERO M., CASANUEVA X., HOUEL C., ÁNGELES MICCONO M. DE LOS, RAMOS R., CUNFF L. LE, BOURSIQUOT J.-M., HINRICHSEN P., ADAM-BLONDON A.-F., 2011 Molecular, genetic and transcriptional evidence for a role of VvAGL11 in stenospermocarpic seedlessness in grapevine. BMC Plant Biol. **11**: 57.

MEUWISSEN T. H., 2009 Accuracy of breeding values of “unrelated” individuals predicted by dense SNP genotyping. Genet. Sel. Evol. **41**: 35.

MEUWISSEN T. H. E., HAYES B. J., GODDARD M. E., 2001 Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics **157**: 1819–1829.

MITA S. DE, THUILLET A.-C., GAY L., AHMADI N., MANEL S., RONFORT J., VIGOUROUX Y., 2013 Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol. Ecol. **22**: 1383–1399.

MORIONDO M., JONES G. V., BOIS B., DIBARI C., FERRISE R., TROMBI G., BINDI M., 2013 Projected shifts of wine regions in response to climate change. Clim. Change **119**: 825–839.

MOSER G., TIER B., CRUMP R. E., KHATKAR M. S., RAADSMA H. W., 2009 A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. **41**: 56.

MUIR W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. Z. Für Tierzüchtung Züchtungsbiologie **124**: 342–355.

MYLES S., BOYKO A. R., OWENS C. L., BROWN P. J., GRASSI F., ARADHYA M. K., PRINS B., REYNOLDS A., CHIA J.-M., WARE D., BUSTAMANTE C. D., BUCKLER E. S., 2011 Genetic Structure and Domestication History of the Grape. Proc. Natl. Acad. Sci. **108**: 3530–3535.

MYLES S., CHIA J.-M., HURWITZ B., SIMON C., ZHONG G. Y., BUCKLER E., WARE D., 2010 Rapid Genomic Characterization of the Genus Vitis. PLoS ONE **5**: e8219.

NAKAYA A., ISOBE S. N., 2012 Will genomic selection be a practical method for plant breeding? Ann. Bot. **110**: 1303–1316.

NEGRUL A., 1938 Evolution of cultivated forms of grapes. Comptes Rendus Dokl. Acade´mie Sci. USSR **18**.

NEGRUL A. M., 1946 *Ampelography of USSR*. Frolov-Bagreev A., Moscow.

NEUENSCHWANDER S., HOSPITAL F., GUILLAUME F., GOUDET J., 2008 quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. Bioinformatics **24**: 1552–1553.

OLLAT N., FERNANDEZ L., ROMIEU C., DUCHENE E., LISSARAGUE J. R., LECOURIEUX D., AGEORGES A., KELLY M., CACHO J., RIVARS J., LAMUELA R., GOUTOULY J. P., LEEUWEN C. VAN, MARGUERIT E., PECCOUX A., BARRIEU F., LEBON E., THIS P., PELLEGRINO A., MARTINEZ-ZAPATER J. M., TORREGROSA L., 2011 Multidisciplinary research to select new cultivars adapted to climate changes. In: Asti and Alba, Italy.

PARADIS E., CLAUDE J., STRIMMER K., 2004 APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics **20**: 289–290.

PARK S. D. E., 2001 Trypanotolerance in West African Cattle and the Population Genetic Effects of Selection.

PARK T., CASELLA G., 2008 The Bayesian Lasso. J. Am. Stat. Assoc. **103**: 681–686.

PASLIER M.-C. LE, CHOISNE R., BACILIERI R., BOURSIQUOT J.-M., BRAS M., BRUNEL D., GASPERO G. DI, HAUSMANN L., LACOMBE T., LAUCOU V., LAUNAY A., MARTINEZ-ZAPATER J., MORGANTE M., RAJ P., PONNAIAH M., QUESNEVILLE H., SCALABRIN S., TORRES-PEREZ R., ADAM-BLONDON A.-F., 2013 The GrapeReSeq 18k Vitis genotyping chip. In: La Serena, Chile.

PAUL H. W., 1996 *Science, Vine and Wine in Modern France*. Cambridge University Press.

PAUQUET J., BOUQUET A., THIS P., ADAM-BLONDON A.-F., 2001 Establishment of a local map of AFLP markers around the powdery mildew resistance gene Run1 in grapevine and assessment of their usefulness for marker assisted selection. Theor. Appl. Genet. **103**: 1201–1210.

PÉREZ P., CAMPOS G. DE LOS, CROSSA J., GIANOLA D., 2010 Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. Plant Genome J. **3**: 106–116.

PÉROS J.-P., BERGER G., PORTEMONT A., BOURSIQUOT J.-M., LACOMBE T., 2011 Genetic variation and biogeography of the disjunct Vitis subg. Vitis (Vitaceae). J. Biogeogr. **38**: 471–486.

PERRIER X., JACQUEMOUD-COLLET J. P., 2006 *DARwin software*.

PINNEY T., 1989 *A history of wine in America. Vol. 1, Vol. 1,*. University of California Press, Berkeley, Calif.; London.

POLAND J. A., RIFE T. W., 2012 Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome J. **5**: 92.

PRITCHARD J. K., STEPHENS M., DONNELLY P., 2000 Inference of Population Structure Using Multilocus Genotype Data. Genetics **155**: 945–959.

PRYCE J. E., DAETWYLER H. D., 2012 Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim. Prod. Sci. **52**: 107–114.

PSZCZOLA M., STRABEL T., MULDER H. A., CALUS M. P. L., 2012 Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. **95**: 389–400.

R CORE TEAM, 2013 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

RESENDE M. F. R., MUNOZ P., RESENDE M. D. V., GARRICK D. J., FERNANDO R. L., DAVIS J. M., JOKELA E. J., MARTIN T. A., PETER G. F., KIRST M., 2012a Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics **190**: 1503–1510.

RESENDE M. D. V., RESENDE M. F. R., SANSALONI C. P., PETROLI C. D., MISSIAGGIA A. A., AGUIAR A. M., ABAD J. M., TAKAHASHI E. K., ROSADO A. M., FARIA D. A., PAPPAS G. J., KILIAN A., GRATTAPAGLIA D., 2012b Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol. **194**: 116–128.

RIAZ S., BOURSIQUOT J.-M., DANGL G. S., LACOMBE T., LAUCOU V., TENSCHER A. C., WALKER M., 2013 Identification of mildew resistance in wild and cultivated Central Asian grape germplasm. BMC Plant Biol. **13**: 149.

RIENTH M., LUCHAIRE N., CHATBANYONG R., AGORGES A., KELLY M., BRILLOUET J. M., MULLER A., PELLEGRINO A., TORREGROSA L., ROMIEU C., 2013 The microvine provides new perspectives for research on berry physiology. Cienc. E Tec. Vitivinic. J. Vitic. Enol. **28**: 412–417.

RINCENT R., LALOË D., NICOLAS S., ALTMANN T., BRUNEL D., REVILLA P., RODRIGUEZ V. M., MORENO-GONZALEZ J., MELCHINGER A., BAUER E., SCHOEN C.-C., MEYER N., GIAUFFRET C., BAULAND C., JAMIN P., LABORDE J., MONOD H., FLAMENT P., CHARCOSSET A., MOREAU L., 2012 Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics **192**: 715–728.

ROOS A. P. W. DE, HAYES B. J., GODDARD M. E., 2009 Reliability of Genomic Predictions Across Multiple Populations. Genetics **183**: 1545–1553.

ROOS A. P. W. DE, HAYES B. J., SPELMAN R. J., GODDARD M. E., 2008 Linkage Disequilibrium and Persistence of Phase in Holstein-Friesian, Jersey and Angus Cattle. Genetics **179**: 1503–1512.

ROYSTON P., 1995 Calculation of unconditional and conditional reference intervals for foetal size and growth from longitudinal measurements. Stat. Med. **14**: 1417–1436.

SCHOEN D. J., BROWN A. H., 1993 Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. U. S. A. **90**: 10623–10627.

SCHWANDER F., EIBACH R., FECHTER I., HAUSMANN L., ZYPRIAN E., TÖPFER R., 2012 Rpv10: a new locus from the Asian Vitis gene pool for pyramiding downy mildew resistance loci in grapevine. Theor. Appl. Genet. **124**: 163–176.

SEGURA V., VILHJALMSSON B. J., PLATT A., KORTE A., SEREN Ü., LONG Q., NORDBORG M., 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. **44**: 825–830.

SITZENSTOCK F., YTOURNEL F., SHARIFI A. R., CAVERO D., TÄUBERT H., PREISINGER R., SIMIANER H., 2013 Efficiency of genomic selection in an established commercial layer breeding program. Genet. Sel. Evol. **45**: 29.

SOLBERG T. R., SONESSON A. K., WOOLLIAMS J. A., MEUWISSEN T. H. E., 2008 Genomic selection using different marker types and densities. J. Anim. Sci. **86**: 2447–2454.

SPITZE K., 1993 Population structure in Daphnia obtusa: quantitative genetic and allozymic variation. Genetics **135**: 367–374.

STORLIE E., CHARMET G., 2013 Genomic Selection Accuracy using Historical Data Generated in a Wheat Breeding Program. Plant Genome **6**: 0.

SUN X., MA P., MUMM R. H., 2012 Nonparametric Method for Genomics-Based Prediction of Performance of Quantitative Traits Involving Epistasis in Plant Breeding (X Wang, Ed.). PLoS ONE **7**: e50604.

SVED J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. **2**: 125–141.

THIS P., LACOMBE T., CADLE-DAVIDSON M., OWENS C. L., 2007 Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene VvmybA1. Theor. Appl. Genet. **114**: 723–730.

THIS P., LACOMBE T., THOMAS M. R., 2006 Historical origins and genetic diversity of wine grapes. Trends Genet. **22**: 511–519.

THIS P., ZAPATER J., PEROS J.-P., LACOMBE T., 2011 Natural Variation in Vitis. In: Kole C (Ed.), *Genetics, Genomics, and Breeding of Grapes*, Science Publishers, pp. 30–67.

TIAN F., BRADBURY P. J., BROWN P. J., HUNG H., SUN Q., FLINT-GARCIA S., ROCHEFORD T. R., MCMULLEN M. D., HOLLAND J. B., BUCKLER E. S., 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. **43**: 159–162.

TOOSI A., FERNANDO R. L., DEKKERS J. C. M., 2010 Genomic selection in admixed and crossbred populations. J. Anim. Sci. **88**: 32–46.

TURNER W. R., BRADLEY B. A., ESTES L. D., HOLE D. G., OPPENHEIMER M., WILCOVE D. S., 2010 Climate change: helping nature survive the human response: Indirect impacts of climate change. Conserv. Lett. **3**: 304–312.

VANRADEN P. M., SULLIVAN P. G., 2010 International genomic evaluation methods for dairy cattle. Genet. Sel. Evol. **42**: 7.

VANRADEN P. M., TASSELL C. P. VAN, WIGGANS G. R., SONSTEGARD T. S., SCHNABEL R. D., TAYLOR J. F., SCHENKEL F. S., 2009 Invited Review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. **92**: 16–24.

VAZQUEZ A. I., ROSA G. J. M., WEIGEL K. A., CAMPOS G. DE LOS, GIANOLA D., ALLISON D. B., 2010 Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. **93**: 5942–5949.

VELASCO R., ZHARKIKH A., TROGGIO M., CARTWRIGHT D. A., CESTARO A., *et al.*, 2007 A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety (B Dilkes, Ed.). PLoS ONE **2**: e1326.

VENUTI S., COPETTI D., FORIA S., FALGINELLA L., HOFFMANN S., BELLIN D., CINDRIC P., KOZMA P., SCALABRIN S., MORGANTE M., TESTOLIN R., GASPERO G. DI, 2013 Historical Introgression of the Downy Mildew Resistance Gene Rpv12 from the Asian Species Vitis amurensis into Grapevine Varieties (JC Nelson, Ed.). PLoS ONE **8**: e61228.

VIGOUROUX Y., JAQUETH J. S., MATSUOKA Y., SMITH O. S., BEAVIS W. D., SMITH J. S. C., DOEBLEY J., 2002 Rate and Pattern of Mutation at Microsatellite Loci in Maize. Mol. Biol. Evol. **19**: 1251–1260.

VISION T. J., BROWN D. G., SHMOYS D. B., DURRETT R. T., TANKSLEY S. D., 2000 Selective mapping: a strategy for optimizing the construction of high-density linkage maps. Genetics **155**: 407–420.

WALLACE J. G., LARSSON S. J., BUCKLER E. S., 2013 Entering the second century of maize quantitative genetics. Heredity.

WANG M., JIANG N., JIA T., LEACH L., COCKRAM J., WAUGH R., RAMSAY L., THOMAS B., LUO Z., 2012 Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theor. Appl. Genet. **124**: 233–246.

WEIR B. S., COCKERHAM C. C., 1984 Estimating F-Statistics for the Analysis of Population Structure. Evolution **38**: 1358–1370.

WIENTJES Y. C. J., VEERKAMP R. F., CALUS M. P. L., 2012 The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics **193**: 621–631.

WONG C. K., BERNARDO R., 2008 Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. **116**: 815–824.

XU Y., CROUCH J. H., 2008 Marker-Assisted Selection in Plant Breeding: From Publications to Practice. Crop Sci. **48**: 391.

YU J., HOLLAND J. B., MCMULLEN M. D., BUCKLER E. S., 2008 Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics **178**: 539–551.

YU J., PRESSOIR G., BRIGGS W. H., VROH BI I., YAMASAKI M., DOEBLEY J. F., MCMULLEN M. D., GAUT B. S., NIELSEN D. M., HOLLAND J. B., KRESOVICH S., BUCKLER E. S., 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38**: 203–208.

