Contribution des stratégies de sélection génomique et phénomique aux programmes d'amélioration génétique de l'abricotier (Prunus armeniaca L.) pour quelques traits d'intérêt

Mariem Nsibi

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'INSTITUT D'ETUDES SUPERIEURES AGRONOMIQUES DE MONTPELLIER MONTPELLIER SUPAGRO

En Génétique et Amélioration des Plantes (Filière : Biologie intégrative, Diversité et Amélioration des Plantes)

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau
Portée par l'Université de Montpellier

Contribution des stratégies de sélection génomique et phénomique aux programmes d'amélioration génétique de l'abricotier (*Prunus armeniaca* L.) pour quelques traits d'intérêt

Présentée par Mariem NSIBI
Le 9 juin 2021

Sous la direction de Jean-Luc REGNARD

Devant le jury composé de

David CHAGNÉ, Science Group Leader, Plant & Food Research, New Zealand	Rapporteur
Eric DUCHÊNE, Ingénieur de Recherche, INRAE Colmar	Rapporteur
Jacques DAVID, Professeur, L'Institut Agro | Montpellier SupAgro	Président du jury
Elisabeth DIRLEWANGER, Directrice de Recherche, INRAE Bordeaux	Examinatrice
Laurence MOREAU, Directrice de Recherche, INRAE Gif s/Yvette	Examinatrice
Laurent BOUFFIER, Chargé de Recherche, INRAE Pierroton	Examinateur
Jean-Marc AUDERGON, Ingénieur de Recherche, INRAE Avignon	Co-encadrant
Jean-Luc REGNARD, Professeur, L'Institut Agro | Montpellier SupAgro	Directeur de thèse
David POT, Chercheur CIRAD, Montpellier	Invité
Christopher SAUVAGE, Responsable de Projet Génétique, Syngenta SAS France	Co-encadrant, invité

Remerciements

Tout d'abord, j'adresse mes remerciements à messieurs Eric Duchêne et David Chagné pour l'honneur que vous m'avez fait en acceptant d'être rapporteurs de ma thèse. Je tiens également à remercier tous les membres du jury d'avoir accepté d'évaluer ce travail.

Je remercie cordialement mon directeur de thèse Jean-Luc Regnard et mes encadrants Jean-Marc Audergon et Christopher Sauvage qui m'ont accompagnée lors de cette expérience enrichissante tant sur le plan professionnel que personnel. Je les remercie très sincèrement pour la confiance et la patience qu'ils m'ont prodiguées ainsi que pour leurs efforts qui ont contribué amplement à l'élaboration de ce projet de thèse dans les meilleures conditions possibles. Je leur en suis infiniment reconnaissante. Outre leurs encouragements et conseils avisés, grâce à eux, j'ai eu l'opportunité de rencontrer de nombreux chercheurs dans le domaine de la biologie, l'économie et de la génétique quantitative, notamment ceux du réseau R2D2 qui œuvrent, par leurs connaissances et savoir-faire, à relever les défis posés à l'agriculture.

Je voudrais également exprimer ma reconnaissance à tous les membres de mes comités de thèse pour l'intérêt porté à mon travail et leurs précieux conseils. En particulier, je tiens à remercier Brigitte Mangin qui m'a accueillie dans son équipe et m'a transmis de précieuses connaissances sur la modélisation de la sélection génomique. Par la même occasion, je remercie Prune Pegot-Espagnet, Pauline Duriez et leurs collègues de bureau pour leur chaleureux accueil, et les précieux moments partagés lors des pauses café, déjeuners et les week-ends. Vous avez rendu mon séjour à Toulouse extrêmement agréable !

Mes remerciements s'adressent également aux différentes équipes scientifiques, techniques et administratives INRAE GAFL et notamment l'équipe DADI animée par Mathilde Causse et Bénédicte Quilot-Turion. Je remercie toutes les personnes qui se sont intéressées à la réalisation de mon travail notamment Joël Chadoeuf, Patrick Lambert, Frédérique Bitton, Jacques Lagnel, Morgane Roth, Emmanuel Le Calonnec, …

Je voudrais également remercier tous les membres de l'équipe Prunus pour leur contribution précieuse aux expérimentations sur terrain et au laboratoire. Je remercie particulièrement Guillaume Roch, Jean Leonetti, Eric Martin, Sabrina Viret, …

Je remercie également Sylvie Bureau et Barbara Gouble pour leur contribution précieuse à la réalisation de mon travail, notamment la partie liée à la spectroscopie infrarouge. Merci d'avoir

partagé vos connaissances avec moi depuis le début de mon stage de master. Je vous adresse toute ma reconnaissance !

Je remercie tous les stagiaires et doctorants que j'ai croisés lors de mon séjour au GAFL : Mariem Omrani, Isidore Diouf, Pierre Sadon, Aimeric Agaoua, Hussein Kanso, Typhaine Briand, Jiantao Zhao, Séverine Monnot, Christina Thenault, Estelle Bineau, Delyan Zafirov…

Je remercie tout particulièrement Audrey Ferro ma collègue de bureau durant mes six premiers mois au GAFL avec qui j'ai passé des moments agréables à Avignon et à Saint-Rémy-de-Provence. Merci pour ta bonne humeur et ta joie de vivre !

Je remercie également mon amie et compatriote Chaima Bengagi pour ton amitié et tes qualités humaines. Je garderai un excellent souvenir de notre voyage à Barcelone !

Ma gratitude s'adresse aux amis que j'ai eu la chance de rencontrer lors des journées dédiées aux jeunes chercheurs du département de Biologie et Amélioration des Plantes INRAE, et lors des formations doctorales à l'université de Montpellier, des séminaires et conférences, pour leurs partages d'expériences de recherche et les échanges instructifs et enrichissants sur nos domaines respectifs.

Je voudrais remercier tous les enseignants que j'ai côtoyés tout au long de mon cursus scolaire et universitaire pour tout ce qu'ils m'ont appris. Mes remerciements s'adressent également à tous mes professeurs de l'Institut National Agronomique de Tunisie qui m'ont initiée au monde de l'agronomie et de l'amélioration des plantes. Qu'il me soit permis de vous présenter à travers ce travail le témoignage de mon grand respect et l'expression de ma profonde reconnaissance. Je remercie également toute l'équipe enseignante de Montpellier SupAgro pour m'avoir transmis la passion de la génétique des plantes. Je leur en suis profondément reconnaissante.

Je remercie aussi mes camarades de la promotion APIMET – SEPMET (2016-17) pour les bons moments partagés notamment lors de nos visites et notre voyage d'études. Je vous souhaite une bonne continuation dans vos vies personnelles et professionnelles !

Enfin, je ne pourrais finir sans adresser mes remerciements les plus chaleureux à tous les membres de ma famille pour leur soutien moral et matériel et leurs encouragements sans faille qui m'ont accompagnée tout au long de ces années. Je les remercie de leur affection, amour et soutien constants qui m'ont été d'un grand réconfort. En espérant qu'ils puissent trouver dans ce travail le fruit de leur sacrifice et dévouement.

Ce travail n’aurait pas été possible sans le soutien du ministère de l’enseignement supérieur et de la recherche scientifique de la Tunisie, l’unité de recherche ‘Génétique et Amélioration des Fruits et Légumes’ et CEP Innovation, qui m’ont attribué une allocation de recherche pour mener ce travail à terme.

Productions scientifiques

a) Publications

a) NSIBI M., GOUBLE B., BUREAU S., FLUTRE T., SAUVAGE C., AUDERGON J.M., REGNARD J.L., 2020. Adoption and optimization of genomic selection to sustain breeding for apricot fruit quality. G3 Genes|Genomes|Genetics, 10(12) : 4513-4529. https://doi.org/10.1534/g3.120.401452

b) Communications en conférences et symposium scientifiques internationaux

a) NSIBI M., CONFOLENT C., GOUBLE B., BUREAU S., BLANC A., ROCH G., LAMBERT P., FLUTRE T., REGNARD JL., SAUVAGE C., AUDERGON JM., 2018. Genomic selection - Which prospects in *Prunus armeniaca*? Preliminary results issued for fruit quality traits and phenology. 9th International Rosaceae Genomics Conference (RGC9), Nanjing, China (June 2018). ⟨hal-01830311⟩

b) NSIBI M., GOUBLE B., BUREAU S., FLUTRE T., REGNARD JL., SAUVAGE C., AUDERGON JM., 2019. Genomic selection – Assessment of genomic selection accuracy on fruit quality traits in apricot. XVIIth International Symposium on Apricot Breeding and Culture, Malatya (Turkey), 06-10 July 2019. ISHS Award de la meilleure présentation.

c) NSIBI M., GOUBLE B., BUREAU S., BLANC A., ROCH G., LAMBERT P., FLUTRE T., REGNARD JL., SAUVAGE C., AUDERGON JM. 2020. Genomic and phenomic selection - Which prospects in *Prunus armeniaca* ? 10th International Rosaceae Genomics Conference (RGC 10, Virtual 2020) December 2020.

c) Présentations et séminaires en France :

a. NSIBI M., CONFOLENT C., LAMBERT P., BUREAU S., GOUBLE B., BLANC A., ROCH G., OMRANI M., FLUTRE T., SAUVAGE C., AUDERGON JM. Sélection génomique - Quelles perspectives chez les Prunus ? Exemple de l'abricotier en adossement au projet FruitSelGen. *Séminaire SelGen 2017 : "La sélection génomique, bilan et perspectives"*, Sept. 2017, Paris, France. 1 p. ⟨hal-02789979⟩

b. NSIBI M., CONFOLENT C., GOUBLE B., BUREAU S., BLANC A., ROCH G., LAMBERT P., FLUTRE T., REGNARD JL., SAUVAGE C., AUDERGON JM., Genomic selection – Which prospects in *Prunus armeniaca* ? Preliminary results issued from fruit quality traits. Poster, Journées Jeunes Chercheurs du département BAP-INRAE, Sept. 2018, Paris, France, 1 p. https://journees.inrae.fr/jjc-bap-2018/

c. NSIBI M. Assessment of genomic selection accuracy on fruit quality traits in apricot. Séminaire présenté à l'Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT-INRAE), le 05 mars 2019.

d. NSIBI M. Assessment of genomic selection accuracy on fruit quality traits in apricot. Séminaire interne GAFL-INRAE, Avignon, le 12 avril 2019, et Genomic selection - which prospects in *Prunus armeniaca*?

e. NSIBI M. Assessment of genomic selection accuracy on fruit quality traits in apricot. Journées Jeunes Chercheurs du département BAP-INRAE, Avignon, 13-14 juin 2019. https://journees.inrae.fr/jjcbap2019-avignon/Program2

NSIBI M. GOUBLE B., BUREAU S., FLUTRE T., REGNARD JL., SAUVAGE C., AUDERGON JM., Assessment of two selection strategies: genomic selection and phenomic selection in order to breed for improved fruit quality in apricot. Séminaire aux Journées du réseau Selgen R2D2 INRAE, Sélection génomique, 15-17 oct. 2019, Obernai, France.

Résumé

L'adossement de l'amélioration génétique des espèces animales et végétales aux progrès technologiques a permis l'avènement de stratégies récentes de sélection s'appuyant sur des données à haut-débit. Le progrès génétique résultant, sous-tendu par l'étude de l'architecture génétique des caractères d'intérêt agronomique, permet désormais un changement de paradigme pour substituer à une démarche de caractérisation des candidats à la sélection une démarche de prédiction de leur réponse à la variation environnementale.

Ainsi, la sélection génomique (SG) fondée sur la contribution de plusieurs dizaines de milliers de marqueurs couvrant tout le génome a révolutionné les programmes de sélection, en surmontant le problème de l'héritabilité manquante rencontré dans les stratégies de sélection assistées par marqueurs (SAM) qui s'appuient sur les locus impliqués dans la variation de caractères quantitatifs (QTL), préalablement détectés par analyse de liaison génétique ou par des études d'association pangénomiques (GWAS). Certaines limites de la SG ont récemment ouvert la voie à la sélection phénomique (SP) qui, en se basant sur la spectroscopie infrarouge, semble offrir une alternative potentiellement aussi efficace et moins coûteuse que la SG.

Dans cette perspective, cette thèse a pour objectif d'évaluer le potentiel de ces avancées scientifiques et technologiques dans les programmes de sélection de l'abricotier *Prunus armeniaca* L. Chez cette espèce fruitière, les méthodes conventionnelles se heurtent à diverses limites biologiques et à la durée des cycles de sélection, qui freinent le progrès génétique. En corollaire, le recours aux données à large échelle s'inscrit dans l'ambition d'accélérer le gain génétique par unité de temps et de créer des variétés performantes afin de répondre aux attentes des différents acteurs de la filière abricot. Dans cette perspective, nous avons évalué la performance des modèles de SG et SP en termes de précision de prédiction des caractères agronomiques. Ce travail s'est appuyé sur une population biparentale en pseudo-F1 (Goldrich × Moniqui) et sur une collection de ressources génétiques présentant un large panel de diversité, pour laquelle de nombreuses données phénotypiques étaient disponibles.

Pour répondre aux objectifs de la thèse, une stratégie de validation croisée a été utilisée afin de prédire des caractères d'intérêt tels que la qualité des fruits, la phénologie (dates de floraison et de maturité) et la sensibilité aux maladies. Nos résultats ont souligné l'intérêt des approches de SG et de SP pour cet ensemble de caractères présentant des architectures génétiques contrastées. Également, nous avons proposé des scénarios d'optimisation des approches de SG et SP en nous appuyant sur la pondération des modèles de prédiction par le biais de la valorisation de l'information apportée par l'architecture génétique des caractères sous investigation. De surcroît, nous avons évalué une stratégie de prédiction multivariée reposant sur des caractères proxies pour prédire des caractères focaux coûteux et difficiles à mesurer. L'implémentation pratique de ces approches dans les programmes de sélection est discutée et les contributions respectives de la SG et la SP sont évaluées.

Abstract

Harnessing technological breakthroughs in data acquisition to genetic improvement of animal and crop species has allowed the advent of novel selection strategies grounded on high-throughput data. The genetic progress issued from unravelling the genetic architecture of agronomically relevant traits resulted in a paradigm shift from extensive characterization of selection candidates to a prediction strategy of their response to environmental variation.

Thus, genomic selection (GS) based on the contribution of several tens of thousands of markers covering the entire genome has revolutionized breeding programs. It has overcome the problem of missing heritability encountered in marker-assisted selection (MAS), which relies on minimal fractions of genetic variance accounted for by a few quantitative trait loci (QTL), identified by linkage analysis or genome-wide association studies (GWAS). The constraints of GS, notably linked to the missing heritability unaccounted for by MAS, have recently paved the way for a novel selection strategy denoted as phenomic selection (PS), which is based on infrared spectroscopy and seems to offer a potentially valuable alternative to SG.

In this perspective, this PhD project aims at evaluating the potential of the recourse to high throughput data in favor of the genetic improvement of apricot tree, *Prunus armeniaca* L., for which conventional breeding is hindered by various biological limitations and long selection cycles and thus tends to impede the genetic progress. Therefore, the drive behind the use of large-scale data is to hasten the genetic gain per unit time and create genetically superior varieties with improved genetic constitution in order to meet the expectations and needs of apricot sector stakeholders.

Within this context, we evaluated the performance of GS and PS models in terms of prediction accuracy for different agronomic traits within a biparental pseudo-testcross population (Goldrich × Moniqui) and a collection of genetic resources encompassing a wide range of diversity.

To this end, a cross-validation strategy was performed to predict a panel of traits of agronomic interest such as fruit quality, phenology (flowering and maturity dates) and disease susceptibility. Our findings highlighted the efficiency of GS and PS approaches for several traits with contrasting quantitative genetic architectures. Furthermore, we proposed scenarios for optimizing the prediction accuracy through weighting GS and PS models using the information provided by the genetic architecture of these traits. In addition, we assessed a multivariate modelling approach where the emphasis was placed on proxy traits in order to predict costly and difficult-to-measure target traits such as ethylene production. The potential place of both strategies within apricot breeding programs is discussed and their respective contributions to selection decisions is evaluated.