ZHAO K., ARANZANA M. J., KIM S., LISTER C., SHINDO C., TANG C., TOOMAJIAN C., ZHENG H., DEAN C., MARJORAM P., NORDBORG M., 2007 An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet **3**: e4.

ZHONG S., DEKKERS J. C. M., FERNANDO R. L., JANNINK J.-L., 2009 Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics **182**: 355–364.

ZOHARY D., 1996 The domestication of the grapevine Vitis vinifera L. in the Near East. In: *The origins and ancient history of wine.*, McGovern PE, Fleming SJ, Katz SH, pp. 31–43.

ANNEXES

1. Annexe I

Données supplémentaires pour l'article « Genome-wide prediction methods in highly diverse and heterozygous species: proof-of-concept through simulation in grapevine »

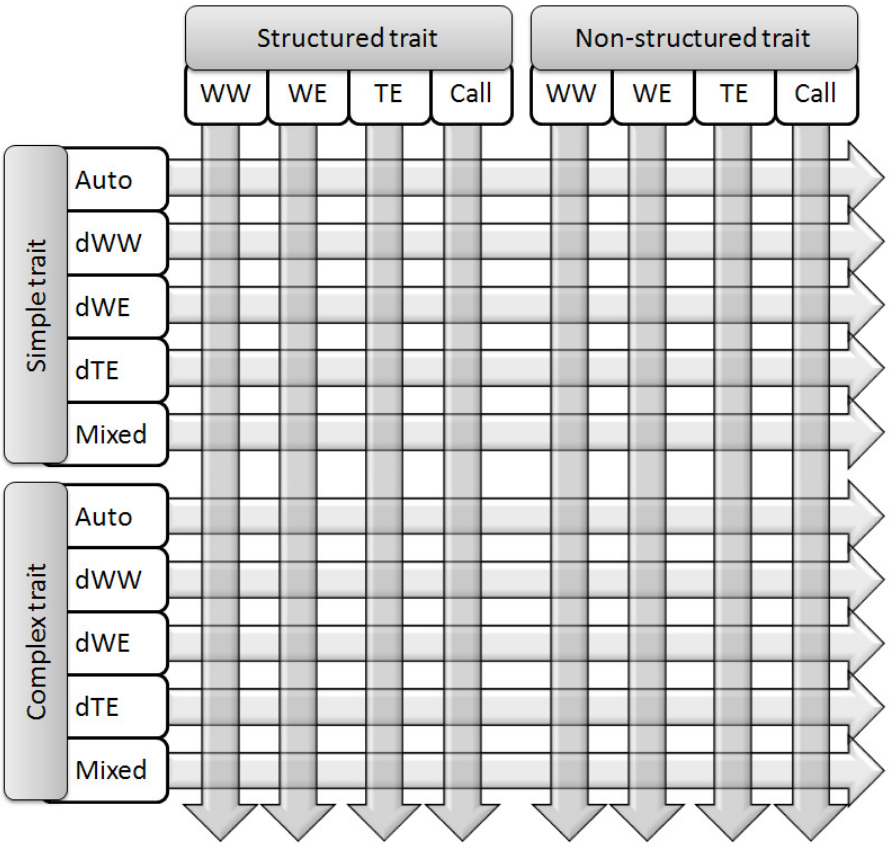


Figure 2.S1 Experimentation plan to study the evolution of the precision in different combinations of training and candidate population, at two levels of genetic architecture's complexity (10 or 100 underlying QTLs) and predicting a structured and a non-structured trait. For each possibility 4 prediction methods were used: sum of cofactors of MLMM (cof) Ridge Regression (RR), Bayesian Ridge Regression (BRR) and a marker assisted RR-BLUP (cofRR).

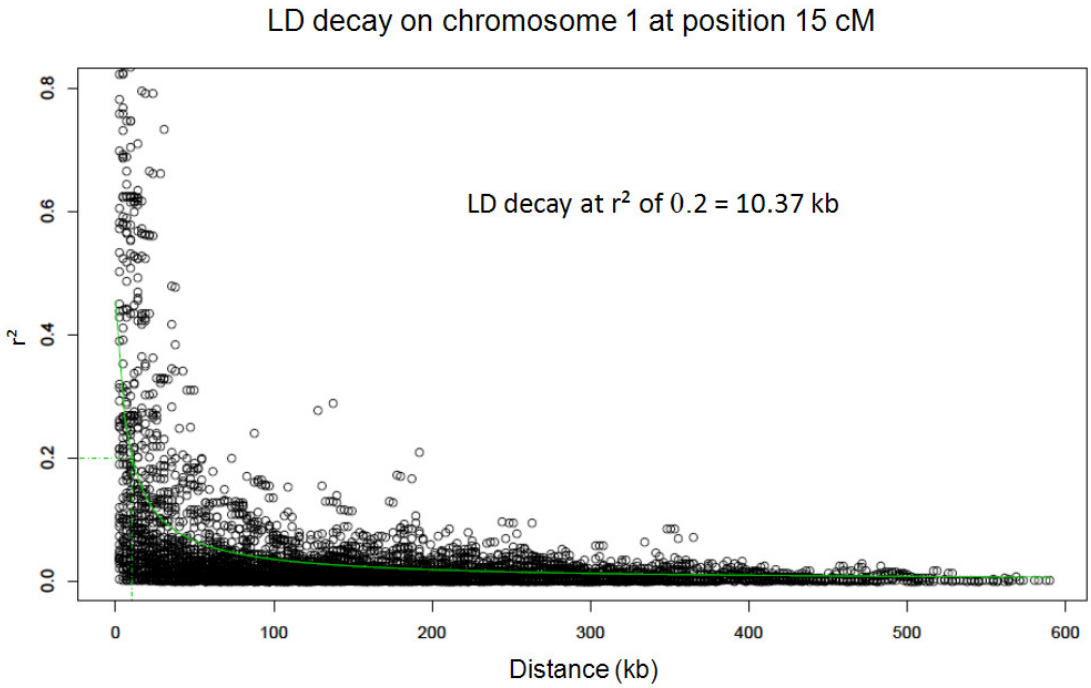


Figure 2.S2 LD decay between each pair of SNP in a 600 Kb window of a **neutral region (around the position 15 cM on the first chromosome) of one simulated replicate. Measures were performed with r^2_{SV} on 3000 individuals The green line represents the LD decay** (HILL and WEIR 1988) **estimated from the raw data (black point).**

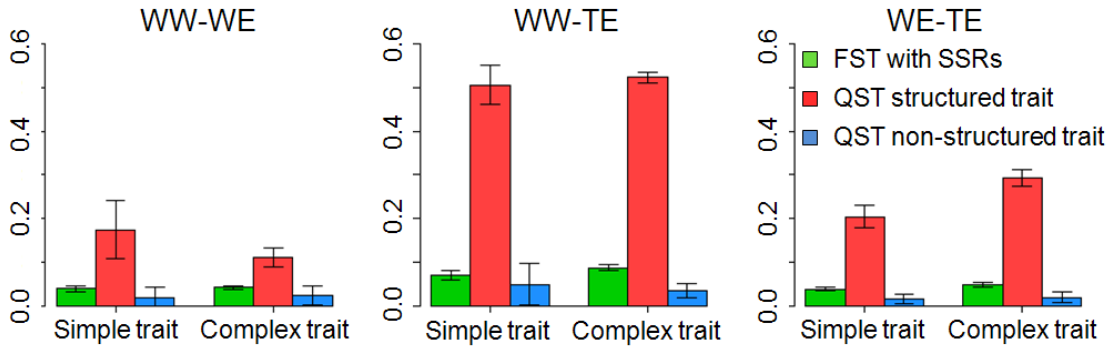


Figure 2.S3 Diversity indices of the simulated data: The mean F_{ST} and Q_{ST} between the three simulated populations (WW-WE-TE) calculated on selected and non-selected traits through 10 replicates for both simple and complex traits. Error bars were calculated with 95% confidence intervals on the estimates of the means.

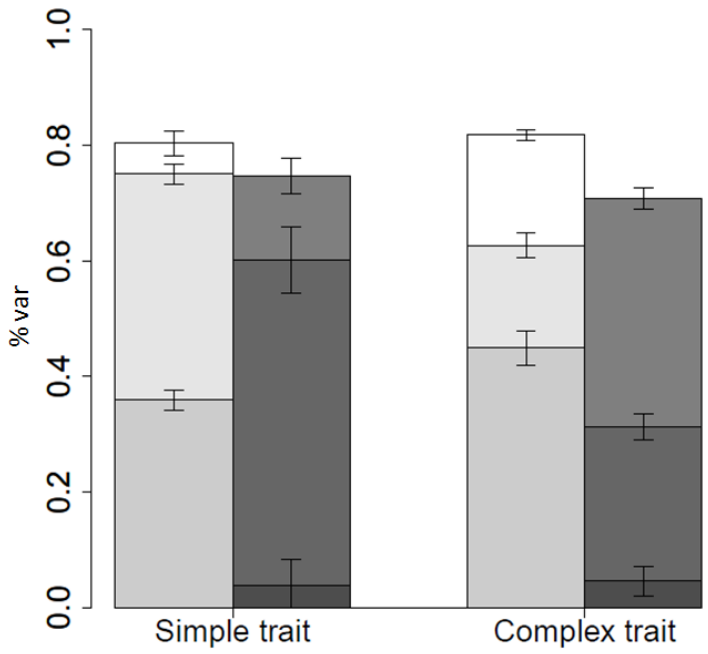


Figure 2.S4 Composition of the variance explained with the best model of mlmm. Results are showed for the structured and non-structured simple and complex traits on 10 replicate of the simulation. The first bars represent structured traits and the second ones represent non-structured traits. The darker color is the part explained by population structure, the intermediate color show the part of cofactors and the lightest represent the part of the polygenic term. To model selection we used mBonf criterion.

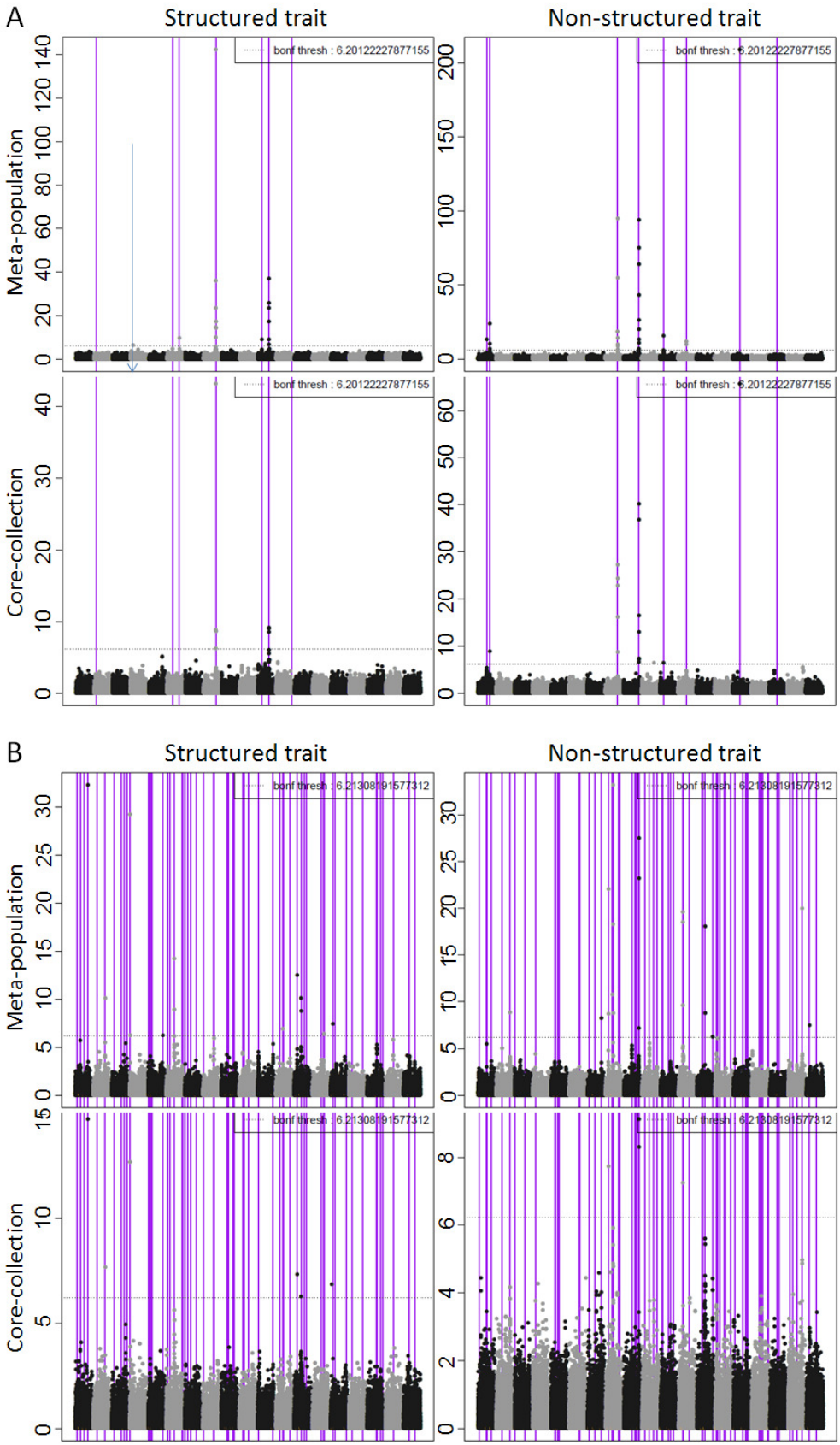


Figure 2.S5 Manhattan-plots of GWAS performed with mlmm on one replicate of the simulation. (A) presents the results for simple (10 QTLs) structured and non-structured trait on the core-collection of 1,000 individuals and on the entire meta-population (3,000 individuals). (B) part presents the results for complex traits (100 QTLs). Violet bars represent QTL loci with MAF>0.05, blue bars are QTLs **with MAF<0.05.**