Table des matières

Liste des tableaux

Liste des figures (hors articles et annexes)

Liste des abréviations

A : valeur génétique additive

ANOVA : analyse de la variance

ATR : Attenuated total reflectance / Réflectance totale atténuée

BIC : Bayesian information criterion / Critère d'information bayésien

BL : Bayesian LASSO / LB: LASSO bayésien

BLUP: Best Linear Unbiased Prediction / Meilleure prédiction linéaire sans biais

B.rot : Scoring of brown rot blossom blight (*Monilinia laxa*) / Notation de la sensibilité au Monilia sur fleur

BRR : Bayesian Ridge Regression / Régression Ridge bayésienne

Canker : Scoring of bacterial canker (*Pseudomonas syringae*) / Notation de la sensibilité au chancre bactérien

CEP : Centre d'Expérimentation de Pépinières

Chr : Chromosome

CIE : Commission Internationale de l'Eclairage

CIM : Composite Interval Mapping / Cartographie d'intervalle composite

Citric.A : Citric acid / Teneur en acide citrique

cM : centimorgan

Ctifl : Centre Technique Interprofessionnel des Fruits et Légumes

CTPS : Comité Technique Permanent de la Sélection de plantes cultivées

D : valeur de dominance

DHS : Distinction, Homogénéité, Stabilité

DNA: desoxyribonucleic acid / ADN : acide désoxyribonucléique

DTGS : Deuterated triglycine sulphate

E : Environment

eBIC : extended Bayesian Information Criterion

ECA : Enroulement chlorotique de l'abricotier (phytoplasme)

FAO : Food and Agriculture Organization of the United Nations

F.weight : Fruit individual weight / Poids individuel du fruit

ΔG : progrès génétique par cycle de sélection

G : Genotype

GAFL : Unité de recherche Génétique et Amélioration des Fruits et Légumes

GBLUP : Genomic best linear unbiased prediction

GBS : Genotyping By Sequencing / Génotypage par séquençage

GDR : Genome Database for Rosaceae

GEBV : GenomicEstimated Breeding Value / Valeur génétique additive génomique

GEVES : Groupe d'Etude et de contrôle des Variétés Et des Semences

Go×Mo : Goldrich×Moniqui pseudo F1 progeny

GS : Genomic Selection / SG : sélection génomique

GWA(S) : Genome-Wide Association (Study) / (Etude d')association pangénomique

H^2 : Broad-sense heritability / Héritabilité au sens large

h^2 : Narrow-sense heritability / Héritabilité au sens étroit

Hue.g : Hue angle of ground colour according to CIE 1976 L*a*b* color space (fruit epidermis, non-blushed side) / couleur de fond de l'épiderme

H-W : Hardy-Weinberg

I : Interaction

IBD : Identity-By-Descent / identité par descendance

INRAE : Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement

kPa : kilopascal

LASSO : Least Absolute Shrinkage and Selection Operator

LD : Linkage Disequilibrium / DL : déséquilibre de liaison

LG : Linkage Group / GL : Groupe de liaison

LOD : Logarithm of the Odds Ratio

MAF : Minor Allele Frequency / Fréquence des allèles mineurs

Malic.A : Malic acid / Teneur en acide malique

MAS : Marker-Assisted Selection / SAM : sélection assistée par marqueurs

Mb : Million base pairs / Million de paires de bases

MCMC : Markov Chain Monte Carlo

Meq : milliequivalent

MIR(S) : Mid-Infrared (Spectroscopy) / (Spectroscopie) moyen infrarouge

MLMM : Multi-locus mixed model / modèle mixte multi-locus

MVN : Multivariate normal

NIR(S) : Near-Infrared (Spectroscopy) / PIR(S): (Spectroscopie) proche infrarouge

NMR : nuclear magnetic resonance / RMN : résonance magnétique nucléaire

OP : Organisation de producteurs (AOP : Association d'OP)

P : Phenotype

PA : Prediction Accuracy / précision de prédiction

PCA : Principal Component Analysis / ACP : Analyse en Composantes Principales

PCR : Polymerase Chain Reaction

P.mildew : scoring of powdery mildew disease (Sphaerotheca pannosa) / notation de la sensibilité à l'oïdium

PPV : Plum Pox Potyvirus (virus de la sharka)

PS : Phenomic Selection / SP : sélection phénomique

QTL (QTLs) : Quantitative Trait Locus (Loci)

RI : Refractive Index / IR : indice réfractométrique

RNA: RiboNucleic Acid / ARN : acide ribonucléique

RR-BLUP : Ridge Regression Best Linear Unbiased Prediction - Régression Ridge

Rust : Scoring of rust disease (Tranzschelia spp.) / Sensibilité à la rouille

SEFRA : Station Expérimentale FRuits d'Auvergne-Rhône-Alpes

SERFEL :

SNES : Station Nationale d'Essais de Semences (structure du GEVES)

SNP : Single Nucleotide Polymorphism

SQPOV : UMR Sécurité et Qualité des Produits d'Origine Végétale (INRAE – Avignon Université)

SSC : Solid soluble content / TSS : Teneur en matière sèche soluble

TA : Titratable Acidity / AT : acidité titrable

TP : Training Population / population d'entraînement

UERI : Unité Expérimentale Recherche Intégrée INRAE Gotheron

UMR : Unité Mixte de Recherche

INTRODUCTION : Combler le gap génotype – phénotype au profit de la sélection variétale

En sélection variétale, la cible de la génétique quantitative est la hiérarchisation des candidats à la sélection et l'identification des individus dotés des meilleures valeurs génétiques et ayant été évalués dans une gamme environnementale diversifiée. Les candidats présentant une performance agronomique supérieure en termes de caractères d'intérêt ciblés par la sélection et répondant aux différents enjeux des filières agronomiques pourront servir de géniteurs pour des croisements ultérieurs.

L'amélioration génétique des plantes opère dans un contexte dynamique caractérisé par des progrès technologiques dont le socle reste très largement ancré sur les ressources génétiques disponibles. Ces progrès s'expriment au plan théorique par la connaissance des traits d'intérêt, la maitrise de leur hérédité et de la manière de les assembler dans des prototypes. Ils sont concrétisés au plan pratique par le développement d'outils et méthodes fiables, robustes, susceptibles d'être déployées sur de larges effectifs à moindre coût. Dans ce contexte, l'amélioration génétique des plantes bénéficie des approches classiques telles que l'échantillonnage mendélien et l'accouplement aléatoire sur lesquels se basent les hypothèses des modèles mixtes employés dans le cadre de la sélection dans l'optique de maximiser le gain génétique et contourner la perte de la diversité au sein des populations candidates. En effet, le concept fondamental de la génétique quantitative repose sur la multiplicité de locus contribuant à la variation continue des caractères à laquelle s'ajoutent la variation issue de l'interférence de l'environnement et de l'interaction de ce dernier avec la gamme variétale évaluée. Il s'agit de la conception de Fisher par rapport à la variation continue des caractères quantitatifs qui a réconcilié les mendélistes et les biométriciens. En génétique mendélienne, le modèle de l'hérédité se base sur des classes phénotypiques distinctes alors qu'en biométrie le phénotype représente un continuum d'observations (Fisher 1918, 1941). Dans ce concept en faveur de la polygénicité du déterminisme des caractères d'intérêt, on relève le fondement de la génétique quantitative qui, sous l'hypothèse d'un accouplement au hasard entre candidats à la sélection et de l'indépendance entre les effets de leurs locus biallélliques, postule que les fréquences génotypiques peuvent être déduites des fréquences alléliques suivant l'équilibre de Hardy-Weinberg. Dans ce sens, le partitionnement de la valeur génétique reflète la contribution de plusieurs facteurs mendéliens. A l'additivité des effets alléliques postulée par le modèle infinitésimal de Fisher s'ajoute une part non transmissible des ascendants vers les descendants

qui dérive de l'interaction allélique intralocus que représente la dominance et interlocus que représente l'épistasie générant ainsi une déviation par rapport à la ségrégation mendélienne, ce qui souligne que la réalité biologique s'écarte de l'hypothèse de l'hérédité additive des caractères. Par ailleurs, outre le masquage des effets de dominance et d'épistasie lors de la décomposition de la valeur phénotypique, maintes sources de variation phénotypique ne sont pas capturées par les modèles classiques donnant lieu à la notion d'héritabilité manquante ; celle-ci mérite d'être investiguée afin de combler le fossé entre la variation génétique et la variation des caractères cibles.

Dans cette perspective, l'emprunt d'approches ayant démontré leur preuve dans d'autres disciplines afin d'étudier le modèle d'hérédité des caractères d'intérêt pour le design de programmes de sélection efficaces s'avère judicieux afin de répondre aux enjeux de l'amélioration variétale contemporaine et concevoir des variétés performantes (Bernardo 2020). Par conséquent, le développement des approches de modélisation de la variation des caractères quantitatifs et son application dans les stratégies de sélection repose sur l'adossement de l'amélioration génétique sur des modèles sophistiqués qui tiennent compte de la complexité des interactions entre les différentes partitions formant le lien entre le génotype et le phénotype. Les efforts déployés en vue de l'élaboration de la carte génotype-phénotype (genotype-phenotype map) (Figure 1) pour les caractères quantitatifs se basent sur l'investigation de l'architecture génétique. Cette architecture sous-tend la variation phénotypique par le biais de l'étude de la liaison et de l'association génétique entre génotype et phénotype grâce aux réseaux de gènes dont l'interaction est modulée par l'environnement (Houle et al. 2010). L'objectif étant d'élucider le déterminisme génétique responsable de la variabilité phénotypique afin d'appréhender la complexité de la notion d'hérédité des caractères agronomiques.

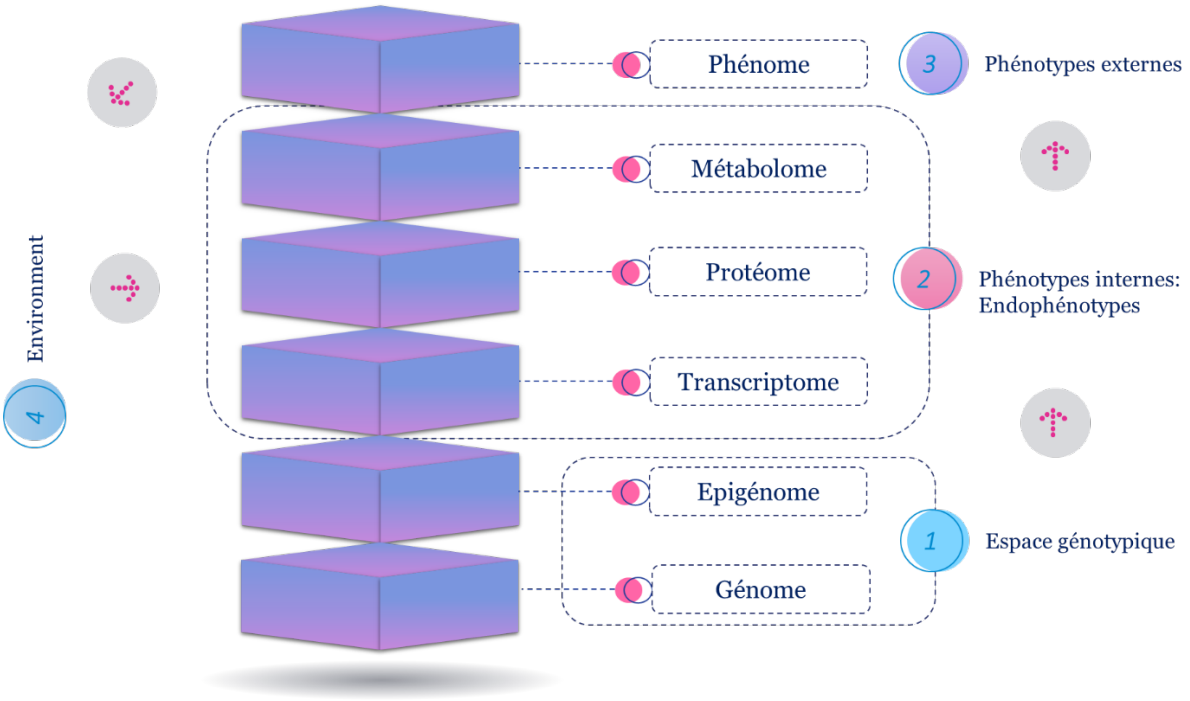


Figure 1: Carte génotype - phénotype

Dans cette optique, grâce à l'essor du séquençage haut-débit du génome qui s'est accompagné d'une baisse marquée et continue des coûts de génotypage, la valorisation de l'information génomique apportée par des marqueurs moléculaires couvrant tous les segments chromosomiques a permis de pallier le problème de l'héritabilité manquante grâce à la sélection génomique dont l'intérêt permet d'appréhender la variation génétique totale y compris celle résultant de locus à effet moyen à faible ainsi que d'allèles rares (Meuwissen et al. 2001). A l'encontre de la sélection assistée par marqueurs, la sélection génomique n'applique pas de sélection sur la base de la significativité statistique portant sur la contribution des marqueurs supposés être en déséquilibre de liaison avec les polymorphismes causaux à la variation du phénotype.

Le recours à la génomique pour disséquer la variation inhérente à l'espace génotype-phénotype doit être complémenté par d'autres disciplines afin de capturer l'héritabilité d'origine non génétique responsable de la variation des caractères cibles. En effet, la variation quantitative des phénotypes dérive d'une variabilité génétique multifactorielle issue de l'interaction de réseaux transcriptionnels, métaboliques et biochimiques qui se caractérisent par leur dynamisme et leur interconnectivité. La dérivation de la variation quantitative à partir des réseaux d'interaction entre locus revêt une importance cruciale dans la compréhension des systèmes biologiques et la prédiction de la composition biochimique des organes de la plante, objet de la biologie prédictive. Par ailleurs, la variation quantitative est privilégiée dans les

schémas de sélection et elle permet l'apport d'une diversité reposant à la fois sur la richesse allélique et le progrès génétique.

Dans un tel contexte, la valorisation des données (phénotypiques) à haut-débit en référence à la spectroscopie répond à la problématique grâce à la capacité des spectres à capturer la variabilité des phénotypes intermédiaires entre le génotype et le caractère observé (endophénotypes, Figure 1). En corollaire, la signature spectrale offre l'opportunité d'appréhender les liens entre le phénotype et le génotype et soutenir le modèle d'hérédité des caractères sur lesquels sont axées la sélection et la création variétale. Par ailleurs, étant fondée sur la spectroscopie infrarouge, la sélection phénomique s'articule autour de l'échantillonnage mendélien et du déséquilibre de liaison entre locus marqueurs et locus impliqués dans la variation des caractères quantitatifs (QTL), à l'instar de la sélection génomique.

Dans ce contexte, cette thèse vise à évaluer la performance des modèles génomiques et phénomiques pour prédire des caractères d'intérêt chez l'abricotier portant sur la qualité des fruits, la phénologie et la sensibilité aux maladies. Nous visons donc à apporter des éléments de réponse quant à l'efficacité de ces deux stratégies de sélection axées sur la prédiction et des éléments de réflexion par rapport à leur intégration dans le schéma de sélection de l'abricotier afin d'aider à la décision et ainsi optimiser le progrès génétique attendu chez cette espèce.

Ce manuscrit de thèse est organisé en six chapitres :

Le **chapitre 1** est une synthèse de la littérature scientifique portant sur des concepts fondamentaux de l'amélioration génétique des plantes et d'approches de sélection fondées sur les avancées technologiques en matière de génotypage et de caractérisation spectroscopique. Il s'agit d'une mise en contexte aidant à situer la création et la sélection variétale chez l'abricotier. Les deux stratégies de sélection génomique et phénomique sont présentées et l'apport de la modélisation statistique à leur mise en œuvre potentielle est exploré.

Dans le **chapitre 2** nous présentons le matériel végétal et les outils statistiques qui ont été valorisés dans le cadre de ce travail. Les données phénotypiques, génotypiques et spectrales acquises dans deux dispositifs expérimentaux contrastés sont présentées et les approches de prédiction sont évaluées.

Dans le **chapitre 3**, nous évaluons la performance des modèles de sélection génomique pour des caractères liés à la qualité des fruits mesurés dans une descendance biparentale. Il est présenté sous la forme d'un article publié dans la revue G3 : Genes | Genomes | Genetics.

Le **chapitre 4** est consacré à l'évaluation de la capacité prédictive des modèles de sélection phénomique, nouvelle stratégie de sélection qui s'intéresse à la spectroscopie dans le proche infrarouge comme outil haut-débit. Des caractères liés à la qualité des fruits ainsi que des caractères phénologiques ont été prédits en utilisant l'information spectrale en comparaison avec l'information génomique. Ce chapitre est présenté sous la forme d'un article soumis à la revue G3 : Genes | Genomes | Genetics.

Le **chapitre 5** est axé sur un panel représentatif de la diversité inhérente à l'espèce, offrant un cadre de validation pour les deux stratégies de sélection. Il se focalise sur des caractères d'intérêt associés à la qualité des fruits, la phénologie et la sensibilité aux maladies.

Dans le **chapitre 6,** nous discutons des principaux résultats acquis ainsi que de la pertinence de la mise en œuvre de la sélection génomique et phénomique dans le programme de sélection de l'abricotier. L'évaluation du cadre de prédiction et d'optimisation est discutée en fonction du matériel végétal mobilisé, des approches de modélisation et des caractères considérés.

Chapitre 1 : Synthèse bibliographique

1.1. De la génétique des populations à la génétique quantitative en vue de l'exploration de l'hérédité des caractères

Dans cette section, nous passons en revue les concepts fondateurs de la génétique des populations et de la génétique quantitative qui ont révolutionné la création variétale depuis les lois de Mendel (1866) jusqu'au modèle de Fisher (1918).

« L'amélioration des plantes peut être définie comme la modification de certains de leurs caractères pour répondre aux besoins de l'Homme. Elle a commencé avec la domestication, et s'est poursuivie essentiellement à partir de la fin du XIX^{e} siècle par l'amélioration dirigée des plantes, intégrant de plus en plus dans ses méthodes et ses outils les progrès des connaissances. Aujourd'hui, l'amélioration des plantes est devenue la science de l'art de la création de variétés ayant des caractères bien définis. » (Gallais 2011).

La sélection et la création variétale opèrent dans un cadre théorique et empirique alimenté par des avancées vis-à-vis de la compréhension des mécanismes biologiques sous-jacents à la variation des caractères d'intérêt agronomique ainsi que la maîtrise de leur transmission au fil des générations. Ces caractères, sur lesquels est axée l'amélioration variétale se partagent en deux catégories.

- Des caractères qualitatifs ou mendéliens présentant une variation discrète ou discontinue résultant de la ségrégation d'un nombre limité de gènes. L'intervention de l'environnement dans l'expression de ces caractères est considérée comme négligeable. Ces caractères présentent une architecture monogénique ou oligogénique.
- Des caractères quantitatifs qui impliquent une variation continue déterminée par une architecture génétique complexe due à la contribution de multiples gènes (polygènes) ou locus à caractère quantitatif (QTL) au contrôle de la variation. Dans le cas de ces caractères, l'environnement est doté d'une forte incidence sur leur expression.

Quant aux caractères mendéliens s'appuyant sur des classes phénotypiques distinctes témoignant de la nature discontinue des caractères, leur déterminisme génétique est conforme aux lois Mendéliennes en termes d'uniformité des hybrides de première génération F1 qui sont issus du croisement de deux lignées pures. Le caractère s'exprimant en F1 correspond au

caractère dominant tandis que le caractère masqué présente le caractère récessif. Il s'agit de la première loi de Mendel. Le croisement des hybrides F1 entre eux aboutit à une deuxième génération F2 comprenant 75% d'individus présentant le caractère dominant et 25% présentant le caractère récessif. C'est la deuxième loi de Mendel ou loi de ségrégation. Quant à la troisième loi, elle postule une indépendance de la transmission des caractères au fil des générations.

En dépit de leur contribution significative à l'établissement des fondements de la génétique du XX^{e} siècle, les lois de Mendel présentent des limites et ont ouvert la voie à l'émergence de concepts réfutant ou complétant la théorie mendélienne de l'hérédité. Prenons l'exemple des expériences de Wilhelm Johannsen (1903) sur des lignées pures de haricot *Phaseolus vulgaris* qui ont mis en évidence une hétérogénéité phénotypique non héritable soulignant la contribution de facteurs autres que le génotype à la variabilité des caractères et introduisant le concept de l'interaction entre le génotype (G) et l'environnement (E) dans lequel il évolue.

A l'appui du schéma mendélien, Fisher a conçu le modèle infinitésimal postulant que l'expression du phénotype dérive de la contribution d'une infinité de locus mendéliens dont chacun explique une proportion infinitésimale de la variation phénotypique, ce qui est le cas de la plupart des caractères ciblés par la sélection des plantes et notamment des arbres fruitiers. Cette théorie représente le concept fondamental de l'approche quantitative de la génétique.

En définitive, en amélioration des plantes, la cible visée est l'identification des individus élites parmi les candidats à la sélection et la création de variétés performantes en termes de volume et de qualité de production et répondant aux enjeux en lien avec la durabilité de la production et de la sécurité alimentaire. C'est dans cette perspective que s'inscrivent les fondements de la génétique quantitative qui visent à perfectionner les différentes méthodes de sélection des caractères quantitatifs, depuis la conception des plans de croisement jusqu'à la mise en place des dispositifs expérimentaux d'évaluation de cette diversité par le biais de la modélisation de la variation issue de la contribution de plusieurs gènes au phénotype d'intérêt.

Ainsi, depuis l'avènement des marqueurs moléculaires, l'amélioration des plantes dispose d'un cadre expérimental propice à l'investigation de l'hérédité des caractères ciblés par la sélection et la création variétale. Ces marqueurs visent à révéler le polymorphisme génétique en lien avec la variation des caractères. Autrement dit, ils permettent de renseigner la variation dans la séquence d'ADN associée à la variation phénotypique. Également, ils offrent l'opportunité d'étudier la structuration de la diversité génétique disponible à l'échelle du génome mais aussi à l'échelle de l'espèce.

Dans ce contexte, outre la modélisation statistique et la génomique, la génétique quantitative s'appuie également sur diverses disciplines telles que la génétique des populations, qui consistent à appréhender la complexité des scénarios évolutifs et leurs conséquences sur l'hérédité des caractères cibles. Ainsi, la population de référence en génétique quantitative est une population en équilibre de Hardy-Weinberg (H-W) où les fréquences alléliques et génotypiques restent constantes au fil des générations. Il s'agit d'une population idéale, panmictique, d'effectif illimité et n'étant soumise à aucune pression évolutive (mutation, sélection ou migration). Or un écart à l'équilibre au sein de la structure allélique de la population idéale est représenté par la notion de déséquilibre de liaison (Figure 2) qui correspond à l'association non aléatoire d'allèles à des locus différents dans une population (Jennings 1917; Lewontin 1964). En corollaire à l'association préférentielle de certains allèles, on constate que certaines combinaisons alléliques sont plus fréquentes que d'autres. Cet écart à l'équilibre de H-W est synonyme à l'évolution des fréquences alléliques et par conséquent des fréquences génotypiques. Il relève de l'existence de forces évolutives.

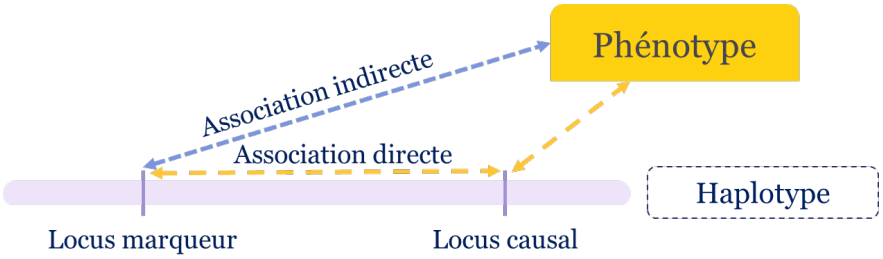


Figure 2: Illustration de la notion du déséquilibre de liaison adaptée de (Balding 2006)

Soient deux locus portant chacun deux allèles A et B en fréquences p_A et p_B, respectivement, avec p_{AB} pour la fréquence de l'haplotype AB.-L'écart à l'association aléatoire est représenté par le paramètre D, le coefficient de déséquilibre de liaison (DL) dont l'équation est :

$$D_{AB} = p_{AB} - p_A p_B \qquad 1.1$$

Mathématiquement, le DL mesure l'écart à l'association aléatoire de deux gènes dans une population donnée. Il s'agit d'un écart à l'indépendance des locus. Ce concept renseigne sur la liaison entre locus marqueurs et locus impliqués dans la variation du caractère quantitatif (QTL).

La modélisation du lien entre marqueur moléculaire et QTL peut être valorisée dans le cadre de l'étude de liaison entre variation génétique causale et variation phénotypique. L'objectif de cette approche est l'exploration de la nature de la variation quantitative des caractères d'intérêt. Elle est fondée sur le principe de recombinaison et de ségrégation entre marqueur et QTL. L'intérêt de cette approche réside dans la facilité de la mise en œuvre des dispositifs expérimentaux en vue d'identifier les composantes génétiques liées à un caractère quantitatif d'intérêt en s'appuyant sur les avancées méthodologiques de la détection de QTL. En revanche, le faible nombre de générations et par conséquent d'évènements de recombinaison présente une limite à la résolution de la cartographie génétique. Pour pallier cette contrainte, le recours à la diversité génétique s'est révélé prometteur dans la quête de déterminants génétiques des caractères cibles notamment via les études d'association qui valorisent l'appariement préférentiel des allèles entre locus marqueurs et QTLs. Cependant, afin d'optimiser la résolution des études d'association génétique, il est intéressant de tenir compte des scénarios évolutifs de l'espèce. L'intégration de la structure de la diversité génétique dans les modèles de régression de la variation phénotypique à la variation génétique quantitative permet d'éviter les signaux d'association artéfactuels causés par les facteurs confondants.

1.2. De la partition de la variance phénotypique à la prédiction de la valeur génétique

1.2.1. Décomposition de la variance phénotypique

La notion de phénotype correspond à l'expression de l'information génomique d'un individu dans l'environnement où il évolue. La décomposition de la variance phénotypique consiste à déterminer la part des facteurs génétiques par rapport aux facteurs environnementaux. Les différences phénotypiques entre individus dérivent de la variation de l'expression des gènes à laquelle s'ajoute l'intervention de l'environnement.

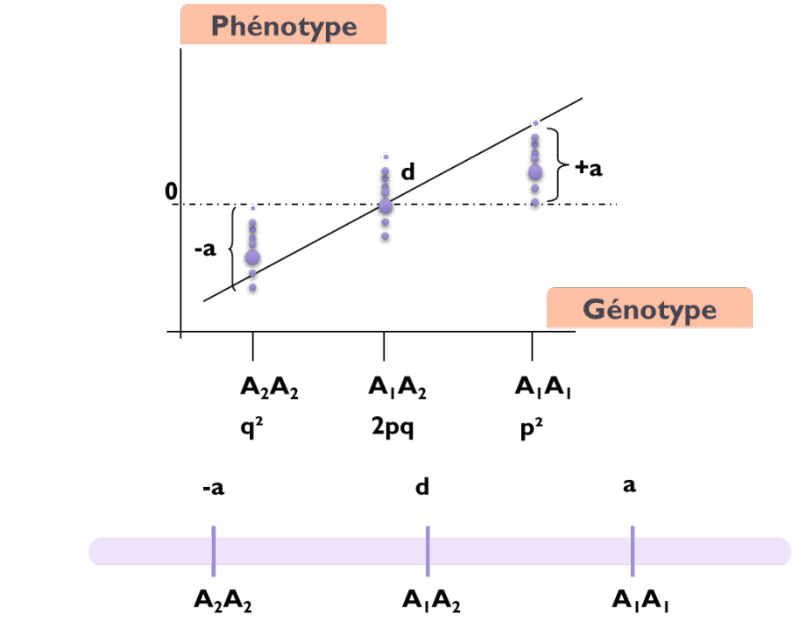
Sous l'hypothèse d'indépendance entre l'effet génétique et l'effet de l'environnement, la valeur phénotypique P s'exprime comme étant la somme d'une valeur génétique G et une valeur environnementale E :

$$P = G + E \qquad 1.2$$

La valeur génétique d'un individu est composée d'une valeur génétique additive *A* correspondant à la somme des effets moyens des allèles et d'une valeur de dominance *D* liée à l'interaction entre allèles d'un même locus, auxquelles s'ajoute une composante liée à l'interaction *I* entre allèles présents à des locus différents. Il s'agit de l'épistasie.

$$G = A + D + I \qquad 1.3$$

La valeur génétique additive d'un individu correspond à la somme des valeurs génétiques additives associées aux génotypes A_1A_1, A_1A_2 et A_2A_2 à chaque locus biallélique. Ces dernières sont déterminées par une régression linéaire entre les valeurs génotypiques et le nombre d'allèles avec α l'effet allélique moyen correspondant à la pente de la droite de régression (Figure 3).



Génotypes	Fréquence allélique	Valeur génétique additive	Dominance
A_1A_1	p^2	$2\alpha_1 = 2q\alpha$	$-2q^2d$
A_1A_2	$2pq$	$\alpha_1 + \alpha_2 = (p-q)\alpha$	$2pqd$
A_2A_2	q^2	$2\alpha_2 = -2p\alpha$	$-2p^2d$

Figure 3: Illustration de la décomposition de la valeur génétique

La valeur génétique additive d'un individu s'exprime donc ainsi :

$$A = \sum_{locus}(\alpha_i + \alpha_j) \qquad 1.4$$

où α_i et α_j les effets moyens associés aux deux allèles *i* et *j*.

Dans une population panmictique, la variance d'additivité correspond à la somme des variances des effets additifs aux différents locus. Quant à la dominance, elle représente la déviation de la moyenne de la population à la somme des effets alléliques additifs. Sous l'hypothèse d'une

interaction nulle entre locus, la variance phénotypique peut être exprimée sous forme de la somme des effets génétiques additifs, des effets liés au masquage des effets des allèles récessifs par les effets des allèles dominants et une variance résiduelle.

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_E^2 \quad 1.5$$

Dans ce contexte, l'héritabilité au sens large représente la proportion de la variabilité génétique par rapport à la variabilité totale observée. Cette variabilité génétique intègre l'additivité et la dominance entre allèles de chaque locus impliqué dans la variation du phénotype en question.

$$h_{SL}^2 = \frac{\sigma_G^2}{\sigma_G^2+\sigma_E^2} \quad 1.6$$

où $cov(G,E) = 0$; $cov(P,G) = \sigma_G^2$

Quant à l'héritabilité au sens strict, elle représente la proportion de la variabilité totale observée due à la variance d'additivité, la part de variance génétique transmise des parents aux descendants. Il s'agit de la fraction héritable de la variabilité phénotypique.

$$h_{SS}^2 = \frac{\sigma_A^2}{\sigma_G^2+\sigma_E^2} \quad 1.7$$

1.2.2. Prédiction de la valeur génétique par des marqueurs moléculaires

La notion de prédiction renvoie à une démarche statistique permettant de déterminer la performance agronomique d'un individu à un stade de développement précoce c'est-à-dire préalablement à l'expression du caractère d'intérêt. Dans ce contexte, la variable réponse est la valeur phénotypique et les variables explicatives correspondent aux marqueurs moléculaires (Lande and Thompson 1990).

L'objectif est donc de mettre en évidence les liens de causalité entre un phénotype et un génotype en valorisant le déséquilibre de liaison entre locus marqueurs et QTL d'intérêt. Par conséquent, la valeur génétique additive correspond aux contributions pondérées des effets des allèles aux marqueurs à la variation du phénotype. Elle est obtenue par régression de la valeur phénotypique sur le dosage allélique correspondant aux marqueurs. Cette valeur génétique renseigne sur la performance agronomique du candidat à la sélection.

1.2.3. Réponse à la sélection

L'objectif ciblé par la sélection variétale consiste à améliorer le gain génétique, qui est défini comme étant le gain de performance d'une population par an, suite à la sélection. Le progrès

génétique en un cycle de sélection ou réponse à la sélection ΔG représente le progrès génétique prévu à l'issue de ce cycle. Il correspond la différence entre le phénotype moyen des descendants des individus sélectionnés et le phénotype moyen de la population candidate à la sélection.

$$\Delta G = \mu_{descendants} - \mu_{candidats} = i \times r_{u,\hat{u}} \times \sigma_a \quad 1.8$$

où :

i : l'intensité de sélection correspond à la différentielle de sélection S 'exprimée en unité d'écart-type phénotypique', c'est-à-dire $i = \frac{S}{\sigma_P}$ où $S = \mu_{sélectionnés} - \mu_{candidats}$ et σ_P est l'écart-type phénotypique.

$r_{u,\hat{u}}$: la précision de la sélection illustrant la fiabilité des valeurs génomiques Genomic Estimated Breeding Values (GEBVs)

σ_a : l'écart-type additif correspondant à la racine carrée de la variance génétique additive

Les facteurs principaux affectant le taux de progrès génétique sont fournis dans l'équation de Falconer (1989), présentée ci-dessous :

$$\Delta G = \frac{ir\sigma_a}{T} \quad 1.9$$

Intensité de sélection

L'intensité de sélection, telle qu'illustrée dans la Figure 4 représente la différentielle de sélection. Deux facteurs principaux affectent l'intensité de la sélection : la taille de la population et la proportion d'individus sélectionnés pour servir de géniteurs à la génération suivante. Concernant le premier facteur, une plus grande intensité i peut être obtenue dans des populations d'effectif important étant donné qu'un nombre plus important de candidats peut être évalué. Également, l'intensité de la sélection dépend du pourcentage d'individus sélectionnés parmi les candidats.

Écart-type génétique

L'écart-type génétique σ_a reflète la variabilité génétique sous-jacente d'un caractère donné au sein de la population.

Durée du cycle de sélection

Il s'agit du temps du cycle de sélection requis pour obtenir une réponse. Une réponse à la sélection provient principalement de la réduction de la durée des cycles de sélection en génotypant les candidats à un stade précoce et en s'affranchissant du phénotypage.

Précision de sélection

La précision de prédiction est affectée par plusieurs facteurs tels que la taille de la population, la structure de la population, le lien entre population d'entraînement et population de validation, la densité des marqueurs, l'architecture génétique des caractères et le modèle statistique.

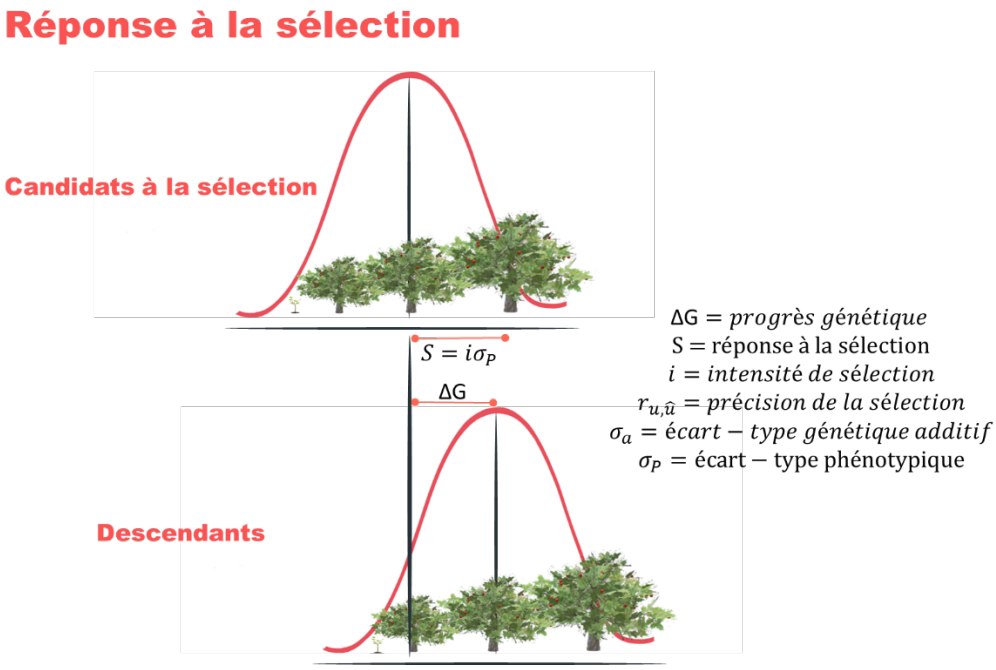


Figure 4: Réponse à la sélection

1.3.Stratégies de sélection fondées sur le haut-débit

1.3.1. Sélection génomique

L'essor des technologies de génotypage a ouvert la voie à la sélection génomique (SG), une approche variante de la sélection assistée par marqueurs mais qui s'intéresse à un grand nombre de marqueurs couvrant tout le génome indépendamment de tout seuil de signification statistique (Meuwissen et al. 2001b). Cette approche se réfère au fondement des décisions de sélection sur

la base des valeurs génomiques estimées (GEBV) qui sont calculées en estimant les effets des marqueurs à partir d'équations de prédiction dérivées d'une population de référence dans le but d'identifier des individus les plus performants (Figure 5).