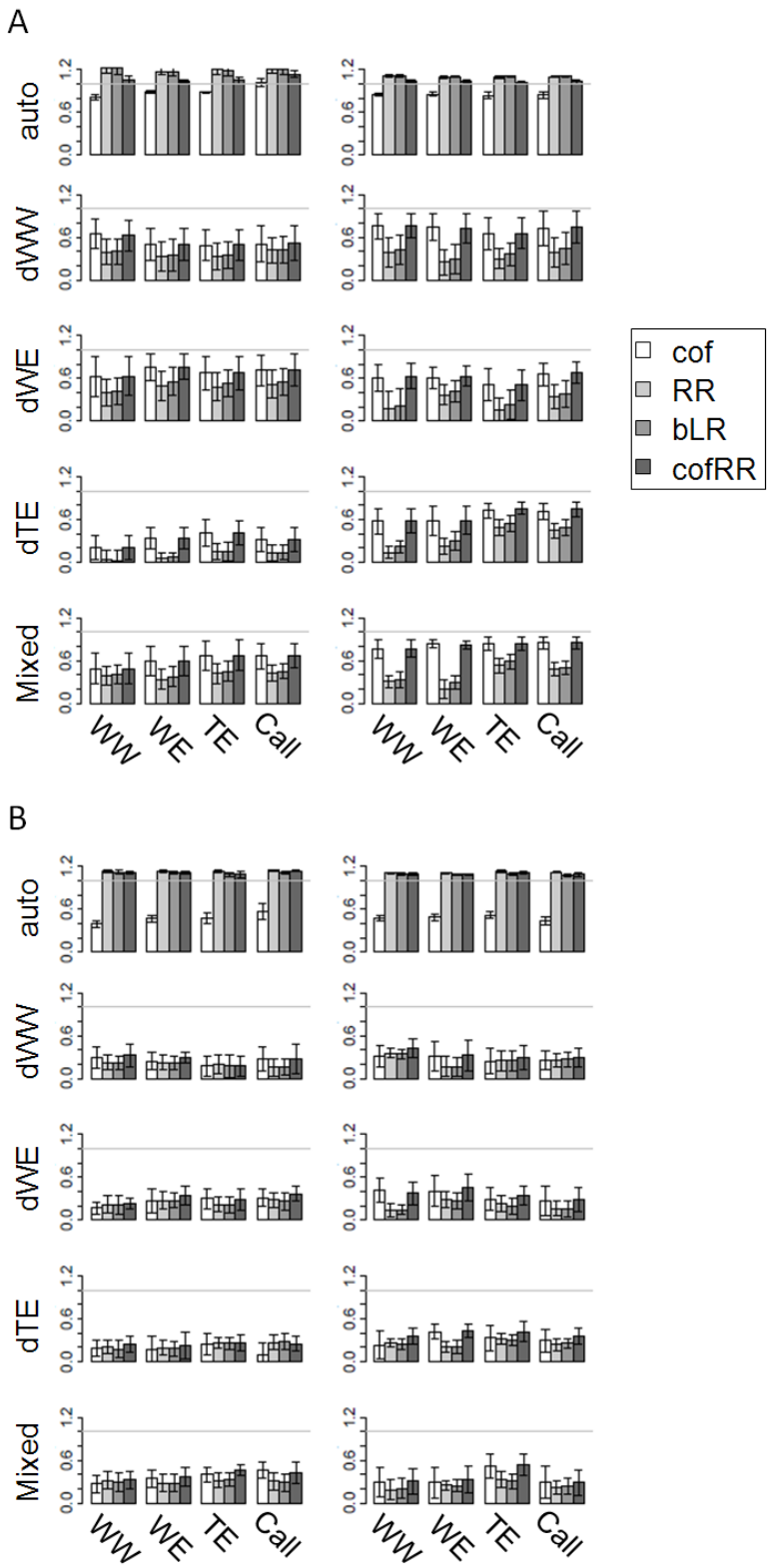


Figure 2.S6 The accuracy of the prediction through different combinations of training and candidate population. At the left side we show structured trait and on the right side the non-structured one. Colors represent the four prediction methods used: the sum of cofactor's effects identified in MLMM ("cof"), Ridge Regression BLUP ("RR"), Bayesian LASSO ("BLR") and marker assisted RR ("cofRR"). We used the three simulated populations (WW, WE, TE) and the entire core-collection (Call) as training population and realized prediction on the same sample (auto) and on each training sup-population (dWW, dWE, dTE, Mixed). On the left side we present the structured trait and on the right side the non-structured one. (A) presents the results on the simple trait (10 QTLs). (B) presents the complex trait (100 QTLs).

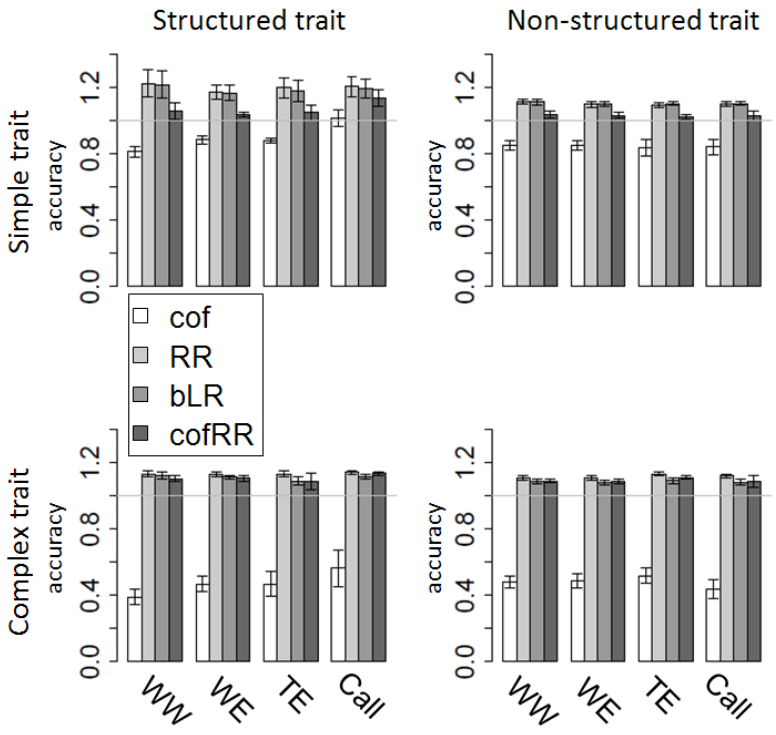


Figure 2.S7 The accuracy of auto-prediction on each training set (WW, WE, TE, Call) for all traits (structured / non-structured and simple or complex) with all four implemented methods (cof, RR, BLR, cofRR).

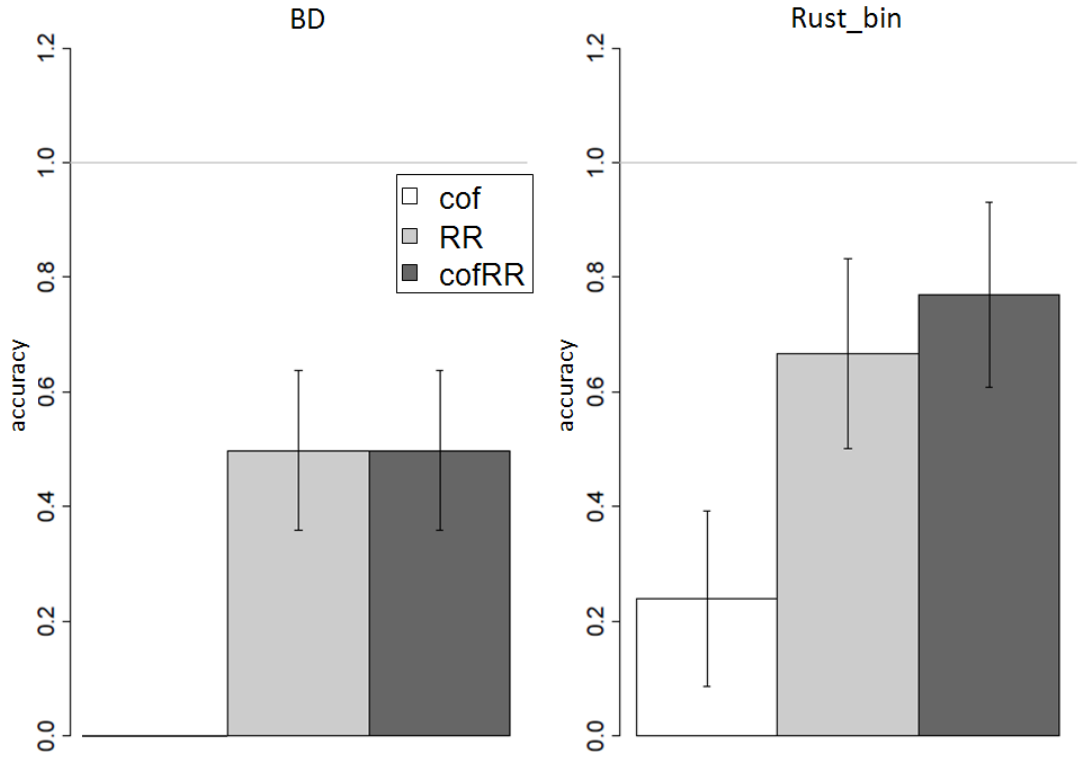


Figure 2.S8 Prediction accuracy for two traits of the pine data using cof, RR and cofRR methods. BD: average branch diameter of six years old trees; Rust_bin: fusiform rust susceptibility by presence or absence of rust. Error bars were calculated with 95% confidence intervals on the estimates of the means.

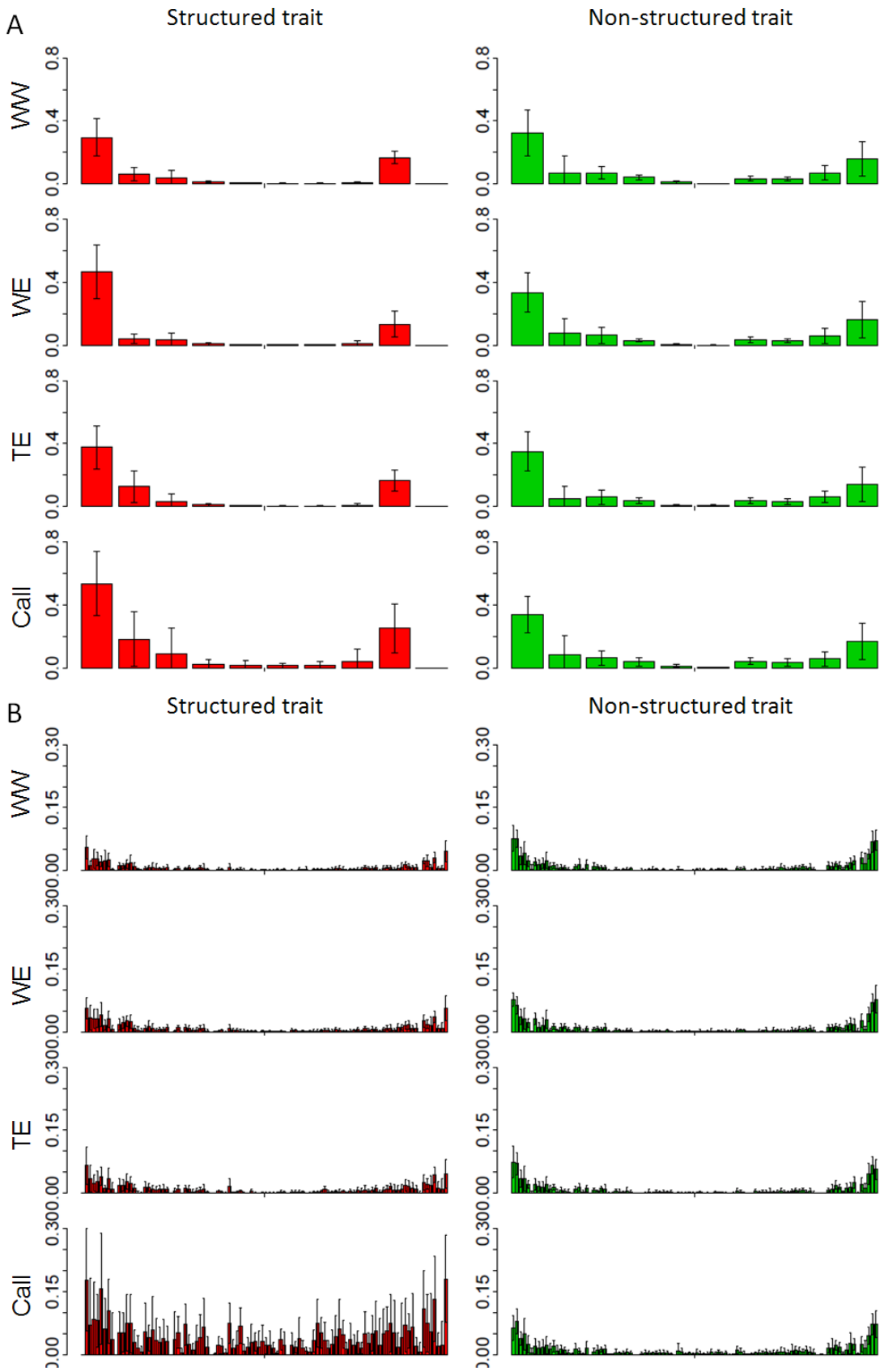


Figure 2.S9 The percentage of the variance explained by each QTL, measured separately on the three sub-population (WW, WE, TE) and on the core-collection (Call). Red bars are structured traits QTLs and greens are non-structured trait's QTLs**. (A) represents simple traits (10 QTLs). (B) complex traits (100 QTLs).**