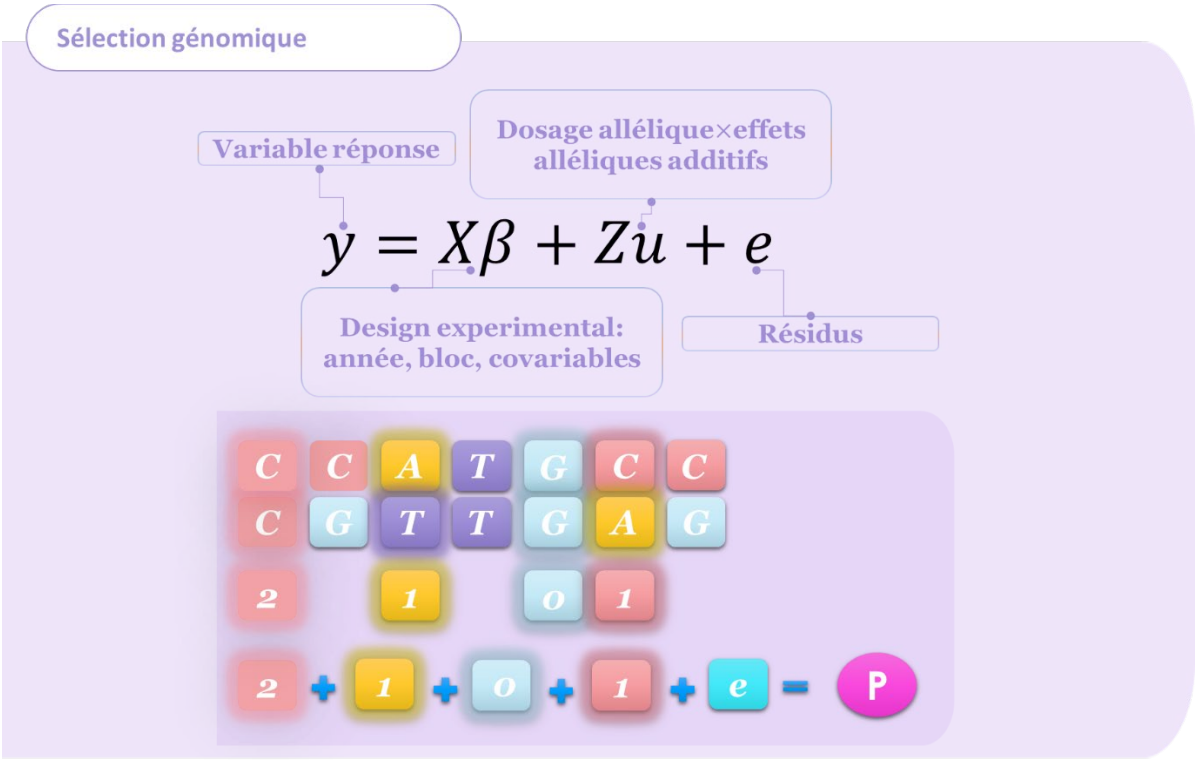


Figure 5: Illustration du modèle de sélection génomique

La mise en place de la SG repose sur la conception d'une population de référence phénotypée pour les caractères d'intérêts agronomiques et génotypée aux locus marqueurs. L'estimation des effets associés aux QTL s'effectue au sein de la population d'entraînement. Par la suite, le calcul des valeurs génétiques GEBV intègre les effets des QTL estimés dans la population d'entrainement et le dosage allélique des candidats à la sélection. La validation de l'approche de sélection génomique repose sur l'évaluation du modèle de prédiction par le biais de la validation croisée (Figure 6).

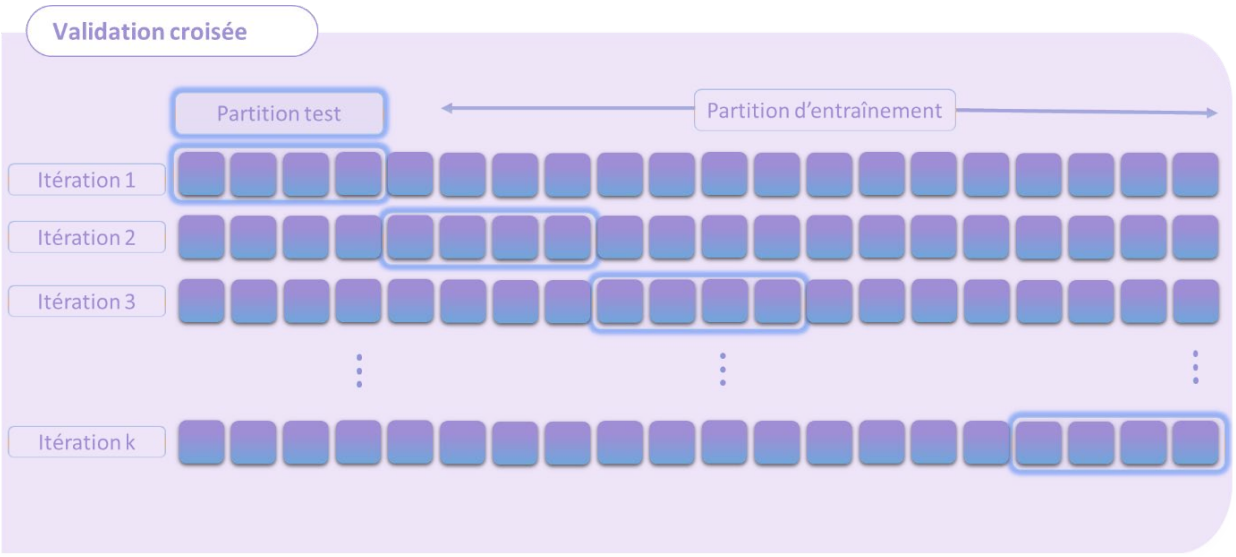


Figure 6 : Illustration du principe de la validation croisée

La sélection génomique se base sur la notion de DL permettant de révéler les locus régissant la variation des caractères cibles étant donné que les mutations causales de ces caractères ne sont pas connues. Lorsque le déséquilibre de liaison est très fort entre les marqueurs et les QTLs, la prédiction des performances peut être qualifiée de robuste (Liu et al. 2015a).

A l'instar de la SAM, la SG utilise comme critère de sélection le génotype des candidats aux marqueurs en DL avec les gènes ou les QTLs d'intérêt. En effet, cette stratégie consiste à sélectionner les individus porteurs des combinaisons alléliques les plus favorables. Par conséquent, les effets alléliques représentent le critère de sélection et les allèles sont donc l'unité d'évaluation des performances agronomiques des candidats à la sélection (Lorenz et al. 2011).

En sélection génomique, les phénotypes ne sont plus utilisés pour sélectionner mais pour concevoir les modèles de prédiction (Heffner et al. 2009; Lorenz et al. 2011). En corollaire, cette approche de sélection pourrait être prometteuse en termes de progrès génétique et notamment via le raccourcissement du cycle de sélection en s'affranchissant du phénotypage. Surmonter le besoin de phénotyper ouvre également la possibilité d'augmenter considérablement le nombre d'individus faisant partie du pipeline de sélection et par conséquent augmenter l'intensité de sélection.

Le recours à la SG est de plus en plus répandu chez les espèces pérennes : l'Eucalyptus (Resende et al. 2012a; Tan et al. 2017; Kainer et al. 2018), l'épinette blanche *Picea glauca* (Beaulieu et al. 2014; Ratcliffe et al. 2017), le pin Loblolly *Pinus taeda (*Resende et al. 2012c*;*

Resende et al. 2012e*)* et l'épinette de Norvège *Picea abies* (Chen et al. 2018). Pour les espèces fruitières, la SG a été implémentée chez le pommier (Kumar et al. 2012a; Muranty et al., 2015; Roth et al. 2020), le poirier (Kumar et al. 2019), la vigne (Fodor et al. 2014b) et les citrus (Minamikawa et al. 2017b).

1.3.1.1. Apports de la sélection génomique

L'intérêt de la sélection génomique réside dans la capacité à contrôler les trois paramètres de l'équation du sélectionneur (Figure 7). En augmentant la précision et l'intensité de la sélection et en raccourcissant l'intervalle de génération, le taux de progrès génétique pour les caractères cibles est susceptible d'augmenter considérablement.

Avant l'avènement des marqueurs moléculaires, les efforts de sélection chez les animaux d'élevage s'appuyaient sur la matrice d'apparentement estimée à partir d'informations généalogiques recueillies sur plusieurs générations. Cependant, ces informations peuvent être indisponibles ou incomplètes. L'inférence des liens de parenté à partir d'un nombre limité de marqueurs moléculaires se traduit par un biais dans l'estimation des paramètres génétiques. Toutefois, l'augmentation de la densité de marquage a permis de réaliser des estimations du niveau d'apparentement, non sujettes au biais apporté par le pedigree (Hayes and Goddard 2008).

Prenons l'exemple des bovins laitiers, il a été démontré que les bovins génomiques d'un an qui disposent d'informations uniquement sur la valeur génomique GEBV présentent un mérite génétique plus élevé que celui des bovins testés sur descendance. Leur utilisation comme reproducteurs s'est traduite par un raccourcissement de l'intervalle de génération de 54 mois à 21 mois (Schefers and Weigel 2012). Tel est le cas ainsi de plusieurs espèces végétales d'intérêt économiques chez lesquelles la SG a fait ses preuves en contribuant à l'amélioration de la précision et au raccourcissement des cycles de sélection et par conséquent à l'optimisation du progrès génétique.

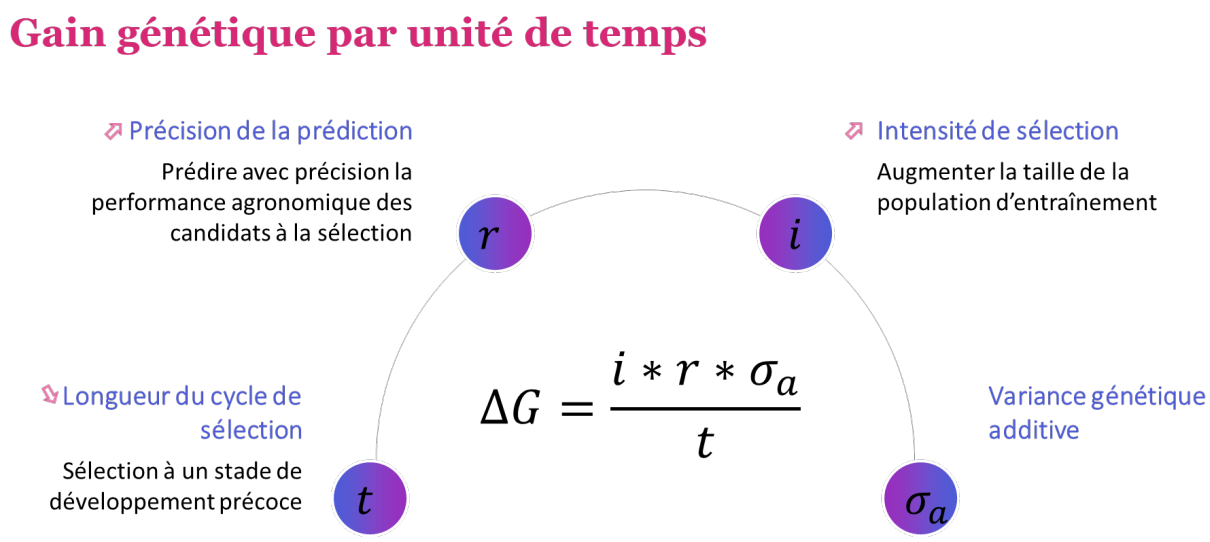


Figure 7: Attentes vis-à-vis de la sélection génomique

1.3.1.2. Modélisation statistique de la SG

La modélisation de la sélection génomique consiste à exploiter le lien entre variation phénotypique et variation génétique dans le but de prédire les valeurs génétiques des candidats à la sélection en valorisant le déséquilibre de liaison entre les locus marqueurs et les polymorphismes causaux des phénotypes d'intérêt. Par conséquent, les locus marqueurs représentent les prédicteurs et le phénotype cible représente la variable réponse à prédire. La prédiction du phénotype par le biais des marqueurs moléculaires consiste à résoudre l'équation du modèle de régression linéaire suivante :

$$y = X\beta + e \qquad 1.10$$

où :

- y est la variable réponse qui représente le phénotype à prédire
- X est une matrice de dimensions $n \times m$
- β est le vecteur des paramètres à estimer
- $e \sim N(0, \sigma_e^2)$ est l'erreur résiduelle

L'estimation du paramètre β s'effectue par la méthode des moindres carrés qui consiste à minimiser la somme des carrés résiduels, c'est-à-dire la distance entre les observations y et les prédictions $\hat{y}$ (Hastie, 2009. La solution de l'équation du modèle de sélection génomique par la méthode des moindres carrés est fournie par l'écriture matricielle suivante :

$$\hat{\beta} = (X^t X)^{-1} X^t y \; e \qquad 1.11$$

La singularité (impossibilité d'inversion) de la matrice X^tX entrave la résolution de l'équation du modèle. En corollaire, un problème de dimensionnalité se manifeste dérivant d'un nombre de paramètres à estimer (effets des marqueurs) qui excède le nombre d'observations (phénotypes) (p >> n). Il en découle un sur-ajustement des effets des marqueurs diminuant la capacité prédictive du modèle de SG. Par conséquent, l'estimation des effets des prédicteurs par le biais de la méthode des moindres carrés est entravée par un déficit de degrés de liberté dont dispose le modèle de régression pour évaluer le lien entre le phénotype et le génotype au locus marqueur. Même si le modèle de prédiction dispose de suffisamment de degrés de liberté pour estimer simultanément les différents paramètres, le biais statistique apporté par un niveau élevé de multicolinéarité entre marqueurs s'avère préjudiciable pour la calibration du modèle (Lorenz et al. 2011).

Pour faire face à la malédiction de la dimensionnalité et pallier le problème de multicolinéarité, maintes approches statistiques ont été conçues pour la sélection génomique. Elles sont regroupées en modèles de régression pénalisée, modèles de sélection de variables et modèles de réduction des dimensions. Cette classification est fondée sur la fonction de pénalité $K(\beta)$ appliquée aux coefficients de régression, qui est contrôlée par le paramètre de pénalisation (shrinkage) λ. Ce paramètre $\lambda \geq 0$ détermine l'étendue du rétrécissement des coefficients de régression. Plus la valeur de λ est élevée, plus la quantité de rétrécissement est importante (Hastie et al. 2009).

1.3.1.2.1. Modèle linéaire mixte

Le modèle linéaire mixte modélise la variable réponse comme étant une combinaison linéaire de variables explicatives aléatoires et fixes. Dans ce contexte, le modèle statistique de référence, en sélection génomique, est un modèle linéaire mixte qui établit le lien entre une variable réponse correspondant au phénotype à prédire et des variables prédictives classées en facteurs fixes représentés par les facteurs environnementaux et des facteurs aléatoires représentés par les marqueurs moléculaires.

$$y = X\beta + Zu + e \qquad 1.12$$

où y est le vecteur du phénotype à prédire, X est la matrice d'incidence des facteurs fixes, β est le vecteur des effets fixes, Z est la matrice renfermant les doses alléliques aux locus marqueurs, u est le vecteur des effets aléatoires attribués aux marqueurs.

Comme dans le cadre de la méthode des moindres carrés, le choix du paramètre de pénalité $\lambda > 0$ repose sur la minimisation de la somme des carrés résiduels. Or, outre la minimisation de l'erreur du modèle dans le cadre d'une stratégie de validation croisée, une méthode alternative consiste à supposer que les effets des marqueurs sont tirés au hasard suivant une loi de distribution normale centrée réduite, en accord avec le modèle infinitésimal de Fisher (1912) et qui renvoie à la résolution des équations du modèle mixte de Henderson (1975), s'écrivant comme suit :

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \quad 1.13$$

Dans ce contexte, le terme λ est défini comme étant le ratio de la variance résiduelle divisée par la variance des effets des marqueurs $var(\beta)$, sous l'hypothèse d'une variance équivalente pour tous les marqueurs (Piepho et al. 2008).

Pour résoudre ce système d'équations, Meuwissen *et al*. (2001) ont proposé une méthode de régression pénalisée fondée sur l'estimation du meilleur prédicteur linéaire non biaisé (Random Regression Best Linear Unbiased Prediction RR-BLUP). La méthode RR-BLUP postule une distribution des effets aléatoires attribués aux marqueurs selon une loi normale avec une variance homogène pour tous les marqueurs. Cette méthode est équivalente au modèle G-BLUP (Genomic BLUP) qui valorise les marqueurs moléculaires uniquement dans le calcul de la matrice de variance covariance des effets aléatoires (VanRaden 2008).

1.3.1.2.2. Méthodes de régression pénalisée

La régression pénalisée impose une contrainte dans le critère des moindres carrés permettant de réduire le nombre de prédicteurs et donc de minimiser l'erreur de prédiction. Lorsque le paramètre de pénalité se rapproche de zéro, la solution converge vers celle obtenue par la méthode des moindres carrés ordinaires, tandis que de fortes valeurs de λ se traduisent par des coefficients de régression qui tendent vers 0.

1.3.1.2.2.1. Régression Ridge

La régression Ridge est une méthode de régression pénalisée introduite par Hoerl and Kennard (1970). Elle est basée sur l'introduction d'une pénalité λ à la diagonale de la matrice X^tX permettant de la rendre inversible, ce qui permet de limiter la variance de l'estimateur $\hat{\beta}$ et par conséquent de réduire la colinéarité entre les prédicteurs. Les coefficients de la régression Ridge sont sujets à une contrainte proportionnelle à leur taille, illustrée par l'équation suivante :

$$K(\beta)_{ridge} = \sum_{j=1}^{p} \beta_j^2 \quad 1.14$$

Cette fonction de pénalité $K(\beta)$ minimise la somme des carrés des résidus.

$$min\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2\} \quad 1.15$$

L'estimation des coefficients $\hat{\beta}$ s'effectue suivant l'équation :

$$\hat{\beta}_{Ridge,\lambda} = (X^t X + \lambda I)^{-1} X^t y \quad 1.16$$

où X est la matrice d'incidence reliant les marqueurs aux individus, λ le paramètre de pénalisation, I la matrice identité et y la variable réponse.

La régression Ridge, encore appelée modèle RR-BLUP, se rapproche du modèle infinitésimal de Fischer avec un paramètre de rétrécissement égal pour tous les marqueurs (Lorenz, 2011). Cette pénalisation est donc homogène pour tous les marqueurs.

1.3.1.2.2.2. Least absolute shrinkage and selection operator LASSO

La méthode LASSO, proposée par Tibshirani (1996), est une méthode statistique qui combine le rétrécissement et la sélection de variables (de los Campos et al. 2009). La réduction de la dimension via le LASSO s'effectue par le biais de l'annulation de certains coefficients et l'ajout d'une pénalité égale à la somme des valeurs absolues des coefficients de régression, comme illustré dans l'équation :

$$K(\beta)_{LASSO} = \sum_{j=1}^{p} |\beta_j| \quad 1.17$$

A l'instar du modèle RR-BLUP, le LASSO fait appel au rétrécissement des coefficients de régression correspondant aux effets des marqueurs. Cependant, la pénalisation est plus forte dans le cadre du LASSO, ce qui se traduit par un plus grand nombre de marqueurs régressés vers 0.

1.3.1.2.3. Méthodes bayésiennes

Pour contourner le problème du nombre de variables prédictives dépassant le nombre d'observations, les méthodes bayésiennes supposent la nullité de certains coefficients du paramètre $\widehat{\beta}$. Elles appliquent donc une sélection de variables.

L'approche Bayésienne a été conçue dans le but de pallier le biais statistique dérivant de la pénalisation homogène des effets génétiques (Meuwissen et al. 2001). Autrement dit, l'inférence bayésienne reposant sur des simulations Markov Chain Monte Carlo (MCMC), postule une hétérogénéité des variances des effets des marqueurs. Dans le cadre bayésien, la majorité des marqueurs en DL avec des QTL a un effet négligeable sur le caractère et une minorité présente un effet fort.

Il s'agit d'une approche robuste en termes d'approximation de la réalité biologique en comparaison avec le modèle infinitésimal qui suppose une contribution d'une infinité de locus à la variance génétique, tous présentant des effets faibles.

Par contraste avec la régression Ridge qui s'écarte de la réalité objective, les modèles bayésiens mettent l'accent sur les marqueurs liés aux mutations causales responsables de l'expression des caractères. Ces marqueurs dotés d'effet fort et dont les variances sont hétérogènes, sont estimés à partir des données expérimentales avec un échantillonnage de Gibbs.

1.3.1.2.3.1. Bayes A

Bayes A postule que les effets des marqueurs suivent une distribution normale avec une variance spécifique à chaque marqueur égale à σ_a^2. Les paramètres de la variance génétique suivent une distribution a priori conforme à une loi χ^2 inverse. Cela se traduit par un nombre important de QTL présentant un effet faible et uniquement quelques QTL avec un effet fort.

$$\sigma_u^2 \sim \chi^{-2}(\vartheta, S) \qquad 1.18$$

1.3.1.2.3.2. Bayes B

Contrairement au modèle Bayes A, Bayes B ne prend en compte que les marqueurs qui présentent un effet significatif sur les caractères. Par conséquent, le modèle Bayes B admet une proportion π de marqueurs qui ne ségrégent pas et présentent donc un effet nul. Par conséquent, une proportion (1-π) de marqueurs est en déséquilibre de liaison avec les polymorphismes causaux. Le modèle Bayes B est équivalent au modèle Bayes A lorsque la proportion de marqueurs à effet nul π est égale à 0.

$$\sigma_u^2 = 0 \qquad avec\ une\ probabilité\ \pi \qquad 1.19$$

$$\sigma_u^2 \sim \chi^{-2}(\vartheta, S) \qquad avec\ une\ probabilité\ 1 - \pi \qquad 1.20$$

1.3.1.2.3.3. Bayes Cπ

Le modèle Bayes Cπ postule que la probabilité π qu'un marqueur présente un effet nul est inconnue (Habier et al. 2011). La conception de ce modèle a été motivée par les inconvénients des deux modèles Bayes A et Bayes B.

1.3.1.2.3.4. LASSO Bayésien

Le LASSO bayésien LB, proposé par (Park and Casella 2008), présente la version bayésienne du modèle LASSO couplant régression pénalisée et sélection de variables. Il s'agit d'une version plus robuste et précise qui assouplit la contrainte associée par rapport au modèle de base. A la différence du modèle LASSO, le LB est basé sur l'échantillonnage de Gibbs, une variante des algorithmes de Monte Carlo par chaînes de Markov.

La figure 8 récapitule les principaux modèles déployés dans le cadre de la sélection génomique, regroupés en deux catégories : des modèles de régression paramétrique et des modèles de régression non paramétrique.

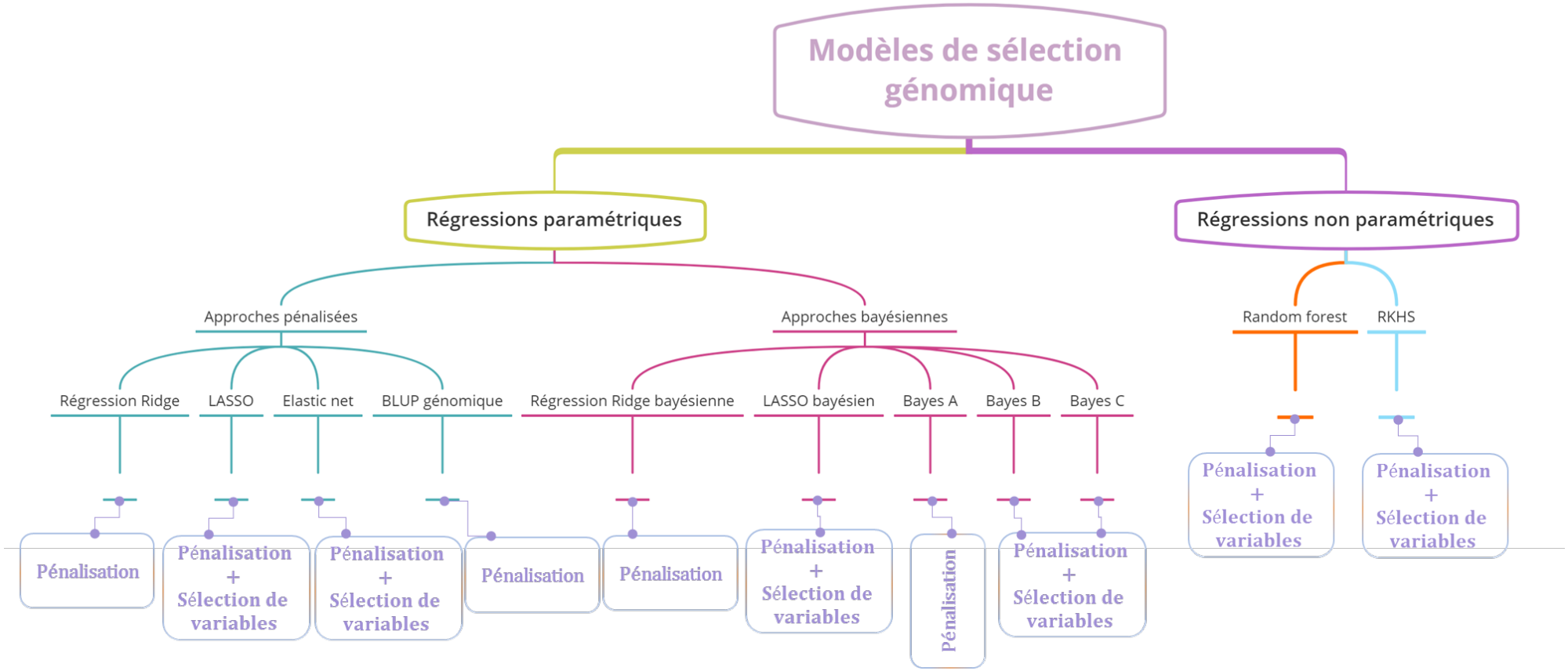


Figure 8: Classification des modèles de sélection génomique

1.3.1.3. Comparaison des modèles de sélection génomique

Le choix du modèle statistique pour la prédiction génomique présente un intérêt non négligeable. La comparaison des différents modèles de SG se révèle justifié compte tenu de l'effet de la méthode de prédiction sur la performance de cette stratégie de sélection. Un modèle statistique est jugé performant lorsqu'il est représentatif de la réalité biologique et par conséquent de la variabilité observée dans les données expérimentales. Parmi les modèles bayésiens, le modèle Bayes B se révèle plus robuste étant donné qu'il traduit fidèlement l'architecture génétique des caractères notamment lorsque la variance génétique est attribuée à un nombre bien déterminé de locus tandis que d'autres ne participent pas au contrôle génétique de ces caractères (Meuwissen et al. 2001 ; Lorenz et al 2011). Certes de nombreux marqueurs peuvent être en déséquilibre de liaison avec les QTL d'intérêt mais certains sont susceptibles d'être situés dans des régions chromosomiques ne participant pas au déterminisme génétique des caractères cibles. Par conséquent, un rétrécissement différentiel des effets des marqueurs s'avère crucial.

Le modèle le plus performant n'existe pas dans la mesure où chaque caractère d'intérêt présente une architecture génétique différente. Lorsque le caractère est déterminé par plusieurs QTL à effets faibles, le modèle RR-BLUP se révèle être supérieur, alors que pour les caractères contrôlés par des QTL majeurs, c'est le modèle Bayes B qui est privilégié (Anderson et al. 2001; Munkvold et al. 2009; Lorenz et al. 2011). En revanche, ces modèles, postulant l'additivité des effets génétiques, peuvent ne pas être performants en termes de précision de prédiction dans le cadre d'une forte contribution des effets génétiques non additifs à l'architecture génétique.

1.3.1.4. Précision de la sélection génomique

Comme indiqué précédemment, la modélisation de la sélection génomique par la régression Ridge est équivalente à l'utilisation du modèle G-BLUP basé sur la matrice d'apparentement génomique, calculée à partir des marqueurs moléculaires (Habier et al. 2007; Hayes et al. 2009). Par conséquent, la précision de la sélection génomique peut être associée à deux composantes principales : une composante est due au déséquilibre de liaison entre locus marqueurs et QTL et une composante est due à la modélisation des liens de parenté.

La précision de sélection génomique associée au déséquilibre de liaison est caractérisée par sa stabilité étant donné qu'elle se dégrade lentement au fil des générations. Ceci est attribué à la rareté des évènements de recombinaison entre individus de la population d'entraînement et les

candidats à la sélection. La perte en précision de prédiction, en lien avec cette composante, est proportionnelle à la fréquence de recombinaison entre marqueurs et QTL (Habier et al. 2007). Quant à la composante associée à la parenté génomique, elle est fondée sur l'inférence de la covariance génétique entre individus. Ce concept de parenté, estimé à partir des marqueurs moléculaires, est dû au partage d'allèles identiques par descendance (Identity By Descent IBD), dérivant par copie d'un même allèle ancêtre. La précision de sélection génomique générée par la capture des liens de parenté, se dégrade rapidement au fil des générations à mesure que la distance génétique entre la population d'entraînement et la population des candidats à la sélection décroît (Lorenz et al. 2011).

Comme souligné par Habier et al. (2007), c'est le DL qui demeure informatif plusieurs générations après l'estimation des effets des marqueurs. Cela laisse place à l'extrapolation du modèle de sélection génomique permettant ainsi de calculer la précision de prédiction de phénotypes exprimés plusieurs générations après la conception du modèle.

En termes de modélisation statistique, les approches de sélection génomique, modélisent différemment les composantes de la précision de prédiction. En effet, il a été démontré que Bayes B accorde plus d'importance aux marqueurs en DL avec les QTL. Néanmoins, le modèle RR-BLUP attribue un poids homogène à tous les marqueurs. En corollaire, la précision de prédiction de Bayes B est plus stable et persiste au fil des générations, en comparaison avec celle de RR-BLUP (Zhong et al. 2009b; Lorenz et al. 2011). Dans ce contexte, Habier et al. (2007) et Habier et al. (2011) privilégient Bayes B pour modéliser la part de la précision de sélection due au déséquilibre de liaison entre marqueurs et QTL. En revanche, RR-BLUP est recommandé lorsque les individus formant la population d'entraînement et les candidats à sélection sont apparentés. La raison pour laquelle RR-BLUP s'avère plus performant que Bayes B dans la modélisation de la parenté génétique additive réside dans la capacité de RR-BLUP à modéliser un nombre de marqueurs supérieur à celui de Bayes B, vu la contrainte liée à la distribution a priori.

Etant donné que la précision de la sélection génomique due au déséquilibre de liaison entre marqueurs et QTL est plus stable au cours des générations, il semble judicieux d'avoir recours à un modèle bayésien tel que le modèle Bayes B notamment lorsque le temps d'acquisition des données phénotypiques est long ou lorsque l'expression du phénotype d'intérêt s'effectue à un stade de développement tardif.

1.3.1.5. Sélection génomique : de l'univarié au multivarié

1.3.1.5.1. Concept de la sélection génomique multivariée

La sélection variétale vise à concevoir des variétés qui cumulent plusieurs caractères d'intérêt. En corollaire, le focus se porte sur maints caractères simultanément afin de répondre à des enjeux de plus en plus diversifiés incluant notamment le contexte d'adaptation au changement climatique et de sécurité alimentaire. Dans cette optique, la construction d'idéotypes variétaux, qui correspondent à des variétés présentant une combinaison optimale de caractères leur procurant la capacité d'être performante dans un environnement bien déterminé, s'avère judicieuse (Donald 1968). En conséquence, orienter la sélection en fonction de variétés qui combinent plusieurs caractères concomitamment implique le recours aux outils de la modélisation multivariée afin de combiner plusieurs variables réponses à expliciter par des facteurs génétiques et environnementaux.

Outre la sélection multicaractère axée sur plusieurs traits privilégiés par la sélection et permettant d'accumuler dans un même génotype plusieurs caractères favorables, la mise en œuvre de la prédiction multivariée permet de s'affranchir au phénotypage des traits coûteux ou difficiles à mesurer soit partiellement ou intégralement. En d'autres termes, la SG multivariée vise à sélectionner des individus non phénotypés ou partiellement phénotypés pour les caractères cibles en se basant sur l'information phénotypique de caractères secondaires. La réduction de l'effectif des candidats à phénotyper pour ces caractères peut se traduire par une optimisation du gain génétique dans la mesure où ils sont intensivement phénotypés pour le caractère proxy. Ainsi, l'inclusion d'informations sur de multiples caractères faciles à mesurer recueillies tout au long des programmes de sélection peut contribuer à l'accélération du processus de sélection (Thompson and Meyer 1986; Bhatta et al. 2020; Lado et al. 2018).

L'engouement pour le haut-débit dans le contexte du phénotypage en vue de l'évaluation des candidats à la sélection peut être valorisé dans la sélection indirecte. Dans cette perspective, maints auteurs ont souligné la pertinence de la mobilisation de données haut-débit telles que la spectroscopie proche infrarouge et la résonance magnétique nucléaire (RMN) afin d'élargir la population d'entraînement. Les informations dérivées des transcriptomes et métabolomes s'avèrent également pertinentes pour prédire des caractères qui leur sont corrélées (Riedelsheimer et al. 2012). En effet, (Guo et al. 2016) suggèrent de recourir à l'utilisation de l'expression des gènes et les informations métaboliques, conjointement avec les marqueurs moléculaires, en tant que prédicteurs en vue d'améliorer le gain génétique pour les caractères complexes sur lesquels sont orientés les programmes de sélection.

De surcroît, le potentiel de la prédiction multivariée dans le cadre de la sélection multicaractère ne se limite pas uniquement au progrès génétique issu de la mobilisation de caractères dont le phénotypage est coûteux et difficile à mettre en œuvre mais également de la valorisation de caractères qui s'expriment à un stade de développement précoce pour sélectionner indirectement les candidats dont les caractères focaux s'expriment tardivement permettant ainsi le raccourcissement du cycle de sélection (Hayes et al. 2017; Fernandes et al. 2018). Ainsi, la sélection indirecte basée sur des caractères secondaires se révèle plus performante que la sélection directe des caractères d'intérêt. En revanche, emprunter des informations issues de deux caractères corrélés permet d'optimiser marginalement la prédiction de la réponse par rapport à un modèle bivarié intégrant un seul caractère corrélé au trait focal (Lado et al. 2018).

Par ailleurs, des études fondées sur des simulations ont souligné l'efficience de l'approche multivariée dans le cadre de la prédiction d'un caractère cible faiblement héritable par le biais d'un caractère proxy fortement héritable, qui lui est génétiquement corrélé (Jia and Jannink 2012b; Guo et al. 2016). Le degré de corrélation génétique entre caractère cible et proxy illustre l'informativité de ce dernier dans le contexte de prédiction du caractère principal. Toutefois, les modèles multivariés intégrant des caractères non corrélés ont tendance à avoir la même performance prédictive que les modèles univariés (Bhatta et al. 2020). En outre, la capacité prédictive des modèles multivariés peut être optimisée en augmentant l'effectif des individus phénotypés pour les caractères proxies. En d'autres termes, l'augmentation du débit de phénotypage pour les caractères proxies en comparaison avec celui des caractères focaux contribue à l'augmentation de la précision de prédiction. Par ailleurs, la mobilisation de caractères corrélés dans le cadre de populations d'entraînement de faible effectif optimise la performance des modèles prédictifs multivariés par rapport aux modèles univariés. Dans ce contexte, la corrélation entre caractères compense la réduction de la taille de la population sur laquelle est entraîné le modèle. A titre d'exemple, Lado *et al.* (2018) ont constaté que la population d'entraînement peut être réduite à 30% de la population totale sans affecter significativement la précision de la prédiction dans la mesure où des caractères corrélés sont intégrés dans le modèle prédictif. En définitive, l'alternative proposée consiste à mettre en œuvre le phénotypage de caractères faciles à mesurer corrélés aux caractères cibles coûteux ou difficiles à mesurer.

1.3.1.5.2. Modélisation statistique multivariée

Dans le cadre de la prédiction univariée, une variable réponse est expliquée par les paramètres du modèle. L'approche multivariée, quant à elle, modélise plusieurs variables réponses et

intègre une matrice de variance-covariance entre caractères. Autrement dit, les modèles multivriées s'intéressent à la prédiction simultanée de plusieurs caractères sur la base du même ensemble de variables d'entrée explicatives (Montesinos-López et al. 2018). Dans ce contexte, le modèle de prédiction génomique multivarié est représenté par l'équation suivante :

$$y = X_t\beta_t + Z_t u_t + e \qquad 1.21$$

où y est le vecteur des phénotypes

X_t est la matrice d'incidence des effets fixes

β_t est le vecteur des effets fixes

Z_t est la matrice d'incidence des effets aléatoires

u_t est le vecteur des effets aléatoires

e est le vecteur des erreurs aléatoires

avec

$$\begin{aligned} y_1 &= X_1\beta_1 + Z_1u_1 + e_1 \text{ pour le caractère 1} \\ y_2 &= X_2\beta_2 + Z_2u_2 + e_2 \text{ pour le caractère 2} \\ &\vdots \\ y_t &= X_t\beta_t + Z_tu_t + e_t \text{ pour le caractère t} \end{aligned} \qquad 1.22$$

Pour le caractère i $(i = 1 \dots t)$, les effets aléatoires u_i et e_i sont supposés suivre une distribution normale avec $u \sim N(0, A\sigma_u^2)$ et $e \sim N(0, I\sigma_e^2)$.