File S1

GrapeSim.RData

File S1 is available for download as RData [.RData] at https://www.dropbox.com/s/x3esk4go6b34713/GrapeSim.RData

This file contains five R objects:

- X: a n by m matrix, where n=number of training individuals, m= number of SNPs, with rownames(X)=individual names, and colnames(X)=SNP names
- Xv: a nV by m matrix, where nV=number of validation individuals, m= number of SNPs, with rownames(Xv)=individual names, and colnames(Xv)=SNP names
- Y_ok: vector of phenotypes of the training set: a vector of length n, with names(Y_ok)=individual names
- Yv_ok: vector of phenotypes of the validation set: a vector of length nV, with names(Yv_ok)=individual names
- PC: a n by k matrix, where m= number of individuals, k= number of groups/PCA axes, with rownames(PC)=individual names colnames(PC) name of groups

File S2

COFRR_FODOR ET AL.R

File S2 is available for download as an R script [.r] at https://sites.google.com/site/vincentosegura/cofrr

Table 2.S1 Results of the GWA performed on the replicate N°25.

		structured trait							structured trait						
		Association sign.		max -	max	detected	distance	r^2_{SV}	Association sign.		max -	max	detected	distance	r^2_{SV}
		all	r^2>=0.05	log10(P)	var.expl	QTL	(kb)		all	r^2>=0.05	log10(P)	var.expl	QTL	(kb)	
Simple trait	meta-population	16	11	142.35	0.26	4	0-201	0.05-0.62	20	13	208.99	0.146	6	2-125	0.05-1
	core-collection	7	5	43.22	0.145	2	9-106	0.17-0.55	16	12	65.69	0.125	4	0-125	0,08-1
Complex trait	meta-population	13	10	32.3	0.052	10	2-135	0,06-0.66	19	4	27.54	0.033	3	0-132	0,15-0.36
	core-collection	6	5	14.71	0.027	5	4-149	0.1-0.72	4	2	9.13	0.029	1	9-45	0,2-0.33

Analyses were on the 3,000 individuals of the meta-population and the 1,000 individuals of the core-collection. Association was considered if the p-value passed the 5% Bonferroni threshold. The maximum of the –log(P-value), the variance explained by the SNP, the number of detected QTLs, the distance and the r^2_{SV} between significant SNP and QTL, were presented only for the associations where the r^2_{SV} between SNP and QTL was at least 0.05.

2. Annexe II

Données supplémentaires pour l'article « Genome-Wide Association Studies (GWAS) and Genomic Selection (GS) in grape for phenotype prediction using a large diversity panel »

Supplementary figures

Figure 3.S1. Repartition of the common and discriminant SNPs between sub-populations for each studied trait using additive (black) or dominant (blue) genotype encoding.

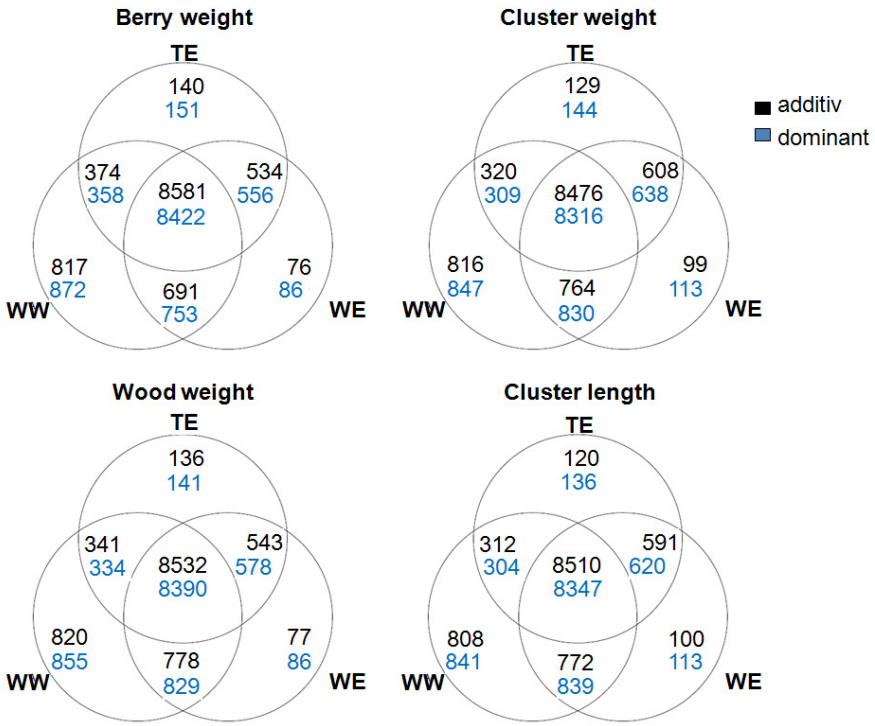


Figure 3.S2. Mean distance between neighboring SNPs. Values calculated on 11,463 SNPs polymorphic in the DP with less than 20% missing data detected on the 19 grapevine chromosomes, 12 random chromosomes ("_rdm") and an additional group of SNP with unknown position ("UN").

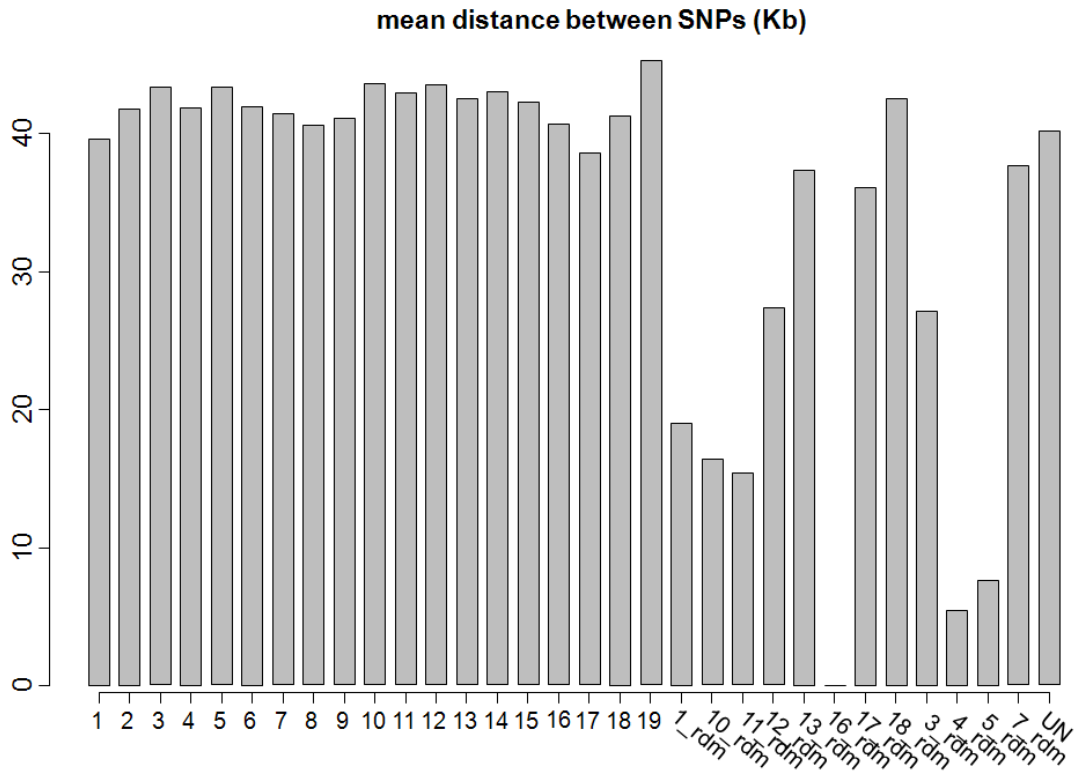
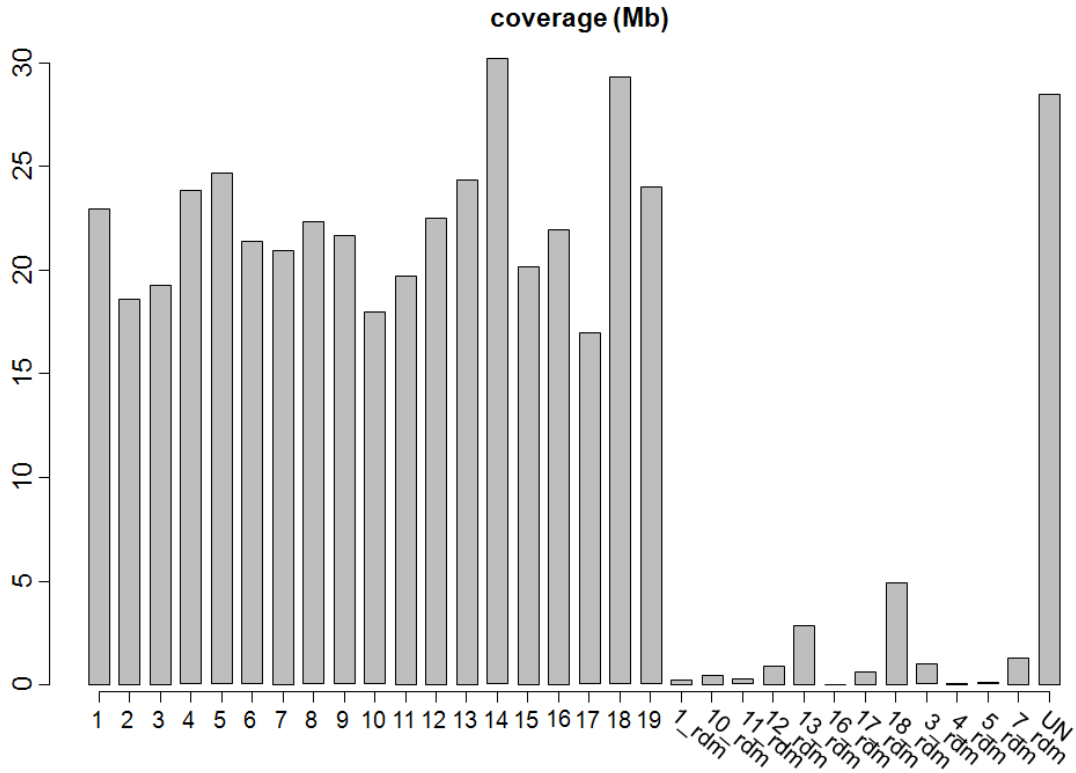
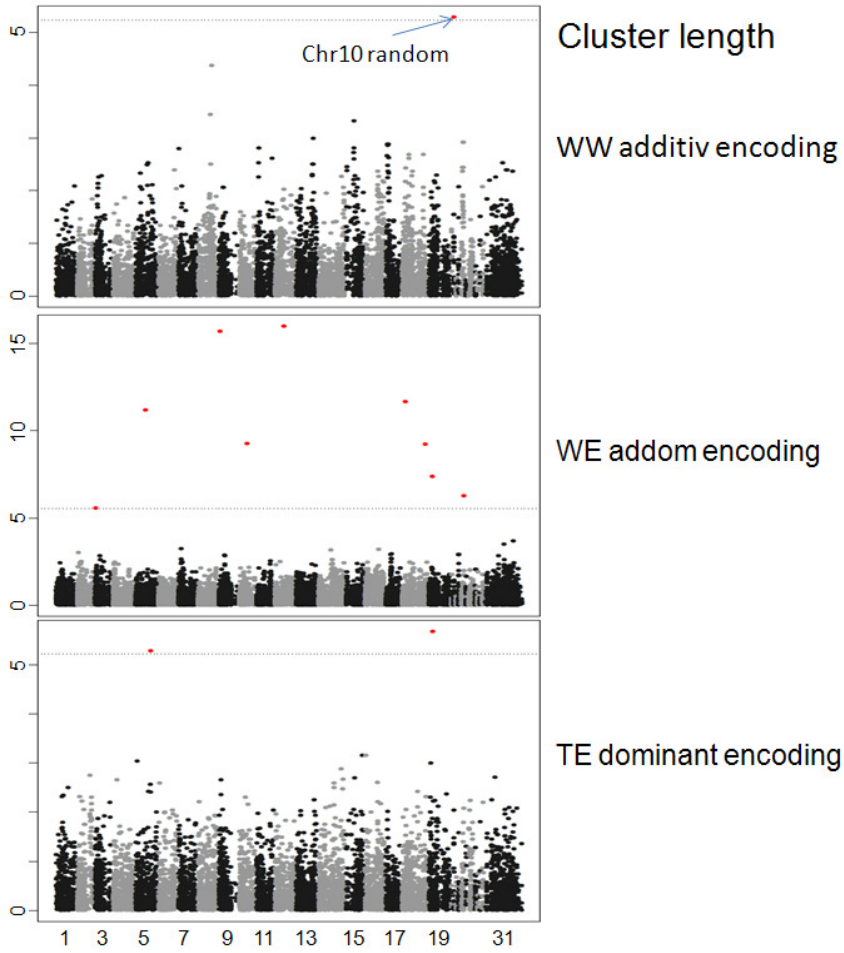


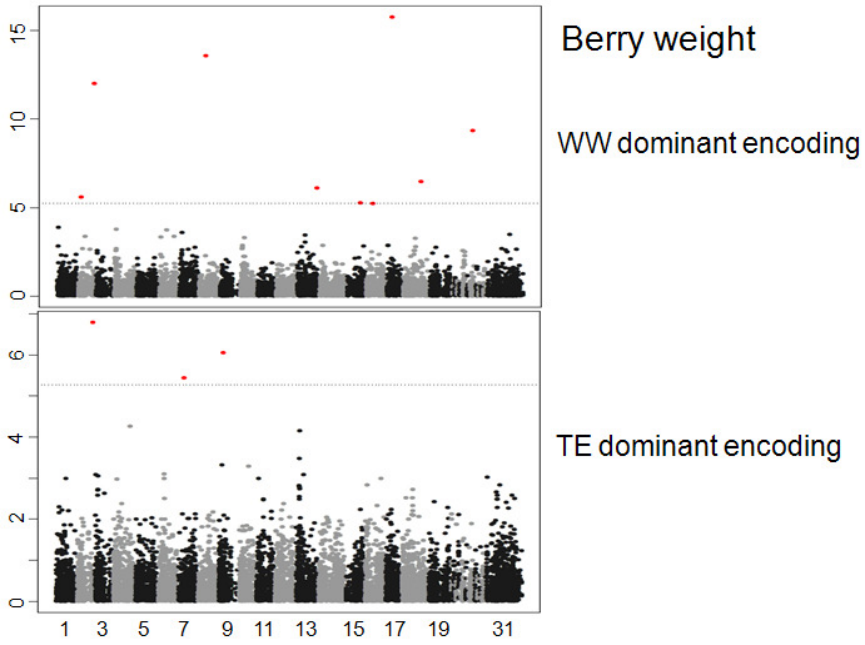
Figure 3.S3. Coverage of the chromosomes of grapevine genome using the filtered SNPs. Calculated on the set of 11,463 SNPs polymorphic in the DP containing less than 20% missing data. The x axis represent the 19 grapevine chromosomes, 12 random chromosomes ("_rdm") and an additional group of SNP with unknown position ("UN") and on the y axis we show the length covered by the markers in mega base-pair.