L'écriture matricielle de la structure de variance – covariance est illustrée dans (Maier et al. 2015; Covarrubias-Pazaran et al. 2018a) :

$$V = \begin{bmatrix} Z_1K\sigma^2_{u_{1_{t1}}}Z_1' + Z_1R\sigma^2_{e_{t1}}Z_1' & \cdots & Z_1K\sigma_{u_{1_{1,t}}}Z_t' + Z_{t1}R\sigma_{e_{1,t}}Z_{ti}' \\ \vdots & \ddots & \vdots \\ Z_1K\sigma_{u_{1_{1,t}}}Z_t' + Z_1R\sigma_{e_{1,t}}Z_t' & \cdots & Z_tK\sigma^2_{u_{1_t}}Z_{ti}' + Z_tR\sigma^2_{e_t}Z_t' \end{bmatrix} \qquad 1.23$$

où K désigne la matrice de covariance pour le $K^{ème}$ effet aléatoire et $R = I$ pour le terme résiduel. Les paramètres $\sigma^2_{u\,k_i}$ et $\sigma^2_{e_i}$ désignent les variances génétique et résiduelle pour le caractère i.

Avec $y \sim$ MVN $(X\beta, V)$, $u_i \sim$ MVN $(0, \sum_u \otimes A)$ and $e_i \sim$ MVN $(0, \sum_e \otimes R)$.

Quant à l'approche bivariée qui implique deux variables réponses, elle est modélisée ainsi :

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad 1.24$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim MVN(0, I \otimes R)$$

Les modèles multivariés exploitent non seulement la corrélation entre les individus, mais également la corrélation entre les caractères, ce qui améliore leur efficacité. Les modèles univariés, en revanche, éliminent toute possibilité d'entraînement à partir des liens existants entre caractères étant donné qu'ils sont conçus indépendamment et traitent chaque caractère séparément.

1.3.1.5.3. Corrélation entre caractères

La connaissance des liens qui existent entre les variables phénotypiques et génétiques revêt une importance primordiale pour la conception du modèle de l'hérédité qui régit le mode de transmission des caractères à la descendance. Ce modèle postule que la variation phénotypique se décompose en une part génétique et une part environnementale modélisées par des variables aléatoires d'espérances nulles et de variances égales à $V(G)$ et $V(E)$, respectivement. En corollaire, la variance phénotypique est égale à la variance de la somme de la valeur génotypique et de la valeur environnementale.

$$V(P) = V(G + E) = V(G) + V(E) + 2\,Cov(G, E) \quad 1.25$$

Sachant que la variance est l'écart entre l'espérance du carré du phénotype et l'espérance du phénotype au carré, la variance phénotypique peut être écrite comme suit :

$$V(P) = E(P^2) - E(P)^2 \quad 1.26$$

$$V(P) = \frac{\sum (P - \bar{P})^2}{n} \quad 1.27$$

Etant donné que la variabilité phénotypique représente le résultat du polymorphisme génétique auquel s'ajoutent l'interférence des facteurs environnementaux, la corrélation statistique entre caractères découle d'une corrélation génétique couplée à une corrélation environnementale. La corrélation génétique dérive du partage du patrimoine génétique entre individus dotés d'un fort degré d'apparentement. La corrélation environnementale souligne le partage d'un même environnement.

Soient deux phénotypes P_I et P_J, la corrélation statistique entre les deux phénotypes peut être écrite ainsi :

$$r_p = \frac{Cov(PI,PJ)}{\sqrt{\sigma^2_{PI}\sigma^2_{PJ}}} \qquad 1.28$$

La covariance phénotypique est la moyenne des produits des écarts de chaque caractère à sa moyenne. Sous l'hypothèse d'indépendance entre facteurs génétiques et environnementaux, la covariance phénotypique entre P_I et P_J s'exprime comme étant (Fisher, 1918):

$$Cov(P_I, P_J) = Cov(G_I, G_J) + Cov(E_I, E_J) \qquad 1.29$$

Dans le cadre de la prédiction multicaractère, la covariance entre caractères s'écrit :

$$\sum PP = \sum GG + \sum EE \qquad 1.30$$

Dans cette équation, $\sum PP$, $\sum GG$ et $\sum EE$ désignent les matrices de variances – covariances phénotypique, génotypique et environnementale, respectivement.

Sachant que la covariance représentant la mesure de la variation conjointe entre variables aléatoires s'exprime ainsi :

$$Cov(X1, X2) = E(X1X2) - E(X1)E(X2) \qquad 1.31$$

La corrélation génétique peut être calculée comme suit :

$$r_{G1,G2} = \frac{\sigma_{1,2}}{\sqrt{\sigma^2_{G1}\sigma^2_{G2}}} \qquad 1.32$$

Quant à la corrélation environnementale, elle être écrite comme suit :

$$r_{E1,E2} = \frac{\sigma_{1,2}}{\sqrt{\sigma^2_{E1}\sigma^2_{E2}}} \quad 1.33$$

Plusieurs hypothèses sous-tendent la corrélation génétique entre caractères :

- la liaison génétique entre caractères

- la pléiotropie où un locus agit simultanément sur plusieurs caractères qui présentent le même déterminisme génétique

- le déséquilibre de liaison issu de la dérive génétique

Quant aux corrélations environnementales, elles sont associées à la répartition non aléatoire des effets du milieu et témoignent de la pléiotropie entre locus si les corrélations évoluent dans le même sens alors qu'elles illustrent une liaison si elles évoluent en sens inverse. L'incidence de la composante G×E sur le déterminisme des caractères est cruciale et mérite d'être intégrée lors de l'estimation de l'héritabilité des caractères dont la décomposition révèle l'existence d'une interaction entre les composantes génétiques et environnementales pouvant contribuer à la modification de l'ordre de classement des génotypes. En effet l'interaction G×E peut modifier l'ampleur et le signe de la corrélation génétique entre deux caractères appréciée dans le cadre d'une variabilité environnementale. Le différentiel d'expression des gènes en réponse à l'environnement renseigne sur la plasticité des génotypes lorsque le contexte environnemental change. Cette dernière conditionne l'adaptation des candidats à la sélection aux fluctuations environnementales.

1.3.2. Sélection phénomique

Malgré l'importance que revêt la sélection génomique, le déploiement de cette stratégie de sélection dans les programmes d'amélioration génétique se heurte à plusieurs contraintes portant notamment sur sa performance économique en lien avec le ratio coûts de génotypage et coûts de phénotypage. Dans ce contexte, la viabilité économique de la sélection génomique a fait l'objet de plusieurs études portant sur plusieurs espèces et un large éventail de scénarios. Elle s'est révélée prometteuse notamment pour des caractères dont l'héritabilité est inférieure à 0.25 à condition que le coût de phénotypage dépasse le coût de génotypage et que le nombre effectif de segments chromosomiques soit inférieur à 100 (Rajsic et al. 2016). Outre la contrainte économique, de nombreux QTL intéressants se situent dans des régions génomiques non codantes n'entraînant aucune altération de la séquence d'ADN. Il s'agit d'éléments de régulation. Il convient de noter que le polymorphisme causal de la plupart des mutations

affectant les caractères complexes demeure insaisissable (Wu et al. 2007; Te Pas et al. 2017). En effet, de nombreux caractères résultent d'un déterminisme moléculaire dicté par la variation de l'expression des gènes plutôt que de la variation des régions génomiques codantes (Cookson et al. 2009). De surcroit, le manque de connaissances sur les éléments de régulation et l'apport de l'épigénétique aux caractères d'intérêt agronomique contribue à la limitation du progrès génétique dérivant de l'adossement de la sélection à la prédiction de la valeur génétique par le biais de marqueurs moléculaires.

En alternative à la sélection conventionnelle ainsi qu'à la SG, diverses approches ont été proposées dans la littérature pour la prédiction des phénotypes. Elles sont fondées sur les endophénotypes que représentent le transcriptome, le protéome et le métabolome. Ces éléments permettent d'appréhender la complexité du phénome en vue d'une meilleure compréhension et interprétation des mécanismes moléculaires et biologiques qui sous-tendent les caractères à hérédité complexe. Par ailleurs, outre la variation de l'information génétique contenue dans le support de l'hérédité, l'expression du phénotype est également modulée par l'environnement dont l'impact se manifeste à plusieurs échelles : le génome, l'endophénome et le phénome. Dans cette perspective, l'ère omique a permis de caractériser la variation d'expression des gènes (la génomique), des ARNm transcrits (la transcriptomique), des protéines issues de l'expression du génome (la protéomique) et des métabolites (la métabolomique).

Elle se fonde sur des approches qui visent à élucider le niveau d'expression des gènes dans le but de comprendre et d'interpréter la variation phénotypique et de capturer les interactions entre les régions génomiques et l'environnement. Par conséquent, l'adoption de ces approches omiques en sélection et notamment dans la prédiction des valeurs génétiques des candidats à la sélection apporte une information supplémentaire par rapport à celle fournie par les marqueurs moléculaires. Toutefois, les coûts associés à l'obtention d'empreintes omiques et donc à la caractérisation des endophénotypes entravent leur déploiement à haute densité. C'est pourquoi (Rincent et al. 2018a) ont proposé une stratégie de sélection fondée sur la spectroscopie dans le proche infrarouge offrant une approche alternative aux approches omiques fondées sur les marqueurs endophénotypiques afin de capturer la variation intrinsèque aux réseaux de régulation complexes entre le génome et le phénotype. Dans ce sens, Rincent et al. (2018), dans leur étude de preuve de concept, ont souligné l'intérêt du recours aux spectres infrarouges pour pallier les inconvénients de la sélection génomique et prédire la performance agronomique des candidats à la sélection en valorisant des ressources disponibles à haut-débit et dont l'acquisition est rentable, facile à mettre en œuvre et non destructive ce qui permet d'élargir les

évaluations expérimentales pour couvrir une gamme diversifiée d'environnements et d'intégrer la variabilité spatio-temporelle dans le cadre de la sélection variétale.

Ainsi, la sélection phénomique est censée tenir compte des formes d'hérédité autres que le système génétique et qui conditionnent la transmission des caractères d'intérêt des géniteurs aux descendants. En effet, la variation au sein du système génétique peut dériver de diverses sources telles que la pléiotropie et de l'épistasie, se traduisant par la déviation par rapport à l'additivité et l'indépendance des actions des gènes (Mackay et al. 2009; Barton 2017). Au-delà de l'architecture génétique, existent d'autres composantes du système biologique qui interfèrent dans le processus d'expression du patrimoine génétique. Parmi celles-ci figure l'environnement qui interagit avec le réseau de gènes contribuant à la modification de la réponse du phénotype. Par conséquent, l'acquisition de spectres couvrant une large gamme de longueurs d'onde dans plusieurs environnements permet d'élucider les différentes échelles spatio-temporelles entre le génome et le phénome et donc de s'appuyer sur le déterminisme génétique couplé au déterminisme environnemental pour prédire les caractères d'intérêt.

En termes de modélisation statistique, la sélection phénomique bénéficie des avancées méthodologiques dérivant de l'intégration de la sélection génomique dans les schémas de sélection variétale. A l'instar de la sélection génomique, la sélection phénomique exploite les similarités génétiques entre les individus. En effet, cette approche de sélection repose sur l'estimation de la covariance génétique entre les individus évalués. Cependant, la covariance entre candidats sur laquelle s'appuie la sélection phénomique, à la différence de la sélection génomique, intègre outre la variation de la ségrégation mendélienne, la variation des séquences génomiques non codantes ainsi que l'interaction avec les facteurs environnementaux. Il s'agit donc d'une approche intégratrice des différents réseaux existant au sein de la carte génotype – phénotype. Par ailleurs, les arguments en faveur de la phénomique sont en concordance avec ceux de la génomique où les partisans de ces approches soulignent l'intérêt du recours aux données haut-débit afin d'appréhender la complexité des phénomènes biologiques (Houle et al. 2010). En conséquence, l'intérêt de la phénomique se fonde sur la disponibilité de données de phénotypage acquises à large échelle dans le cadre de programmes de sélection actuels. Cet intérêt permet de valoriser l'information génomique incluse dans 'la boite noire' sur laquelle repose le potentiel de la sélection génomique dont la supériorité dérive de l'expansion des dimensions des entrées du modèle de prédiction par rapport à la sélection assistée par marqueurs.

Au-delà des réseaux de liens de la carte génotype – phénotype, la spectroscopie occupe une place privilégiée dans plusieurs domaines. Un large éventail d'activités de recherches se concentre sur l'utilisation des spectres afin de prédire la composition chimique des systèmes biologiques (Bureau et al. 2009b, 2009c; Gebreselassie et al. 2017). En revanche, contrairement à l'utilisation classique des spectres infrarouges dans la prédiction des phénotypes d'intérêt, la sélection phénomique s'intéresse à des caractères indépendants des caractéristiques biochimiques des échantillons.

Dans l'article fondateur de la sélection phénomique, Rincent *et al.* (2018) ont démontré la capacité de longueurs d'onde couvrant la gamme spectrale allant de 8000 cm^{-1} à 10000 cm^{-1} à prédire des caractères d'intérêt économique et agronomique chez le blé et le peuplier. Dans ce cadre, deux scénarios ont été évalués. Dans le premier scénario, l'acquisition des spectres et la prédiction ont été effectuées dans le même environnement. Alors que dans le deuxième, l'environnement dans lequel a été conçu le modèle d'entraînement était différent de l'environnement dans lequel ont été recueillis les spectres. Les capacités prédictives des modèles de prédiction phénomique se sont révélées plus performantes que celles de la prédiction génomique notamment chez le blé. Par ailleurs, même si les absorbances sont spécifiques à un environnement donné, les prédictions phénomiques ont la même fiabilité dans des environnements différents. Par conséquent, un modèle de prédiction conçu dans un environnement bien déterminé peut être mobilisé pour prédire un phénotype dans un environnement différent.

1.4. Contexte de la création et la sélection variétale chez l'abricotier

1.4.1. Présentation de l'espèce

L'abricotier *Prunus armeniaca* L. est une espèce fruitière angiosperme dicotylédone appartenant à la famille des *Rosaceae*, à la sous-famille des *Prunoideae* et au genre *Prunus L* (Figure 9) regroupant plus de 200 espèces d'arbres et arbustes cultivés pour leurs productions fruitières ou pour leur valeur ornementale (Lichou and Jay, 2012). C'est une espèce diploïde comportant huit paires de chromosomes (2n = 2x = 16) et un génome de petite taille (294 Mb/n) (Arumuganathan and Earle, 1991) fortement synténique avec celui du pêcher *Prunus persica* (L.).

Plusieurs modes de multiplication s'offrent à l'abricotier suivant son origine géographique et l'objectif sous-jacent :

- Le mode le plus répandu est la ***multiplication sexuée par semis***. La graine résulte d'une autofécondation, si la variété mère est autofertile, ou d'une allofécondation (fécondation croisée) dans le cas contraire. Ce mode a permis la dissémination de l'espèce depuis le centre d'origine, il a conduit à la sélection naturelle de populations d'arbres à grand développement participant à l'ombrage des cultures dans les écosystèmes oasiens du Maghreb. Il est à la base de la sélection et de l'amélioration de l'espèce cultivée en permettant d'introduire de la variabilité génétique (recombinaison) à partir d'un cultivar connu par des caractéristiques intéressantes pour la sélection.
- L'***hybridation contrôlée*** par le biais de la pollinisation croisée entre un cultivar mâle et un cultivar femelle, après la castration des organes mâles sur les fleurs de la variété mère. Ce mode de reproduction, à la base de la plupart des variétés modernes, est mis en œuvre afin d'intégrer des caractères d'intérêt agronomique chez des géniteurs et d'accompagner leur transmission aux descendants. Cependant, il se heurte à des contraintes biologiques liées à l'incompatibilité pollinique qui est fréquente au sein de l'espèce.
- La ***multiplication végétative par greffage*** **:** elle permet de reproduire des accessions présentant des caractéristiques recherchées (hybrides et variétés patrimoniales) à l'identique pour les déployer en verger en vue de les conserver au fil des générations.

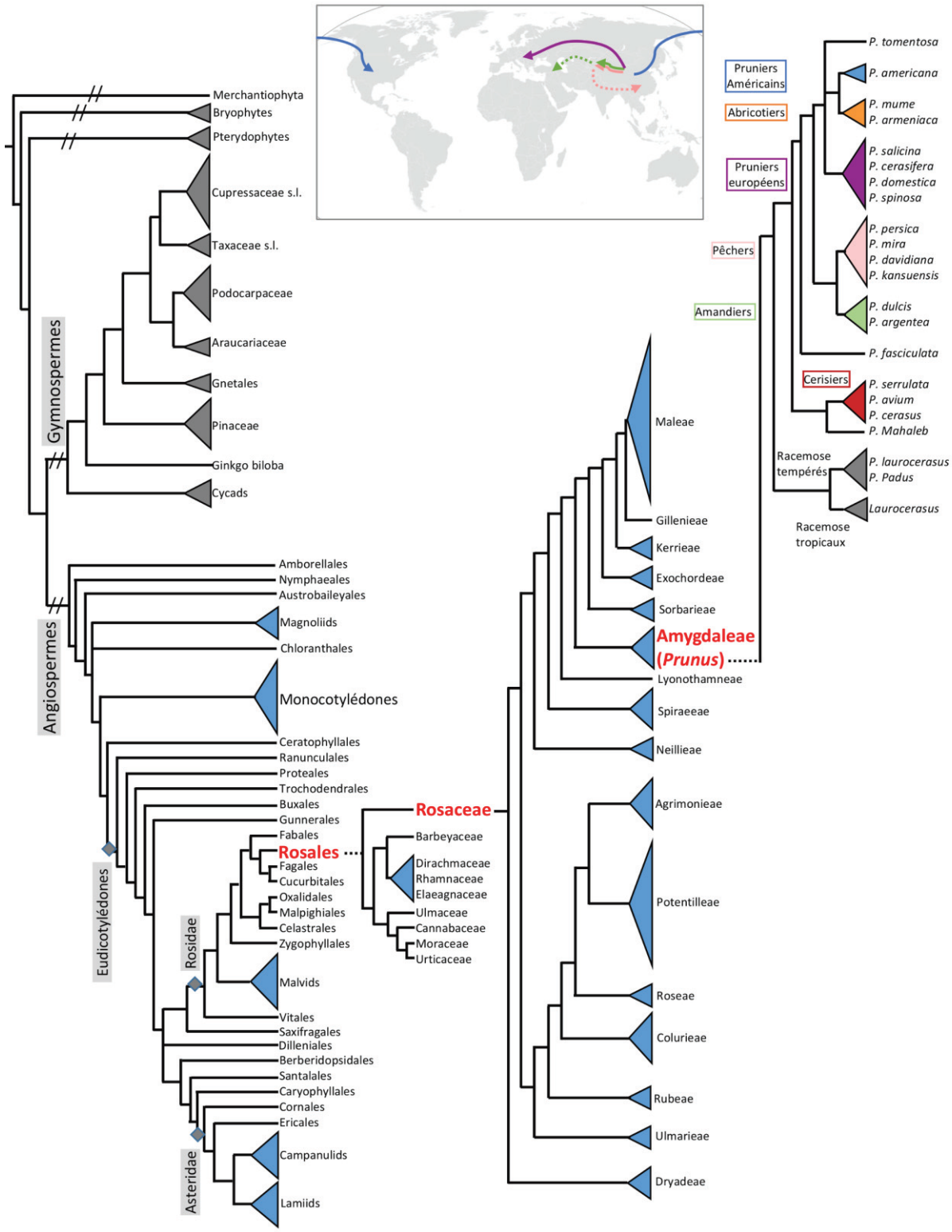


Figure 9: Représentation graphique de la position des *Prunus* et leurs routes de diversification (Van Ghelder 2019)

Illustration des relations phylogénétiques entre les espèces appartenant au genre *Prunus* et leurs routes de propagation à partir de leur centre d'origine (Asie centrale).

1.4.2. Economie de l'abricot

L'abricotier est une espèce fruitière qui contribue de manière substantielle à l'économie agricole française. La tendance à la hausse de la production d'abricot est une réponse à une demande accrue tant du marché local français que des marchés d'export, ce qui lui confère une valeur stratégique sur le plan économique. En l'occurrence, la France occupe une place prépondérante dans l'union européenne et dans le monde en termes de volume de production. En Europe, la France se place en troisième position après l'Italie et l'Espagne et avant la Grèce, malgré une modeste récolte de 93 000 tonnes en 2020 (Agreste 2020), mais une moyenne interannuelle de 130 000 tonnes de 2016 à 2018, soit 3.25% de la production mondiale qui avoisinait 4.0 millions de tonnes sur cette période (FranceAgriMer 2020) (Figure 10). La France est le huitième producteur d'abricot dans le monde, le bassin méditerranéen assurant quant à lui, plus de la moitié de la production mondiale (FAO 2018).

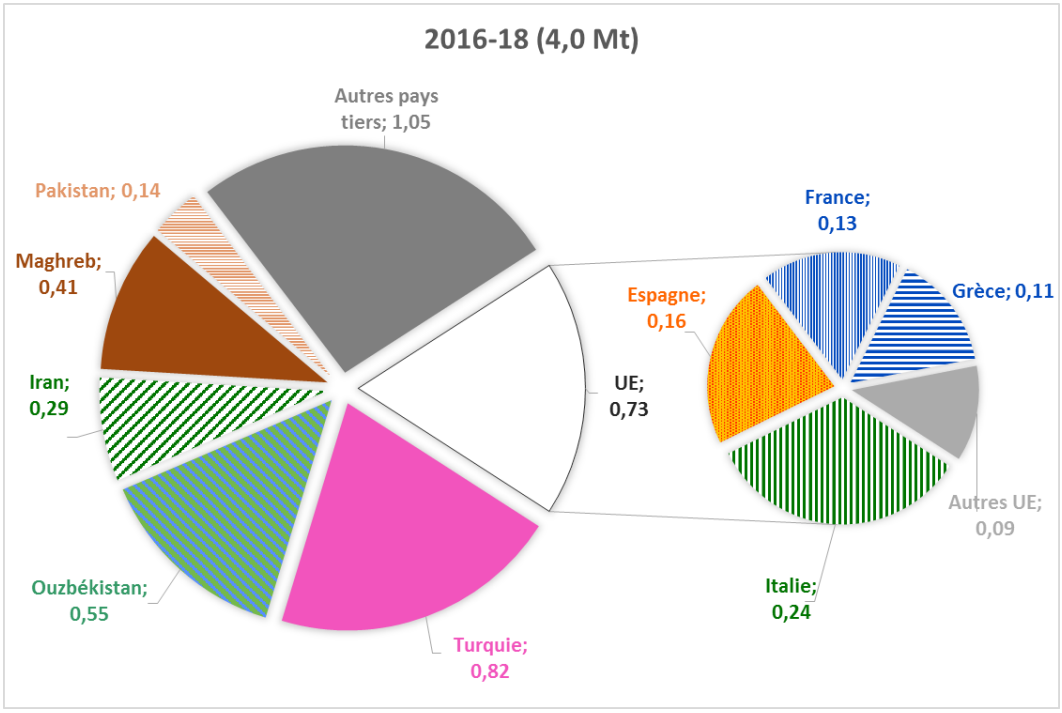


Figure 10: Production d'abricot par pays (1000 tonnes)

La production de l'abricot garde un attrait non négligeable pour les arboriculteurs français, ce qui justifie une grande stabilité de la superficie de cette espèce dans les exploitations avec 12 280 ha en 2019, en légère progression par rapport aux 12 014 ha cultivés en 2015, soit environ 7% du territoire agricole dédié à l'arboriculture fruitière avec un poids relatif estimé à 9% en valeur dans les exportations françaises de fruits (FranceAgriMer 2020) (Figure 11A).

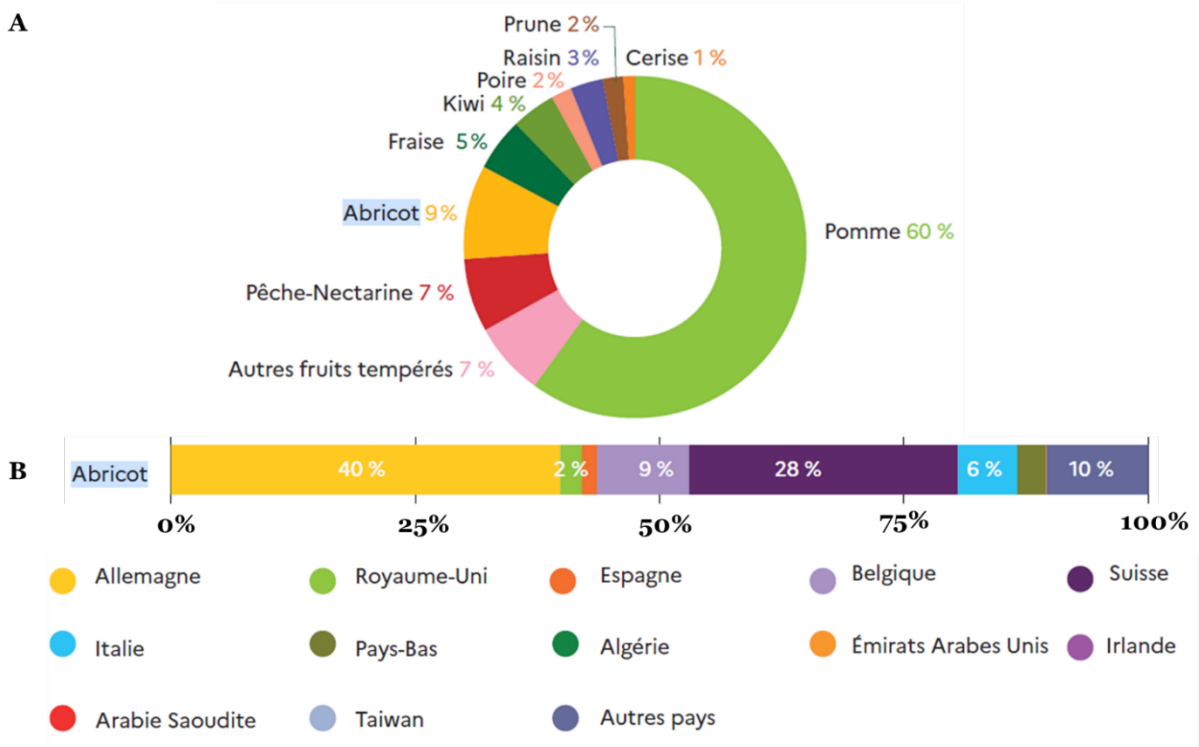


Figure 11: Poids relatif des différents fruits tempérés dans les exportations françaises (A) et pays clients pour l'abricot (B) pour l'année 2019 (% en valeur) (FranceAgriMer 2020)

En revanche, la filière abricot évolue dans un contexte fragile où les aléas climatiques et phytosanitaires impactent très fortement la production, pénalisant les tonnages produits et récoltés, et la qualité de l'offre. En effet, pour la troisième année consécutive, la campagne d'abricot a présenté une baisse de la production due aux conditions climatiques caractérisées par des épisodes de gel et de grêle ou des anomalies florales consécutives à des hivers doux. L'année 2020 est ainsi l'une des plus faibles productions de la décennie avec une réduction des tonnages estimée à -28% à l'échelle européenne et -29% en France en comparaison avec la campagne de 2019. A l'échelle nationale, le potentiel de production accuse une perte de 30% en région Provence-Alpes-Côte d'Azur, de 29% en Occitanie et de 27% en Auvergne-Rhône-Alpes par rapport à l'année 2019. Face à ce déficit de production, le marché français s'est trouvé sous-approvisionné en 2020, et dépendant des importations (abricot d'origine espagnole, à plus de 90%). Malgré une récolte en baisse, le chiffre d'affaires national pour ce fruit a augmenté en 2020 grâce à la fluidité des ventes et à la fermeté des prix (FranceAgriMer 2020). La production de l'abricot français s'inscrit dans un marché national, et européen, avec des débouchés en Allemagne (40%), en Suisse (28%), en Belgique et en Italie (Figure 11B). L'abricot français destiné à l'export se situe en aval du calendrier de production espagnol ce qui lui laisse une place sur le créneau tardif (Juillet – Août) alors que les fruits espagnols sont très présents sur le

créneau précoce (mai – juin). Dans ce contexte, la France est un acteur important de l'approvisionnement des marchés internationaux en abricot avec des volumes expédiés compris entre 20 000 et 60 000 tonnes.

1.4.3. Exigences culturales et climatiques

En France, les vergers d'abricotier couvrent trois grandes régions de production, le Roussillon, la basse Vallée du Rhône et la moyenne Vallée du Rhône (Figure 12). Chacune d'entre elles s'appuient sur un pool de variétés locales traditionnelles : populations dérivées de Rouge du Roussillon dans le Roussillon, populations dérivées de Colomer, Fournes dans la basse vallée du Rhône, populations dérivées de Bergeron et Polonais (ou Orangé de Provence) dans la Moyenne vallée du Rhône. Ces régions sont caractérisées par des conditions pédoclimatiques contrastées allant d'une zone à hiver doux dans le Roussillon à une zone au climat continental dans la vallée du Rhône. Les panels variétaux traditionnels bien adaptés aux différentes régions d'origine sont inadaptées dans les autres régions, ce qui est une caractéristique majeure de l'espèce. L'un des enjeux de la culture réside dans l'élargissement du calendrier variétal dans chacune des régions de production en prenant en compte les caractéristiques d'adaptation locale. Cet enjeu spécifique s'ajoute à une demande générique des consommateurs et plus largement de la société pour l'installation de vergers économes en intrants, notamment en produits phytosanitaires. L'ambition de changer les paradigmes de la production fruitière est d'actualité. Cette ambition de conception de variétés d'abricotier adaptées à des itinéraires techniques plus respectueux de l'environnement se heurte à plusieurs contraintes en lien avec les caractéristiques biologiques et génétiques de l'espèce. En effet, l'adoption de nouveaux systèmes de culture est contrainte par la sensibilité de l'espèce à une large gamme d'agents pathogènes entravant la volonté de réduire l'utilisation des produits phytosanitaires pour maintenir sa productivité. Peuvent s'ajouter les problématiques de protection du verger, ceci dépendant des exigences culturales inhérentes à l'espèce. A titre d'exemple, l'installation d'un verger d'abricotier est tributaire non seulement du choix de la variété mais également de celui du porte-greffe. Afin de parfaire ce choix, le porte-greffe doit répondre à la problématique d'adaptation aux conditions pédoclimatiques du verger et de compatibilité avec le greffon. L'incompatibilité de greffe se manifeste par un décollement entre le porte-greffe et le greffon lorsque les variétés sont greffées sur des porte-greffes botaniquement éloignés de l'espèce abricotier (greffage sur pêcher ou prunier). Par ailleurs, la production d'abricot étant concentrée en France dans des zones géographiques restreintes et basée sur une gamme variétale réduite en raison d'une plasticité d'adaptation variable, cette situation induit une fragilité majeure dans

la plupart des zones de production. A titre d'exemple, le cultivar 'Bergeron' était à la fin des années 2010 une variété de base, représentant 29% de la structure variétale du verger français en 2009 (Lichou and Jay 2012) avec une zone de production dominante dans la région Rhône-Alpes.

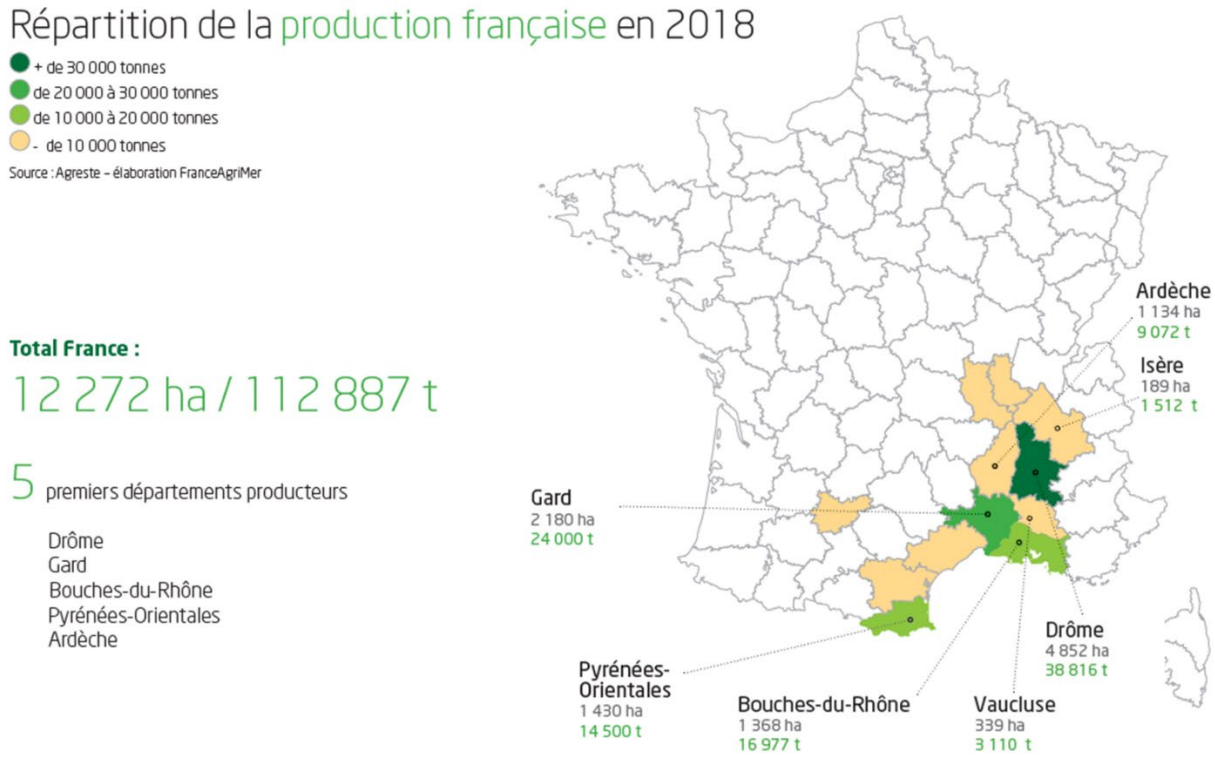


Figure 12: Répartition de la production française (FranceAgriMer 2019)

Au plan climatique, l'abricotier est caractérisé par des exigences qui différent selon les variétés. Parmi celles-ci figurent les besoins en froid pour la levée de la dormance auxquels s'ajoutent les besoins en chaleur pour la croissance des ébauches florales et ensuite pour le développement et la maturité des fruits. Par conséquent, des déficits en températures basses peuvent engendrer des perturbations de débourrement, des anomalies florales et des chutes de bourgeons. Egalement, une satisfaction lente des besoins en chaleur retarde le déroulement de la croissance végétative et les processus physiologiques sous-jacents, et par conséquent une modification des calendriers de maturité des fruits, voire des chutes de fruits durant le processus de croissance.