Fiure 3.S4. Manhattan-plots for cluster weight. Red spots are significant associations using the 5% Bonferroni threshold (cofactors).



Fiure 3.S5. Manhattan-plots for berry weight. Red spots are significant associations using the 5% Bonferroni threshold (cofactors).



Supplementary tables

Table 3.S1. Viruses detected on the diversity panel (DP). This table presents the varieties of the diversity panel. “Code “is their identification in the Vassal collection, “Pop” is the group of origin confirmed by BACILIERI *et al.* (2013) and remaining columns present the results of sanitary tests for five viruses: grapevine fan leaf virus (CNa), grapevine leaf roll associated viruses type 1 to 3 (GLRaV1, GLRaV2, GLRaV3) and grapevine fleck virus (GFkV).

Code	Name	Pop	CNa	GLRaV1	GLRaV2	GLRaV3	GFkV
0Mtp1480	Pasiga = CG 26-858	TE	0	0	0	0	0
0Mtp1557	VariŽtŽ d'oasis Bou Chemma 46	TE	0	0	0	0	0
0Mtp1650	Big Perlon	TE	0	0	0	0	0
0Mtp166	C 50-171 (Ramming)	TE	0	0	0	0	0
0Mtp29	Angoor Kalan	TE	0	0	0	0	0
1221Mtp2	Ichkimar	TE	0	0	0	0	0
1227Mtp2	Khoussa•nŽ blanc	TE	0	0	0	0	0
1247Mtp1	Baresana = Korithi aspro	TE	0	0	0	0	0
126Mtp2	Vassarga bielaia	TE	0	0	0	0	0
1367Mtp1	Teta de Vaca	TE	0	0	0	0	0
1563Mtp1	Ag isioum	TE	0	0	0	0	0
1715Mtp1	Kasoufi de la Bekaa	TE	0	0	0	0	0
1744Mtp1	Chahnani	TE	0	0	0	0	0
1972Mtp1	Garrido macho	TE	0	0	0	0	0
1992Mtp1	Abouhou	TE	0	0	0	0	0
2281Mtp1	Aetonyki	TE	0	0	0	0	0
2503Mtp1	Araxeni tchernii	TE	0	0	0	0	0
2505Mtp1	Assyl kara	TE	0	0	0	0	0
2635Mtp1	Koz ouzioum	TE	0	0	0	0	0
2644Mtp1	Mskhali	TE	0	0	0	0	0
2709Mtp1	Zimsko belo	TE	0	0	0	0	0
2774Mtp1	Sultanina	TE	0	0	0	0	0
2844Mtp1	Lialmigui	TE	0	0	0	0	0
2856Mtp1	Sourkhak biely	TE	0	0	0	0	0
2991Mtp1	Achlamiche	TE	0	0	0	0	0
2995Mtp1	Assouad Abou khislŽ	TE	0	0	0	0	0
3005Mtp1	Attiki	TE	0	0	0	0	0
308Mtp9	Muscat d'Alexandrie	TE	0	0	0	0	0
626Mtp1	Akiki	TE	0	0	0	0	0
0Mtp1513	Grain	WW	0	0	0	0	0
0Mtp171	Cahours	WW	0	0	0	0	0
0Mtp1724	Plant de Vic 98-N-3 (Collection Torres S.A.)	WW	0	0	0	0	0
0Mtp1747	Galotta	WW	0	0	0	0	0
0Mtp739	Morenoa	WW	0	0	0	0	0
1217Mtp1	Claverie Coulard	WW	0	0	0	0	0

167Mtp12	Altesse	WW	0	0	0	0	0
1837Mtp4	Arvine	WW	0	0	0	0	0
203Mtp1	Saint-Laurent	WW	0	0	0	0	0
2327Mtp1	B 5-6	WW	0	0	0	0	0
2713Mtp1	Mireille	WW	0	0	0	0	0
2725Mtp1	Bacchus	WW	0	0	0	0	0
2890Mtp1	Baserri	WW	0	0	0	0	0
2953Mtp1	Labrusco	WW	0	0	0	0	0
3009Mtp1	Balsamino blanc	WW	0	0	0	0	0
1232Mtp2	Gros Cabernet	WW	0	0	0	0	NA
403Mtp1	Aouillat	WW	0	0	0	0	NA
0Mtp1004	San Lorenzo	WW	0	0	0	NA	0
0Mtp89	Belle Denise	WW	NA	0	0	0	0
2066Mtp1	Humagne blanc	WW	NA	0	0	0	0
0Mtp1154	Urbanitraube noir (Collection Oberlin)	WE	0	0	0	0	0
0Mtp1758	Canella	WE	0	0	0	0	0
1628Mtp3	Dinka zšld	WE	0	0	0	0	0
1784Mtp1	Karitsiotis	WE	0	0	0	0	0
2174Mtp1	Pozsonyi feher	WE	0	0	0	0	0
2243Mtp4	Crimposie	WE	0	0	0	0	0
2247Mtp2	Feteasca regala	WE	0	0	0	0	0
2287Mtp1	Verdeca = Lagorthi	WE	0	0	0	0	0
2298Mtp2	Basicata	WE	0	0	0	0	0
2304Mtp1	Platyracho	WE	0	0	0	0	0
2745Mtp1	Heroldrebe = We S 130	WE	0	0	0	0	0
284Mtp4	Velteliner rouge	WE	0	0	0	0	0
80Mtp1	Gros Bourgogne	WE	0	0	0	0	0
1827Mtp1	Mission	WE	0	0	0	0	0
2318Mtp1	Strophyliatico	WE	0	0	0	NA	0
0Mtp1068	Skiadopoulo	WE	0	0	NA	1	1
601Mtp1	Aromriesling	WW	0	NA	0	1	0
0Mtp561	…reg Kadarka	WE	0	NA	0	1	1
1636Mtp1	Rudezusa	WE	NA	0	0	1	0
0Mtp831	Okatac faux	WE	NA	1	0	0	0
577Mtp1	Muscat de Terracina	WE	NA	1	0	0	0
1644Mtp1	Babic	WE	NA	1	0	1	0
2043Mtp1	Bianco d'Alessano	WE	NA	1	0	1	0
0Mtp1491	Kichmich tcherni	TE	0	0	0	0	1
0Mtp1761	Caprugnone	WE	0	0	0	0	1
0Mtp703	Matrassa blanc faux	TE	0	0	0	0	1
0Mtp995	Roussette basse de Seyssel	WW	0	0	0	0	1
1653Mtp1	Ruzevina	WE	0	0	0	0	1
1667Mtp3	Negru vertos	WE	0	0	0	0	1
1797Mtp1	Phokiano	WE	0	0	0	0	1
1888Mtp1	Balbut bijeli	WE	0	0	0	0	1
1893Mtp1	Hadari	TE	0	0	0	0	1

2057Mtp1	Impigno	WE	0	0	0	0	1
2657Mtp1	Soiaki	TE	0	0	0	0	1
2659Mtp1	Tagobi	TE	0	0	0	0	1
2683Mtp1	Sorok Let Oktiabria	WE	0	0	0	0	1
372Mtp1	Gibert	WW	0	0	0	0	1
0Mtp633	Lambrusco del Caset	WW	0	0	0	1	NA
2500Mtp1	Alexandroouli	TE	0	0	0	1	NA
0Mtp1005	Sao Mamede	WW	0	0	0	1	0
0Mtp1072	Souzao faux	WW	0	0	0	1	0
0Mtp1176	Verdelho tinto	WW	0	0	0	1	0
0Mtp1195	Magdeleine noire des Charentes	WW	0	0	0	1	0
0Mtp1213	Vulpea faux (Collection Ravaz)	WE	0	0	0	1	0
0Mtp1293	Plant de Pedebernade 1	WW	0	0	0	1	0
0Mtp1676	B 40-97 (Ramming)	TE	0	0	0	1	0
0Mtp313	Domina = Geilweilerhof 4-25-7	WW	0	0	0	1	0
0Mtp36	Arinto tinto (Collection Soares Franco)	WW	0	0	0	1	0
0Mtp416	Gharbi	TE	0	0	0	1	0
0Mtp469	Hagnos Zšld	WE	0	0	0	1	0
0Mtp537	Grosse MŽrille	WW	0	0	0	1	0
0Mtp636	Lameiro	WW	0	0	0	1	0
0Mtp752	Mourisco (Collection EVV Amandio Galhano)	WW	0	0	0	1	0
0Mtp799	Nevoeira	WW	0	0	0	1	0
0Mtp835	Osteiner	WW	0	0	0	1	0
0Mtp875	Pero Godal	WW	0	0	0	1	0
0Mtp952	Ramisco	WW	0	0	0	1	0
114Mtp4	Pagadebiti	WE	0	0	0	1	0
124Mtp1	Rimin se	WE	0	0	0	1	0
1266Mtp2	Raboso piave	WW	0	0	0	1	0
1287Mtp2	Lagrein	WW	0	0	0	1	0
1303Mtp3	Catarratto bianco lucido	WE	0	0	0	1	0
1494Mtp1	Verdelho tinto femelle	WW	0	0	0	1	0
1576Mtp2	Heunisch schwarz	WW	0	0	0	1	0
1578Mtp2	Kšvidinka	WE	0	0	0	1	0
157Mtp3	Corbeau	WW	0	0	0	1	0
18Mtp8	Carignan	WE	0	0	0	1	0
1909Mtp1	Bou Khanzir noir	TE	0	0	0	1	0
2003Mtp1	Azizi el Ja•a	TE	0	0	0	1	0
2013Mtp1	Khalt	TE	0	0	0	1	0
2070Mtp2	CacabouŽ	WW	0	0	0	1	0
2112Mtp1	Razdani	TE	0	0	0	1	0
2136Mtp2	Galbena de Odobesti	WE	0	0	0	1	0
2225Mtp5	Barlinka faux (Almeria)	TE	0	0	0	1	0
2282Mtp2	July Muscat	TE	0	0	0	1	0
2371Mtp1	Fumin	WW	0	0	0	1	0
2543Mtp1	Maingonnat 3 L 1	WE	0	0	0	1	0
2655Mtp1	SapŽrŽ otskhanouri	TE	0	0	0	1	0

2708Mtp1	Emerald seedless	TE	0	0	0	1	0
2842Mtp1	Ag Kiourdach	TE	0	0	0	1	0
2874Mtp2	Centennial seedless	TE	0	0	0	1	0
2892Mtp1	Donzelinho	WW	0	0	0	1	0
2893Mtp2	Corrin Seedless	TE	0	0	0	1	0
2968Mtp1	Mourtaou	WW	0	0	0	1	0
3016Mtp1	Massirart	WW	0	0	0	1	0
328Mtp2	Petit Verdot	WW	0	0	0	1	0
411Mtp1	Camaraou noir	WW	0	0	0	1	0
442Mtp2	NŽgret pointu	WW	0	0	0	1	0
537Mtp1	Terret Bouschet	WW	0	0	0	1	0
585Mtp62	Chasselas	WW	0	0	0	1	0
629Mtp1	Darkaia noir	TE	0	0	0	1	0
86Mtp2	Tibouren	WW	0	0	0	1	0
9Mtp3	Morrastel	WW	0	0	0	1	0
0Mtp1553	Morenzi	WE	0	0	0	1	1
0Mtp832	…kŸz gszŸ faux	TE	0	0	0	1	1
1531Mtp3	Carcajolo	WW	0	0	0	1	1
1662Mtp3	Rosa menna di vacca	WE	0	0	0	1	1
2471Mtp1	Tsitsa Kaprei	TE	0	0	0	1	1
2472Mtp1	Galbena uriasa	WE	0	0	0	1	1
569Mtp1	Moscato giallo	WE	0	0	0	1	1
632Mtp2	Frankenthal rouge foncŽ	TE	0	0	0	1	1
1805Mtp1	Araklinos	WE	0	0	1	0	0
2342Mtp1	Albaranzeuli bianco	WE	0	0	1	0	0
0Mtp500	Plant de Ponteilla (Jaubert)	WE	0	0	1	0	1
1483Mtp3	Souzao	WW	0	0	1	0	1
0Mtp406	Mourv□dre GoulŽ	WE	0	0	1	1	0
1248Mtp2	Zeini abiad	TE	0	0	1	1	0
2317Mtp1	Staphidampelo	WE	0	0	1	1	0
252Mtp1	Poulsard	WW	0	0	1	1	0
413Mtp1	Courbu	WW	0	0	1	1	0
635Mtp1	Dattier noir	TE	0	0	1	1	0
975Mtp3	Pardina = Pirovano 130	TE	0	0	1	1	1
0Mtp220	Chami abiad	TE	0	1	0	0	0
0Mtp318	Doppel Augen	TE	0	1	0	0	0
0Mtp604	KolossiŽ gebirgig	TE	0	1	0	0	0
0Mtp982	Rosaky rose faux	WE	0	1	0	0	0
1186Mtp1	Chirai obak	TE	0	1	0	0	0
1218Mtp1	Tavkveri	TE	0	1	0	0	0
1654Mtp1	Posip bijeli	WE	0	1	0	0	0
1801Mtp1	Skylopnichtis	WE	0	1	0	0	0
1815Mtp1	Tachtas	TE	0	1	0	0	0
2074Mtp1	Siah	TE	0	1	0	0	0
2218Mtp1	Avarengo	WW	0	1	0	0	0
329Mtp1	GhemŽra	WE	0	1	0	0	0

74Mtp31	Ugni blanc	WE	0	1	0	0	0
848Mtp1	Santa Morena	TE	0	1	0	0	0
0Mtp811	Noir de CrimŽe faux	TE	0	1	0	0	1
227Mtp1	Roublot	WW	0	1	0	1	NA
0Mtp1073	Starinky	WE	0	1	0	1	0
0Mtp1156	Urmi dinka	WE	0	1	0	1	0
0Mtp1235	Tzimliansky belyi	WE	0	1	0	1	0
0Mtp300	Dili kaftar	TE	0	1	0	1	0
0Mtp569	Kara oglan faux	WE	0	1	0	1	0
0Mtp715	MŽdouar	TE	0	1	0	1	0
1620Mtp1	Tantovina	WE	0	1	0	1	0
1631Mtp1	Blank blauer	WE	0	1	0	1	0
1673Mtp5	Chaouch blanc	TE	0	1	0	1	0
1753Mtp1	Verico	TE	0	1	0	1	0
219Mtp2	Arbane	WW	0	1	0	1	0
2886Mtp1	Blanchier	WW	0	1	0	1	0
2982Mtp1	Ahmeh sal apyr□ne	TE	0	1	0	1	0
433Mtp1	Razachie rosie	WE	0	1	0	1	0
1277Mtp6	Primitivo	WE	0	1	0	1	0
0Mtp1474	Koutlaksky belyi	TE	0	1	0	1	1
1648Mtp1	Nincusa	WE	0	1	0	1	1
1814Mtp1	Vidiano	WE	0	1	0	1	1
2621Mtp2	Nieddera	WE	0	1	0	1	1
443Mtp19	Mauzac	WW	0	1	0	1	1
1284Mtp1	Montepulciano	WE	0	1	0	1	1
1493Mtp1	Touriga	WW	0	1	1	1	0
0Mtp1218	Wildbacher de Hongrie (Collection Ravaz)	WW	1	0	NA	1	0
3000Mtp1	Asprouda Zakinthou	WE	1	0	NA	1	0
0Mtp1033	Sasca	WE	1	0	0	0	0
0Mtp1129	Totika	WE	1	0	0	0	0
1629Mtp1	Beregi rozsas	WE	1	0	0	0	0
0Mtp408	Garbanega faux (Istituto San Michele)	TE	1	0	0	1	0
0Mtp440	Graeco	TE	1	0	0	1	0
0Mtp562	Kakotryghis	WE	1	0	0	1	0
0Mtp64	Avgoustiatis	WE	1	0	0	1	0
0Mtp775	Muscate (Collection Ravaz)	WE	1	0	0	1	0
0Mtp886	Plyto	TE	1	0	0	1	0
1570Mtp1	Mezesfeher	WE	1	0	0	1	0
2597Mtp1	Peikani	TE	1	0	0	1	0
424Mtp2	Lauzet B	WW	1	0	0	1	0
0Mtp1074	Stavroto faux (Collection Lykovrissi)	WE	1	0	0	1	1
0Mtp581	Khikhvi	TE	1	0	0	1	1
0Mtp796	Nero grosso	TE	1	0	0	1	1
1258Mtp1	Verdea	WE	1	0	0	1	1
2348Mtp1	Cococciola	WE	1	0	0	1	1
2460Mtp2	Negru mare	TE	1	0	0	1	1

50Mtp1	Alfrocheiro preto	WW	1	0	1	0	0
1245Mtp1	Freisa	WW	1	0	1	0	1
0Mtp1512	Plant de Chaudefonds 53 (Fardeau)	WW	1	1	0	0	0
0Mtp610	Korza erevani	TE	1	1	0	0	0
2373Mtp1	Greco bianco	WE	1	1	0	1	0
613Mtp1	Kaisermuskat	WW	1	1	0	1	0
0Mtp961	Ribote rose	WE	1	1	0	1	1
0Mtp1733	Fortunato	TE	0	0	0	1	0
0Mtp270	Corason de Kabritte blanc	WW	0	0	0	1	0
11Mtp3	Piquepoul noir	WW	0	0	0	1	0
1237Mtp1	Rossara trentina	WE	0	0	0	1	0
1261Mtp1	Coda di volpe bianca	WE	0	0	0	0	0
129Mtp12	Chatus	WW	0	0	1	1	0
1301Mtp1	Verdiso	WW	1	0	0	1	0
1307Mtp1	Inzolia	WE	0	1	0	1	0
1314Mtp1	Perricone	TE	0	1	1	1	0
1338Mtp1	Bellone	WE	0	1	0	0	0
1354Mtp2	Bonamico	WE	0	0	0	1	0
1365Mtp5	Ohan□s	TE	0	0	0	0	1
139Mtp1	Dureza	WW	0	0	0	1	NA
154Mtp1	Joubertin	WW	1	0	0	1	0
1583Mtp3	Affenthaler	WW	0	1	0	0	0
1695Mtp1	Assoued kere	TE	1	0	1	0	0
176Mtp1	Mondeuse blanche	WW	1	0	0	1	0
1844Mtp2	Malvasia istriana	WE	0	0	0	1	0
188Mtp1	M□cle de Bourgoin	WW	0	0	0	1	NA
1Mtp3	Rivairenc = Aspiran noir	WW	0	0	0	1	0
2104Mtp1	Bogazkere	TE	0	0	0	1	1
2107Mtp1	Dimrit	WE	0	0	0	1	1
210Mtp1	Gouais blanc	WE	0	0	0	1	0
226Mtp7	Gascon	WW	0	0	0	1	0
2349Mtp1	Lambrusco Marani	WW	0	1	0	1	0
2418Mtp1	Vermentino nero	WE	0	0	0	1	1
257Mtp16	Savagnin blanc	WW	0	0	0	1	0
261Mtp2	Argant	WE	1	0	0	1	0
2694Mtp1	Landroter	WW	0	1	0	1	0
26Mtp2	Clairette	WW	0	0	NA	1	0
2747Mtp1	Arinarnoa	WW	0	0	0	1	0
2902Mtp1	Gantziandan	TE	0	1	0	0	0
2972Mtp1	Codivarta	WE	0	0	0	1	1
332Mtp1	Semillon	WW	0	0	0	1	0
344Mtp2	Blanc Auba	WW	0	0	0	1	0
349Mtp1	Penouille	WW	0	0	0	1	1
380Mtp1	Baroque	WW	0	0	1	1	1
43Mtp1	Olivette rose	TE	0	1	1	1	0
446Mtp3	Len de l'El	WW	0	0	0	1	0

44Mtp1	Alba imputotato	WE	0	1	0	0	0
556Mtp2	Katta-kourgan	TE	0	0	0	1	0
657Mtp1	Kalili	TE	0	0	0	1	1
666Mtp3	Danugue	TE	1	0	0	1	0
672Mtp2	Molinera gorda	TE	1	0	0	1	0
725Mtp1	Kolliniatico	WE	0	1	0	1	0
727Mtp1	Nehelescol	TE	0	0	1	1	0
735Mtp1	Dabouki	TE	0	0	1	0	0
744Mtp1	Kolontar	WE	1	0	1	1	0
746Mtp1	Kizil	TE	0	1	0	1	0
749Mtp2	Coarna alba	WE	0	0	0	1	0
789Mtp1	Genk Uzum	WE	0	1	0	0	0
791Mtp1	Chaptal	TE	0	0	1	1	0
830Mtp1	Salicette (Collection Parc de la T?te d'Or)	WW	0	0	0	0	0

Table 3.S2. Statistical analysis of phenotypic data. Type of transformation (transform.) and equation of the best mixed model fitted on the four studied traits.

Trait	Transform.	Mixed model
Berry weigh	ln	$\mu + G_i + b_j + y_k + (G \times y)_{ik} + (G \times b)_{ij} + e_{ijk}$
Cluster weight	-	$\mu + G_i + b_j + y_k + (G \times y)_{ik} + (b \times y)_{jk} + m_2 + m_3 + m_4 + m_5 + e_{ijk}$
Pruning weight	Root-square	$\mu + G_i + b_j + y_k + (G \times b)_{ij} + (b \times y)_{jk} + e_{ijk}$
Cluster length	Root-square	$\mu + G_i + m_3 + +m_5 + e_{ijk}$

Where μ the overall mean, G_i the random effect of genotype *i*, b_j the fixed effect of block *j*, y_k the fixed effect of year *k*, m_1 to m_5 the fixed effects of viruses (grapevine fan leaf virus, grapevine leaf roll associated viruses type 1 to 3 and grapevine fleck virus) and e_{ijk} the residual error effect. $(G \times b)_{ij}$ was the interaction between genotype *i* and block *j*, $(G \times y)_{ik}$ the interaction between genotype *i* and year *k* and $(b \times y)_{jk}$ the interaction between block *j* and year *k*.

Table 3.S3. The list of significant SNPs from GWAS. The column "cofactor" shows the percentage of variance explained by the SNP, the sign of its effect on the trait (+/-), the encoding (D for dominant and A for additive). Bold letters are associations explaining more than 10% of the variance. "ad" marks SNP detected when jointly analyzing additive and dominant encoding (addom). The sign * indicate the SNP significant (5% Bonferroni and FDR) in the emmax step.

Chromosome	Trait	Population	cofactor	position (bp)
Chr2	berry weight	WW	**23% + D**	**2 760 746**
			13% + D	**18 305 755**
		TE	**24% - D**	**17 903 397**
Chr3	cluster length	WE	**21% + A ad**	**574 803**
Chr5	cluster length	WE	**14% - A ad**	**11 894 106**
		TE	**18% + D**	**17 966 666**
Chr7	berry weight	TE	**14% - D**	**6 265 676**
Chr8	berry weight	WW	**12% - D**	**7 340 158**
Chr9	berry weight	TE	**17% + D**	**4 865 125**
	cluster length	WE	**12% + D ad**	**891 660**
Chr10	cluster length	WE	7% + D ad	8 922 806
Chr12	cluster length	WE	9% + D ad	10 296 309
Chr13	berry weight	WW	10% - D	22 816 709
Chr15	berry weight	WW	8% + D	15 155 685
Chr16	berry weight	WW	6% + D	7 104 351
Chr17	berry weight	WW	4% - D	5 772 346
Chr18	berry weight	WW	3% + D	20 659 438
	cluster length	WE	**12% - A ad**	**3 680 788**
			5% + D ad	27 301 869
Chr19	cluster length	WE	6% - D ad	3 851 881
		TE	**20% - D**	**4 147 367**
Chr10 rdm	cluster length	WW	**28% - A ***	**180 028**
Chr13 rdm	cluster length	WE	5% + D ad	2 902 290
Chr18 rdm	berry weight	WW	2% - D	3 575 533

Table 3.S4. List of the genes identified around the associations. Boldface letters indicates genes …

Population	Trait	Chromosome	Position	Genes included in a window of 10 kb around the position	Putative function
TE	berry weight	Chr02	17903397	VIT_02s0087g00490	10-deacetylbaccatin III 10-O-acetyltransferase
				VIT_02s0087g00500	MAP kinase 9
		Chr07	6265676	**VIT_07s0005g03380**	Translocase inner membrane subunit 44-2 ATTIM44-2
				VIT_07s0005g03390	Unknown protein
		Chr09	4865125	**VIT_09s0002g05170**	Unknown protein
				VIT_09s0002g05180	No hit
	cluster length	Chr05	17966666	**VIT_05s0062g00010**	PUMILIO 5 (APUM5)
		Chr19	4147367	**VIT_19s0014g03920**	Ubiquitin-specific protease 14 (UBP14)
WE	cluster length	Chr03	574803	**VIT_03s0038g00640**	Unknown
				VIT_03s0038g00650	Coenzyme Q10 homolog B
		Chr05	11894106	VIT_05s0051g00830	Dihydroxy-acid dehydratase
		Chr09	891660	**VIT_09s0002g01150**	RelA/SpoT protein (RSH3)
				VIT_09s0002g01160	Transcription initiation factor TFIID subunit D7
		Chr10	8922806	VIT_10s0003g05010	No hit
		Chr12	10296309	**VIT_12s0057g01540**	Regulator of chromosome condensation (RCC1)
		Chr18	3680788	**VIT_18s0001g04040**	PHD finger transcription factor
		Chr18	27301869	VIT_18s0041g02160	Lipase GDSL
		Chr19	3851881	**VIT_19s0014g03740**	Membrane-associated mannitol-induced
				VIT_19s0014g03750	No hit
		Chr13_rdm	2902290	**VIT_13s0019g01290**	Crossover junction endonuclease MUS81
WW	berry weight	Chr02	2760746	VIT_02s0025g03220	Trihelix DNA-binding protein (GT2)
				VIT_02s0025g03230	Fringe protein
		Chr02	18305755	VIT_02s0087g00740	Unknown protein
		Chr08	7340158	VIT_08s0105g00260	Digalactosyldiacylglycerol synthase 1
		Chr13	22816709	**VIT_13s0064g00950**	Unknown protein
		Chr15	15155685	VIT_15s0048g01030	Retrovirus Pol polyprotein
				VIT_15s0048g01040	DNA repair protein

		Chr16	7104351	**VIT_16s0013g01140**	Phosphofructokinase
		Chr17	5772346	VIT_17s0000g05280	UPF0737 protein AFP3
		Chr18	20659438	VIT_18s0072g01080	DnaJ homolog, subfamily C, member 8
				VIT_18s0072g01090	TIR-NBS-LRR type R protein 7
		Chr18_rdm	3575533	**VIT_18s0001g03900**	R protein L6
				VIT_18s0001g03920	No hit
	cluster length	Chr10_rdm	180028	VIT_10s0116g00380	Ovate family protein 3 OFP3
				VIT_10s0116g00400	SEC14 cytosolic factor

3. Annexe III

Rapport de formation « Valorisation des compétences – un nouveau chapitre de la thèse® »

Durant la troisième année de ma thèse j'ai suivi une formation de 10 jours, organisée par L'ABG-Intelli'agence (http://www.intelliagence.fr/). L'exercice proposé par cette formation était de réaliser un inventaire et une mise en valeur des compétences qui sont nécessaires au bon déroulement de la thèse. Il n'a aucun caractère académique. Il présente toutes les activités abordées en lien avec le projet de la thèse, même si elles n'ont pas abouti sur des résultats présentables dans la partie académique du manuscrit. Cependant, sa réalisation m'a éclairée sur de nouveaux aspects de mon projet de thèse, qui ont soulevé ma curiosité, et complète mes acquis scientifiques et techniques issus de ces travaux.

Je tiens à remercier mon directeur de thèse et à mon encadrant d'avoir donné leur accord pour communiquer ce document non-scientifique dans le manuscrit de thèse.

Le rapport final de cette formation présente une analyse critique de la manière dont j'ai conduit et géré ma thèse en tant que projet. J'en tire des conclusions quant aux qualités personnelles et aux savoir-faire que j'ai développés, sachant que ces aptitudes me serviront tout au long de ma vie professionnelle.

Ce bilan est organisé autour des trois points suivants :

- une analyse critique du déroulement de la thèse comme expérience professionnelle de gestion de projet,
- une identification et une illustration des différents acquis professionnels et compétences personnelles qui ont été développés pendant la thèse,
- identification des perspectives de valorisation en termes de pistes professionnelles envisagées à moyen terme.

Valorisation des compétences, NCT®

Agota FODOR

Ecole doctorale : Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement (SIBAGHE)

Organisme de rattachement : Montpellier SupAgro

Mentor : Nathalie Camus

Créer les cépages de demain avec les outils d'aujourd'hui

Date de la présentation orale : le 20 juin 2013

Sujet académique de la thèse : La Sélection Génomique appliquée chez la vigne, évaluation et utilisation

Directeur de thèse : Patrice This

Encadrant : Loïc Le Cunff

Date probable de soutenance de thèse : décembre 2013.

Cadre générale et enjeu de la thèse

Aujourd'hui la viticulture française doit faire face à plusieurs défis. En effet, la vigne est aujourd'hui une des espèces les plus fortes utilisatrices de produits phytosanitaires en Europe. Elle sera confrontée comme toutes les cultures aux évolutions du climat qui pourraient engendrer de profondes modifications et ce notamment dans la zone méditerranéenne. Enfin, elle doit faire face à une compétition de plus en plus soutenue de la part des autres pays notamment ceux du nouveau monde. A l'heure actuelle, l'organisation de la viticulture laisse peu de place à l'introduction de nouveaux cépages sur le marché : En effet les AOCs imposent un unique lieu de production mais aussi une liste de cépages pour produire un vin d'appellation. Cependant, la création de nouvelles variétés apparaît comme une solution de plus en plus incontournable pour répondre à ces défis.

L'utilisation des données moléculaires (information de code ADN) dans l'amélioration des espèces est aujourd'hui de plus en plus répandue. Si on arrive à bien comprendre le lien entre l'information génétique et le phénotype observé sur la plante, nous pouvons très tôt sélectionner les individus intéressants issus d'un croisement. Jusqu'à nos jours plusieurs méthodes ont été développées, testées et mises en pratique chez diverses espèces animales et végétales, et donnent des résultats prometteurs.

Dans ma thèse j'étudie la faisabilité de prédire les phénotypes de vigne en observant uniquement leur ADN en s'appuyant sur deux méthodes récentes appelées génétique d'association sur tout le génome (GWAS) et prédiction génomique (GS). Ces deux méthodes se basent sur l'observation de l'information génétique sur tout le génome mais elles l'utilisent différemment. Les résultats de ce travail ont pour objectif d'apporter aux sélectionneurs des réponses quant au choix d'utilisation de ces méthodes en fonction de l'objectif de leurs programmes d'amélioration (complexité du caractère à améliorer, des moyens financiers et le matériel végétal disponible).

Mon projet de thèse est composé de deux axes principaux : (i) une étude de faisabilité, définition des conditions optimales et des limites des deux méthodes en utilisant un jeu de données simulé (ii) la comparaison des deux méthodes sur un jeu de donnée issu du vignoble (données réelles). Durant ces travaux j'ai mis en place des collaborations pour créer des outils (notamment un pipeline d'analyse) qui peuvent être utilisés dans des programmes d'amélioration.

La thèse dans son contexte

Ma thèse est financée dans le cadre d'une convention CIFRE entre l'INRA de Montpellier et l'Institut Français de la Vigne et du Vin (IFV, entreprise semi-publique). Elle se déroule au sein de l'unité mixte technologique Géno-Vigne® qui regroupe plusieurs instituts et équipes : l'équipe Diversité et Adaptation de la Vigne et des Espèces Méditerranéennes (DAVEM) de l'UMR AGAP, composée de personnel de l'INRA et de Montpellier SupAgro, l'unité expérimentale du domaine de Vassal, et le pôle matériel végétal de l'IFV. Je suis localisée au sein de l'UMT Géno-VIgne® sur le campus de Montpellier SupAgro.

Mes travaux s'intègrent dans un projet plus large, lauréat de l'appel à projet CASDAR 2010 recherche finalisée et innovation ouvert au institut technique agricole ; qui a pour but le développement d'outils moléculaires, de méthodologies et de matériel végétal innovant au service de la création variétale chez la vigne. Ce projet s'appuie sur d'autres projets plus vastes menés dans l'équipe DAVEM comprenant la mise en place du dispositif expérimental et l'utilisation de puissantes méthodes de phénotypage pour l'étudier les stress environnementaux.

Moi dans ce contexte

J'ai toujours été intéressée par la biologie et plus précisément par les plantes. Après le baccalauréat j'ai choisi de poursuivre mes études dans le domaine de l'horticulture à l'Université Corvinus de Budapest (Budapest, Hongrie). Certains de mes professeurs, notamment en génétique, m'ont fasciné par leur approche scientifique de problèmes complexes et m'ont donné le goût de la recherche et la création de connaissances. Je me suis spécialisée en génétique et amélioration des plantes pour m'orienter vers la recherche appliquée. Ces réussites à l'université et mes stages m'ont encouragé et réveillé l'ambition de réaliser un projet de recherche de trois ans, une thèse.

L'envie de voyager, découvrir et apprendre à connaître différentes cultures m'a amené à faire un séjour Erasmus pendant mes études de Master. J'ai choisi la France et l'école de Montpellier SupAgro car leur programme en amélioration des plantes semblait pouvoir apporter les compléments nécessaires à ma formation. J'ai beaucoup apprécié la qualité de l'enseignement et l'expérience professionnelle acquise au cours de stages, durant mon séjour en France. La possibilité de me former à la recherche dans un milieu de renommée internationale ayant des moyens techniques humains et financiers plus importants que dans mon pays m'a motivée pour continuer mon parcours en France malgré les défis rencontrés avec la langue.

Mon parcours universitaire m'a permis d'avoir un premier aperçu de la recherche publique. Pour élargir mes expériences vers la recherche privée j'ai cherché une thèse en partenariat entre le secteur public et privé.

La vigne est une espèce pérenne horticole qui rend son étude difficile mais du fait de son impact économique, des moyens importants sont mis à la disposition de la recherche pour mieux comprendre son fonctionnement et pour travailler ses caractéristiques. J'ai choisi une thèse sur cette plante pour acquérir des connaissances modernes sur l'amélioration des plantes en restant fidèle à l'axe de l'horticulture dans mon parcours.

Déroulement, gestion et coût de mon projet

Préparation et cadre du projet

Le premier axe de ma thèse est la réalisation d'une étude de faisabilité basée sur des simulations. Il s'agit de la reproduction virtuelle du matériel existant pour observer l'influence de certains paramètres (nombre de marqueurs moléculaires, architecture génétique et héritabilité du caractère) sur l'efficacité des deux méthodes de sélection comparées. L'objectif est d'avoir une première idée de l'efficacité et des limites de ces deux méthodes chez la vigne mais aussi d'identifier les conditions idéales pour leur application.

Les connaissances acquises et les outils d'analyses développés dans la première partie vont aider à orienter et à réaliser le deuxième axe de la thèse qui est le test (ou la validation) de la théorie sur les données réelles. Pour ce faire, différents outils et ressources ont été mis à la disposition de l'équipe grâce aux collaborations.

1.1. Le dispositif expérimental représentant la diversité de la vigne a été mis en place dans le cadre du projet ANR DL-Vitis coordonnée par l'équipe DAVEM.
1.2. Des méthodes de phénotypage haut-débit ont été mises au point dans le cadre du projet VitSeq pour des caractères impliqués dans la tolérance à la sécheresse.
1.3. Le génotypage du matériel végétal a été réalisé en partie dans le cadre d'un projet Européen KBBE GrapeReSeq, qui permettait aussi le développement de l'outil de génotypage (une puce 18 KSNP) et en partie payé par le projet CASDAR.

Conduite du projet

L'encadrement de ma thèse était partagé entre mon directeur de thèse et mon encadrant direct représentant de l'entreprise IFV. Pendant les deux premières années nous avons fait des réunions ponctuelles qui sont devenues des événements réguliers en dernière année. J'ai appris à préparer des rapports d'étapes, résumer les points importants de l'avancement des différentes activités liés à mon projet, soulever des problèmes rencontrés et proposer des solutions possibles.

L'école doctorale dans laquelle se déroule ma thèse impose une réunion annuelle entre le doctorant, et un comité de pilotage composé des spécialistes de différents domaines en lien avec la thèse. Pour ces réunions j'ai préparé des rapports d'avancement détaillés et des exposés oraux suivis d'une discussion sur l'orientation scientifique et pratique du projet de thèse.

Pour approfondir certains aspects de la thématique complexe de ma thèse je suis allée chercher des compétences au sein de mon équipe, mais aussi en dehors.

Pour le premier axe de ma thèse (étude de simulation) j'ai consulté des experts en génétique des populations pour élaborer des scenarios pertinents. Pour résoudre des problèmes issus de la complexité de mes scénarios j'ai contacté l'auteur du logiciel de simulation. Suite à nos interactions régulières dans la troisième année de ma thèse nous avons mis en place une collaboration dans laquelle j'ai contribué à tester et développer la nouvelle version du logiciel.

Les simulations sont capables de produire de grandes quantités de données en relativement peu de temps. La gestion et l'analyse de ces données demande une excellente organisation et des outils optimisés. Pour améliorer mes outils d'analyse j'ai trouvé des collaborateurs dans le domaine de la bioinformatique – qui m'était inconnue avant ma thèse.

Le deuxième axe de ma thèse s'appuie sur des données issues en partie d'un autre projet (ANR DL-Vitis) et en partie des données que nous produisons dans le cadre du projet CASDAR. Dans le but d'organiser, optimiser et suivre les travaux et les analyses en commun en respectant la valorisation des résultats propres à chaque projet, j'ai mis en place des réunions et j'ai rédigé des rapports d'étapes.

Dans la deuxième année de ma thèse je me suis chargée de l'organisation de la récolte et des analyses sur le raisin frais. C'est une période de pointe, car en 2 mois il faut traiter près de 1200 échantillons en respectant un protocole complexe. Vu la quantité de travail, des saisonniers ont été embauchés et j'ai réussi à mobiliser la majorité de mon équipe pour diviser les charges. J'ai organisé plusieurs équipes pour repartir les tâches. J'ai coordonné leur travail entre deux sites différents (terrain et laboratoire) en veillant sur les contraintes météorologiques et le temps de manipulation car les échantillons s'abiment s'ils ne sont pas traités dans un délai limité.

Dans le cadre d'une collaboration avec la plateforme de phénotypage haut-débit de Bordeaux (Métabolome) j'ai effectué un séjour de 4 semaines à l'INRA de Bordeaux, pour analyser la qualité des baies de raisin en travaillant avec des biochimistes.

Estimation et prise en charge du coût du projet

ESTIMATION DU COUT CONSOLIDE DE LA THESE

Montants en euros TTC

	Nature de la dépense	Détails *		Coûts totaux (euros TTC)			
				Nombre d'unités	Coût unitaire moyen	Quote-part utilisation	Total
1	Ressources Humaines	Salaire brut	Charges				
1.1	Doctorant	2197	1161	36	3358	100%	120888
1.2	Encadrant 1	5073	3479	1,5	8552	100%	12828
1.3	Encadrant 2	3192	1686	3	4878	100%	14634
1.4	Autre personnel (hors sous-traitance)						
	8 techniciens de recherche (INRA)	12592	6208	4	18800	50%	37600
	1 main d'œuvres (INRA)	1532	644	4	2176	50%	4352
	1 chargé de recherche (INRA)	2884	1327	3	4211	50%	6316
	1 ingénieur de recherche (INRA)	2657	1223	2	3880	50%	3880
	1 ingénieur d'étude (INRA)	1995	982	1	2977	100%	2977
	2 technicien IFV	5796	3062	1	8858	100%	8858
	2 stagiaires IFV	1250	660	5	1910	100%	9550
	autres personnes IFV (entretien du matériel végétal)	2898	1531	9	4429	50%	19930
1.5	Sous-traitance						
	mesure delta C13			1	26910	100%	26910
	Reséquençage Genotoul (10 lanes)			10	2860	100%	28600
	génotypage Illumina puce SNP 20K			1	40000	33%	13200
	Sous-total Ressources Humaines						240307
2	Consommables	Détails					
2.1	Fournitures expérimentales	manip GBS sur 300 individus		300	33,34	100%	10000
		manip sur jus de raisin		1	1100	100%	1100
2.2	Fournitures de bureau			3 ans	80	100%	240
	Sous-total Consommables						11340
3	Infrastructures	Loyer total par ans					

3.1	Loyers, entretien et charges des locaux	4000		3	4000		12000
3.2	Loyer d'une serre	5000		6 mois	5000		2500
	Sous-total Infrastructures						14500
4	Matériel (amortissements)	Taux d'amortissement par ans				% dédié à mon projet	
4.1	Matériel d'expérimentation (dont les appareils, ordinateurs et logiciels spécialisés)	10%		2 ans	30000	33%	2000
4.2	Ordinateur de bureau	33%		3 ans	100	100%	100
4.3	Logiciels de bureau	33%		3 ans	150	100%	150
4.4	Matériel informatique avec entretien (serveur de calcule)	33%		1,5 an	20000	100%	10000
4.5	Chambre de culture	10%		3 ans	30000	50%	4500
	Sous-total Matériel						16750
5	Déplacements	Transport	Hébergement + autres frais				
5.1	Missions en France	300	1000	par ans	1300		3900
	Sous-total Déplacements						3900
6	Formation	Détails					
6.1	Formations	Ecole chercheur		2	400		800
		Autres modules de formation		3	350		1050
6.2	Autres frais (Inscription à l'Université, Sécurité Sociale étudiante, etc.)	Inscription		3	370		1110
	Sous-total Formation						2960
7	Documentation et communication	Détails					
7.1	Affranchissements, Internet, téléphone			Par ans	100		300
7.2	Frais de publication et	Publication d'article scientifique		2	1600		3200
		Relecture en anglais		2	200		400
	Sous-total Documentation et communication						3900
8	Charges financières (intérêts des emprunts)						0

	Sous-total Charges financières					0
9	Charges exceptionnelles					0
	Sous-total Charges exceptionnelles					0
10	TOTAL	293675				

Compétences, savoir-faire, qualités professionnelles et personnelles illustrées par des exemples

Mission principale : Mettre en place et déterminer les limites et les opportunités offertes par une méthode de sélection à la pointe de la technologie (sélection génomique) chez la vigne pour contribuer à un programme de création variétale novateur.

Mission 1. Effectuer une veille scientifique

- Participer à plusieurs séminaires et formations correspondants à mes questions de recherche.
- Repérer, analyser et comprendre les communications scientifiques liées à mes axes de recherche. (Notamment sur les concepts, les outils et méthodologies existants, et sur les résultats des études similaires à la mienne.)

Qualités et compétences développées : **Synthèse et transfert** des enseignements dans mon contexte de travail et adaptation à ma problématique de thèse.

Mission 2. Mettre en place les scénarios pour simuler le matériel végétal

- Choisir et apprendre à maitriser le logiciel de simulation.
- Etudier l'histoire évolutive de la vigne et consulter les spécialistes de ce domaine.
- Implémenter et évaluer les différents scénarios.
- Optimiser la simulation en fonction des paramètres connus.
- Mettre en place une collaboration pour améliorer le logiciel de simulation.

Mission 3. Mettre en place les outils d'analyse informatique

- Définir des étapes d'analyse nécessaires.
- Créer ou adapter les outils d'analyses pertinents.
- Vérifier le bon fonctionnement des outils.
- Mettre en place une collaboration avec un bio-informaticien pour améliorer les performances et la généricité des outils créés.

Qualités et compétences développées : **Autonomie** dans le domaine de la simulation et de la sélection génomique. **Démarchage de collaborateurs potentiels** pour renforcer les compétences indispensables pour mon projet. **Explications simples de mon sujet** aux collaborateurs. **Argumentation** sur la pertinence de mes choix (logiciel, paramètres) en **restant à l'écoute** des remarques et des propositions des collaborateurs. Synthèse et intégration des nouvelles informations à mon travail. **Gestion d'importants volumes de données** avec les outils d'analyse

appropriés. **Organisation des expérimentations** et **estimation du temps** d'analyse pour respecter le timing. **Communication** sur l'avancement du projet et des nouveaux résultats aux collaborateurs et à mes supérieurs.

Mission 4. Organiser et coordonner la récolte et la production des données phénotypiques

- Développer des protocoles d'analyse.
- Planifier et contribuer à la réalisation des analyses en laboratoire.
- Organiser et conduire la récolte
- Prévoir et gérer les personnes disponibles pour les travaux.
- Tenir informer les responsables des autres projets sur le déroulement des travaux.

Qualités et compétences développées : **Planification et organisation** des expérimentations sur le terrain en fonction des moyens disponibles et des contraintes externes (personnes disponibles, maturité de la vigne, traitement phytosanitaires sur les parcelles voisines, conditions météorologiques). **Définition des objectifs**, organisation logistique entre le terrain et le laboratoire, **gestion de priorités et d'imprévus**. **Encadrement et suivi de saisonniers** et des équipes de techniciens. **Etablissement et respect des protocoles** de travail. **Communication** avec l'équipe de récolte et information des responsables des autres projets sur l'avancement des travaux.

Mission 5. Contribuer au développement d'un protocole de génotypage par séquençage (GBS) adapté à la vigne

- Rechercher de l'information bibliographique sur les techniques à mettre en place.
- Participer à la mise en place des tests en laboratoire.

Qualités et compétences développées : Organisation et mise en place des tests en laboratoire. Planification des expérimentations et travail pour optimiser les méthodes utilisées.

Mission 6. Analyser et interpréter les données

- Classer et comparer les résultats des analyses.
- Rédiger des rapports d'étape sur les résultats.
- Organiser et animer des réunions avec mes supérieurs sur l'avancement des analyses. Discussion sur les points problématiques.
- Solliciter des spécialistes des différents domaines (génétique quantitative, bio-statistiques) pour contribuer à la réflexion.
- Prendre des décisions sur la direction à prendre et définition de deadline pour les futures analyses.

Qualités et compétences développées : **Extraction des données exploitables** à partir des résultats bruts. **Analyse critique** des résultats, organisation et **présentation à l'écrit et à l'oral**. **Organisation et animation des réunions** scientifiques avec mes supérieurs et d'autres experts. **Interprétation des résultats** en tenant compte des remarques et propositions des participants et prise de **décisions** pour la suite du projet.

Mission 7. Communiquer les résultats de la thèse à l'oral et à l'écrit.

- Définir et structurer le contenu de la communication avec mes supérieurs.
- Interagir avec mes supérieurs et les collaborateurs pour améliorer le manuscrit.
- Préparer un exposé oral et soutenir les résultats de ma thèse devant un jury spécialisé dans mon domaine de recherche.

Qualités et compétences développées : Organisation et **présentation des travaux et résultats** d'un projet de 3 ans. **Evaluation de l'importance** des résultats. **Synthèse** des résultats d'analyses pour en **dégager des messages** clairs.

Résultats scientifiques et techniques :

- Etudes de génétique d'association (GWAS), et mise en place des modèles de sélection génomique (GS) chez la vigne.
- Phénotypage du matériel végétal sur le terrain et en laboratoire sur 2 années. Analyses statistiques des données.
- Développement d'un pipeline d'analyse génétique (GS et GWAS) haut débit.
- Mise en place de protocoles de génotypage par séquençage (GBS) et détection de SNP.

Résultats en gestion et pilotage de projet :

- Coordination et réalisation de campagnes d'analyses.
- Organisation, structuration et traitement d'importants jeux de données.
- Communication : rédaction de rapports d'avancement, de la thèse et d'articles scientifiques en anglais. Organisation et animation de réunions.

Résultats relationnels et managériaux :

- Mise en place et coordination de collaborations interdisciplinaires.
- Développement d'un réseau d'experts scientifiques de différents domaines à l'échelle nationale et internationale.
- Encadrement de techniciens et de saisonniers.

Résultats, impact de la thèse

Ma thèse est une première étude sur la faisabilité et l'application de la prédiction génomique chez la vigne. Les connaissances acquises seront publiées dans des journaux scientifiques et serviront de base aux futurs projets au sein de laboratoire (un chargé de recherche va être recruté pour poursuivre mes analyses).

J'ai développé des outils d'analyses qui permettent de réaliser des prédictions génomiques et de la génétique d'association sur tout le génome à la fois pour des données simulées et réelles. En les rendant plus accessibles aux utilisateurs, ils représenteront un appui précieux aux programmes de sélection. Cet outil peut aider à choisir la stratégie la mieux adaptée pour répondre à l'objectif du programme compte tenu des moyens engagés. Le choix optimal permet de conduire un programme plus efficient en économisant du temps et de l'argent.

J'ai contribué à la définition d'un protocole de génotypage par séquençage (GBS) appliqué à la vigne. Cette méthode est capable d'assurer un marquage dense sur tout le génome par une technique relativement simple et bon marché.

Cette thèse m'a aidé à élargir et à approfondir mes connaissances scientifiques et techniques dans le domaine de l'amélioration des plantes (et plus précisément de la vigne, difficile à manipuler de par sa nature pérenne). Je la considère comme un premier pas dans le monde de la recherche y compris une première expérience de gestion de projet scientifique long (sur 3 ans). Enfin, ces trois ans de thèse m'ont apporté des réflexions précieuses pour la construction de mon projet professionnel en me faisant découvrir et développer des nouvelles compétences.

Identifications de pistes professionnelles

Durant ma thèse j'ai développé une **expertise dans un domaine pointu** et j'ai construit mon réseau professionnel en mettant en place des **collaborations** interdisciplinaires. Un poste de « **chargé de recherche** » dans un institut public me permettrait de valoriser ces compétences en mettant en place des projets scientifiques fondamentaux ou appliqués entre plusieurs partenaires.

Je suis une personne très **curieuse**, je suis née en Hongrie et c'est la découverte d'une autre culture et une autre langue qui m'ont poussée à venir en France et à poursuivre mes études en thèse. Pendant la réalisation de mon projet de thèse je me suis découvert une grande affinité pour la **gestion scientifique** et le **management d'équipe**. Aujourd'hui j'aimerais utiliser et développer ces types de compétences, dans le domaine de mon expertise professionnelle, la génétique et

l'amélioration des plantes, à travers d'un poste de « **chef de projet R et D en amélioration des plantes**» dans une entreprise.

Résumé

L'ambition de cette thèse était de proposer un nouvel élan pour la création variétale chez la vigne, incluant les connaissances et les derniers outils de la recherche. En effet, la viticulture française comme d'autres filières agricoles doit aujourd'hui faire face à 3 grands défis: la réduction des intrants phytosanitaires (plan Ecophyto 2018), les changements climatiques, et de nouveaux concurrents, notamment les pays du Nouveau Monde. La création variétale, qui a été peu exploitée chez la vigne, peut être une des solutions pour répondre à ces défis.

S'appuyant sur un génotypage dense, plusieurs outils et concepts innovants – réunis sous le terme de sélection génomique (GS) – ont vu le jour ces dernières années en sélection animale, qui permettent de prédire les phénotypes des individus seulement génotypés.

Afin d'atteindre notre but, nous avons évalué et comparé l'efficacité de la GS et de la sélection assistée par marqueur (SAM) « classique », basée sur la génétique d'association « genome-wide » (GWAS) chez la vigne. Le potentiel théorique des deux méthodes a été évalué dans une étude de simulation, puis sur des données réelles.

Nous montrons que la GS est plus pertinente que la SAM « classique » pour prédire les phénotypes et ce pour des caractères complexes et / ou structurés. Cependant la GS couplée aux résultats issus de la GWAS semble être une méthode intéressante lorsque le marquage moléculaire est non limitant. Finalement, nous discutons des conditions d'utilisation de la GS en termes économiques et d'efficacité au cours du temps. Nous proposons trois scénarios fonction de l'investissement de départ et des besoins en termes de création variétale.

Abstract

The aim of this PhD project was to provide a new impulse for grapevine breeding, applying the latest knowledge and research tools on this species. Indeed, French viticulture, as well as various other agricultural sectors, faces today three major challenges: how to reduce phytosanitary inputs (Ecophyto 2018), impact of climatic changes and new competitors on the market, especially New World countries. Plant breeding in grapevine has not been much exploited until today, but could be a solution to these challenges.

Several innovative tools and concepts have seen the light in animal breeding in the last decade. Using high density genome-wide marker information and advanced statistical models, phenotypes can be predicted for individuals that were merely genotyped. The method termed genomic selection (GS) is implementing this type of approach.

To achieve our aim, we evaluated and compared the efficiency of GS and "classical" marker-assisted selection (MAS), based on genome-wide association study (GWAS) for grapevine. The theoretical potential of the two methods was evaluated in a simulation study but also on real data.

We show that GS is more relevant than "classical" MAS to predict phenotypes of complex and / or structured traits. However, the combination of GS with results from GWAS seems to be of particular interest if the number of molecular markers available is adequate. Finally, we discuss GS implementation in terms of economic aspects and efficiency over time; we propose three scenarios differing by the initial investment required and the breeding objectives to be reached.