1.4.4. Acteurs français de la filière abricot

A la lumière de telles contraintes, la réponse variétale aux enjeux en lien avec l'extension des aires de production et la durabilité des vergers dans un contexte de réchauffement climatique

va être prise en compte par un continuum des acteurs Recherche et Développement en lien avec les organisations professionnelles (Figure 13):

- **L'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE)** procède à la caractérisation de la diversité génétique et l'identification des déterminants génétiques associés aux caractères d'intérêt notamment la qualité des fruits, la résistance aux maladies et la régularité de la production. Il est en charge :
    - des ressources génétiques (maintenance, caractérisation, gestion) ; il s'agit d'une mission nationale conduite pour partie sous mandat du GEVES pour les activités de DHS (étude des nouvelles variétés en vue de leur inscription au catalogue national et européen des variétés cultivées) et de leur protection (sous mandat de l'OCVV), pour partie sous mandat du Ministère en charge de l'Agriculture à des fins de préservation de la diversité génétique.
    - des programmes de recherche destinés à répondre aux questions des professionnels (caractérisation du matériel végétal, notamment de la qualité des fruits et de la sensibilité aux bioagresseurs) et à développer les outils et méthodes mobilisables par ces acteurs (voir infra).
    - des programmes de prébreeding destinés à identifier et mettre à disposition les géniteurs et les méthodes à mobiliser dans le cadre des programmes de breeding pour adresser les cibles « agroécologie » et « adaptation au changement climatique » priorisées par l'Institut.
- **Le centre technique interprofessionnel des fruits et légumes (CTIFL)** est l'autorité compétente en charge de l'évaluation variétale et de la certification fruitière (contrôle de l'authenticité et de l'état sanitaire des variétés en multiplication) se basant sur la sélection conservatrice, la production des plants de base et des greffons et le contrôle sanitaire des variétés et porte-greffe certifiés. Le CTIFL coordonne des réseaux d'expérimentations axés sur l'évaluation des variétés et porte-greffe. Il vise à caractériser l'intérêt agronomique et commercial des innovations variétales en se basant sur des essais multisites et pluriannuels.
- L'évaluation variétale porte en premier lieu sur le comportement agronomique et la qualité des fruits, et elle est complétée depuis peu par l'accès à des caractères de sensibilité aux bioagresseurs : la sensibilité à la sharka et au chancre bactérien (en partenariat avec les Stations Régionales d'expérimentation).

- Les Stations Régionales d'expérimentation : CENTREX, Sud Expé (ex-SERFEL), SEFRA qui assurent l'évaluation des performances agronomiques des variétés candidates dans les différentes régions de production sous la coordination du CTIFL.
- Dans le cadre de sa démarche de recherche partenariale, INRAE s'appuie aussi sur des partenariats public-privés. Citons le Centre d'Expérimentation des Pépinières (CEP) qui regroupe des pépiniéristes ayant pour mission l'édition sous certification de variétés et porte-greffe fruitiers obtenus par INRAE (signature des contrats d'édition exclusive en 2008). CEP a pour mission de multiplier et protéger, par le dépôt de Certificats d'Obtention Végétale (COV), les obtentions INRAE ayant témoigné d'un potentiel commercial. Citons aussi deux sociétés dérivées de CEP, Novadi et CEP Innovation en charge des programmes de breeding et d'évaluation chez les espèces fruitières. Le partenariat INRAE avec CEP Innovation et Novadi initié autour les espèces pommier et abricotier, intègre aujourd'hui d'autres espèces fruitières majeures telles que le pêcher, le cerisier, le poirier et les porte-greffe, il est formalisé dans le cadre d'une convention générale de partenariat. En place depuis plus de 10 ans, ce partenariat s'opère dans une optique de valorisation des efforts déployés par la recherche publique et d'accompagnement de l'innovation variétale conçue par INRAE. Ainsi, le secteur privé participe à la création de variétés qui répondent aux enjeux des différents circuits de la filière abricot s'inscrivant dans le cadre de la conception d'idéotypes variétaux qui intègrent la résistance aux maladies notamment à la sharka, l'autofertilité pollinique, la qualité des fruits et la phénologie.
- Au terme de l'évaluation variétale le **Comité Technique Permanent de la Sélection des plantes cultivées (CTPS)**, instance paritaire sous la tutelle du ministère de l'agriculture, examine les demandes d'inscription des variétés au catalogue officiel des variétés cultivées en France et au niveau européen. Les nouvelles obtentions, pour être inscrites, doivent être différentes des variétés déjà commercialisées et exemptes de maladies de quarantaine. Pour les variétés fruitières, le CTPS fonde ses décisions de proposition d'inscription au catalogue sur des critères de distinction, d'homogénéité et de stabilité (épreuves DHS) et sur l'état sanitaire des variétés candidates. Concernant la valeur agronomique, technologique et environnementale (VATE), contrairement à d'autres plantes agricoles, elle ne fait pas partie du processus d'inscription et n'est évaluée qu'en post-inscription.
- **Les producteurs** peuvent rejoindre les organisations de producteurs (OP), lesquelles peuvent se regrouper en association d'OP (AOP) pour bénéficier de soutiens financiers

dans le cadre de l'organisation commune des marchés de l'Union Européenne. L'Association d'Organisation de Producteurs « Pêches et Abricots de France qui réunit et met en marché près de 60% de la production nationale d'abricot, est ainsi un interlocuteur représentatif des professionnels dans la définition des choix et des orientations de recherche et développement au service de la filière.

- En aval de la filière figurent les différents **opérateurs des circuits de distribution et de commercialisation** auxquels s'ajoute **le consommateur** qui représente un maillon dont les préférences orientent les activités en amont de la filière et sont pris en compte lors de la redéfinition des choix variétaux et des débouchés de la production.

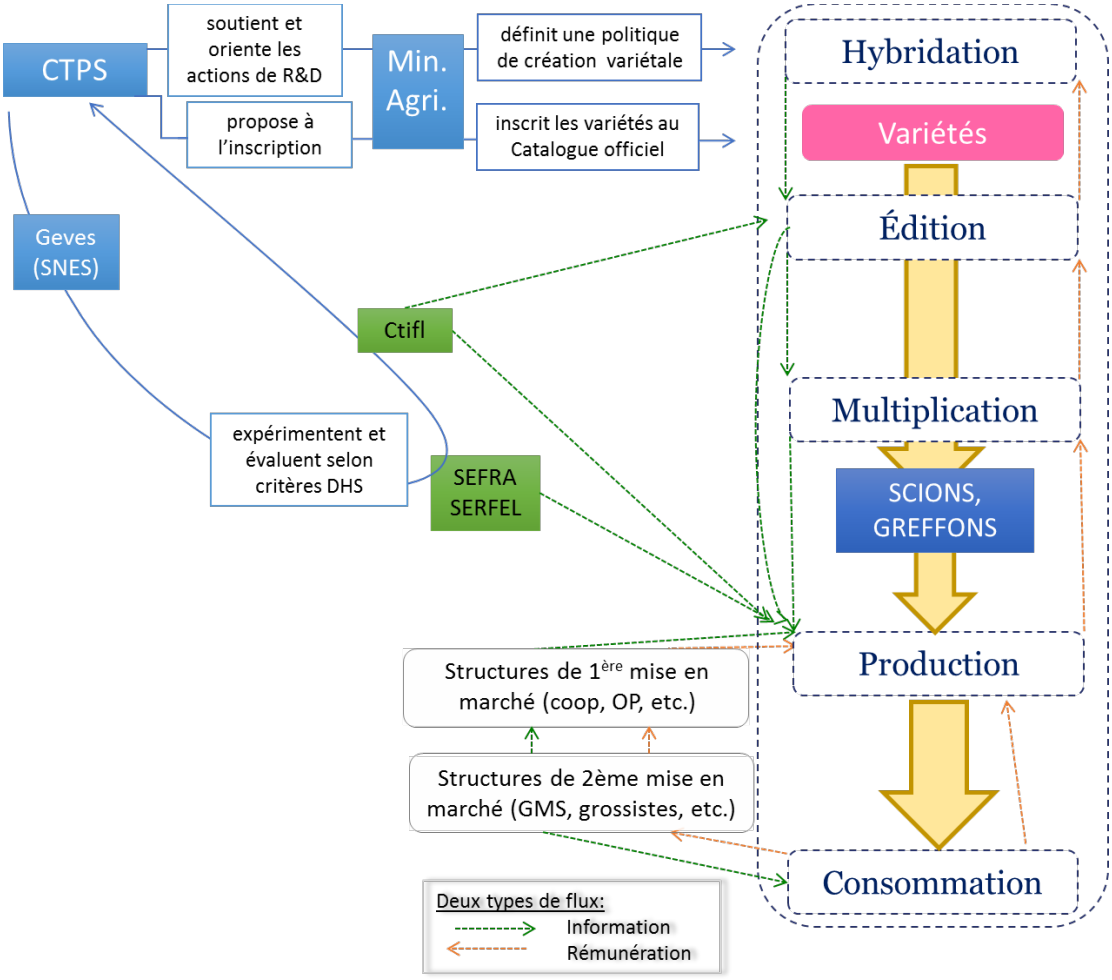


Figure 13: Organisation de la filière abricot (Lamine et al. 2017)

Abréviations : GMS : Grandes et Moyennes Surfaces; coop : Coopératives; OP : Organisations de Producteurs; DHS : Distinction Homogénéité Stabilité; SEFRA : Station d'Expérimentations Fruits Rhône-Alpes; SERFEL (intégrée dans SudExpé en 2017): Station d'Expérimentation Régionale pour les Fruits et Légumes; Min Agri. : Ministère de l'Agriculture et de l'Alimentation; SNES : Station Nationale d'Essais de Semences.

C'est dans ce cadre que s'inscrit l'organisation de la filière abricot s'appuyant sur la contribution coordonnée des acteurs publics et privés dont les activités s'inscrivent dans une réflexion visant à rénover le secteur tout en garantissant sa performance et sa durabilité et de répondre aux nouveaux enjeux en lien avec la segmentation variétale indispensable à la dynamisation du marché. En effet, la mise à disposition de plusieurs variantes d'un même produit offertes à des fourchettes de prix différentes permet de consolider la gamme variétale existante et d'élargir le calendrier de production et de commercialisation. De surcroît, la segmentation du marché vise à élargir la palette variétale par des produits à date de maturité échelonnée, depuis des variétés très précoces jusqu'à des variétés plus tardives afin de compléter l'offre et pallier la brièveté de la période de commercialisation de l'abricot. Il convient de signaler que l'enjeu de l'allongement du calendrier de production se heurte à la prolongation de la période d'exposition des variétés aux bio-agresseurs et aux fluctuations climatiques ce qui se traduit dans le cas de stress biotiques par le recours à l'utilisation des produits phytosanitaires afin de garantir la protection des variétés tardives vis-à-vis de certaines maladies (Lamine et al. 2017). Ainsi, l'intensification des systèmes de culture rime avec mise en jeu des leviers sur lesquels s'articule la filière à savoir la durabilité de l'offre variétale et le respect de l'environnement.

1.4.5. Création et sélection variétale

L'essor de l'abricotier est associé à la diversité de la palette variétale grâce aux efforts déployés dans la conception des programmes de création et de sélection des variétés s'articulant autour de la recherche et la fixation de combinaisons alléliques favorables par le biais de l'observation des performances des candidats à la sélection et l'identification des individus conformes aux objectifs de sélection afin de concevoir des variétés répondant aux attentes du marché et des acteurs de la filière. Dans ce contexte, les enjeux de la création variétale chez l'abricotier peuvent se résumer comme suit :

- L'élargissement des gammes variétales pour chacune des régions de production afin de compléter la gamme variétale existante avec une démarche de segmentation produit (abricots classiques, bicolores, rouges et demain blancs) tout en maintenant la qualité des fruits afin de pallier les problèmes liés à la culture monovariétale. C'est le premier objectif qui a été travaillé afin de donner la possibilité d'une extension de la production à l'échelle de chacune des exploitations.
- L'écologisation de la production qui a été initiée en ciblant les maladies les plus graves et conduisant à la disparition des vergers comme la sharka et le chancre bactérien. Elle

se poursuit aujourd'hui par l'élargissement des cibles en vue d'une réduction des traitements phytosanitaires (monilia sur fleurs, rouille, oïdium et tavelure sur fruits).

- L'adaptation au changement climatique en relation avec la gestion des problèmes adaptatifs dont nous avons indiqué précédemment qu'il s'agissait d'un des facteurs limitant au déploiement de calendriers variétaux par région de production, et qu'il n'était pas indépendant de l'origine phylogénétique des variétés.

1.4.5.1.Sélection conventionnelle

Déployés par INRAE depuis les années 1970, les programmes d'amélioration de l'abricotier ont tout d'abord été basés sur des travaux de prospection et sélection clonales (Huet 1961) avec la mise en comparaison des différentes variétés cultivées et la vérification de leur état sanitaire. Cette approche, après avoir permis l'identification des clones les plus performants (Rouge du Roussillon A157, Bergeron A660) encore aujourd'hui cultivés, a été complétée par les premiers travaux de création variétale qui ont visé dans un premier temps à élargir le calendrier variétal par des croisements entre variétés issues des phylum européen et méditerranéens (Carraut and Crossa-Raynaud 1981). Des variétés comme Héléna du Roussillon, Royal Roussillon puis Mariem, Malice, Ivresse sont ainsi venues compléter les gammes variétales dans chacune des régions de production. Dans le même temps des ressources génétiques issues d'autres régions de culture (autres pays méditerranéens, USA, Canada, Italie, Grèce et Iran) étaient introduites et étudiées par comparaison avec les variétés locales afin d'étendre la gamme des variétés cultivées. Des variétés comme Goldrich, Hargrand, Orangered® Barth ont émergé et ont été introduites et leurs fonds génétiques ont été mobilisés en croisement pour élargir les calendriers variétaux. Avec elles, des caractères nouveaux tels que fruit doux, bicolore, de gros calibre ont été introduits, mais aussi l'incompatibilité pollinique et l'émergence d'anomalies florales inconnues dans les fonds génétiques nord-méditerranéens. C'est alors qu'ont été initiés les travaux concernant la résistance variétale aux bioagresseurs (Sharka et Chancre bactérien, notamment). Ils ont établi les bases de travaux conduits par INRAE pour introgresser ces caractères dans les fonds génétiques cultivés, approche qui est un des fondements du programme conduit actuellement, avec le maintien de la diversité génétique, la maitrise de la qualité des fruits et la régularité de production.

En prenant du recul par rapport à l'évolution du programme, nous constatons que les processus de sélection et de création variétale chez les espèces fruitières pérennes et notamment l'abricotier sont entravés par des caractéristiques inhérentes à la biologie des espèces pérennes fruitières telles que la pérennité et la longueur de la période juvénile qui varie de 3 à 8 ans. De

plus, le mode de reproduction de l'abricotier (allogamie préférentielle) qui entrave la fixation des allèles recherchés et l'hétérozygotie qui masque les allèles récessifs potentiellement intéressants, s'ajoutent aux contraintes du progrès génétique. Outre les contraintes biologiques, la sélection se heurte à la méconnaissance de l'héritabilité et l'architecture génétique des caractères d'intérêt et à la forte contribution des facteurs environnementaux à la variation phénotypique. Dans un tel contexte, le schéma de sélection de l'abricotier se base classiquement sur l'évaluation et la prospection des ressources génétiques disponibles dans des vergers d'évaluation du comportement agronomique, le déploiement de croisements destinés à recombiner les traits d'intérêt et l'identification des présélections dans les populations hybrides par la sélection massale, puis leur évaluation dans le cadre d'essais multisites et pluriannuels pour identifier les meilleures sélections sur la base des comportements agronomiques par rapport à des variétés de référence. La sélection massale, en s'appuyant sur la caractérisation phénotypique de matériels évalués dans des dispositifs expérimentaux multilocaux et pluriannuels, est ralentie par la durée des cycles de sélection qui est imposée par le besoin d'une régularité des performances interannuelles. Par conséquent, l'augmentation de la fréquence des allèles favorables dans le cadre de la sélection massale est progressive et lente.

En définitive, le pilotage des programmes d'amélioration chez l'abricotier est fondé sur la sélection massale et la sélection récurrente. Ce schéma repose sur l'hybridation contrôlée via la fécondation croisée entre géniteurs dotés de caractéristiques phénotypiques complémentaires en vue de leur transmission aux descendants qui présenteront une complémentarité de caractéristiques favorables issues de leurs parents. Il s'agit d'un processus long et difficile à mettre en œuvre s'articulant sur des cycles qui varient entre 7 et 15 ans. A l'instar de la sélection, le processus de création variétale est d'autant plus long que l'idéotype recherché s'éloigne des variétés de référence. Dans ce sens, le temps requis pour concevoir et évaluer une nouvelle variété est compris entre 15 et 20 ans.

1.4.5.2. Sélection assistée par marqueurs

La sélection assistée par marqueurs (SAM) est une stratégie de sélection fondée sur l'identification de polymorphismes génétiques responsables de la variation du caractère d'intérêt. Cette stratégie fait appel à la génétique quantitative pour éclairer le contrôle génétique du caractère cible dans le but d'introgresser des combinaisons alléliques favorables dans les variétés élites. La SAM est fondée sur les études de liaison et d'association génétiques entre marqueurs et caractères. Après avoir identifié les QTL qui définissent l'architecture génétique des caractères désirables, elle va se baser sur l'utilisation de marqueurs étroitement liés aux

QTL identifiés. La SAM est mobilisée depuis près de 10 ans chez l'abricotier dans le cadre de l'innovation variétale s'inscrivant dans une démarche de sélection mono-caractère ciblée pour protéger l'espèce de contaminations par la sharka dont l'extension inexorable mettait en péril le devenir de la culture. Dans ce sens, le recours à la SAM, couplé à une caractérisation phénotypique, en création variétale a permis la mise au point de la gamme 'Aramis' regroupant des variétés résistantes à la sharka, dans le cadre du partenariat entre INRAE et CEP innovation. Elle est aussi mobilisée pour éliminer les variétés auto-incompatibles afin de ne maintenir en sélection que les variétés n'ayant pas besoin de pollinisation croisée en verger de production. Les travaux de recherche et de prébreeding actuels visent à élargir le spectre des caractères pris en compte afin de réduire l'utilisation des intrants, notamment les pesticides, dans l'optique de la conception de vergers écoresponsables et ceci concerne tout particulièrement la sensibilité au chancre bactérien, au monilia sur fleurs, à la rouille et dans une moindre mesure à l'oïdium.

En définitive, la sélection chez l'abricotier est axée sur le développement de variétés répondant aux besoins du marché et associant productivité, qualité et attractivité du fruit, moindre sensibilité aux bioagresseurs. Elle repose sur la mise au point de variétés où constitution génétique et performances agronomiques sont prises en compte. Elle est fondée sur une stratégie intégrant sélection conventionnelle (pour les performances agronomiques) et SAM (pour quelques caractères d'intérêt). Elle mobilise des descendances biparentales dérivées de croisements entre variétés qui présentent des caractères agronomiques contrastés permettant ainsi d'introgresser des caractères d'intérêt via des rétrocroisements. Afin d'élargir la base génétique des populations sélectionnées, le recours à la diversité s'effectue par le biais de la mise en place de plans de croisements où sont combinées des accessions issues de collections de ressources génétiques représentatives de la diversité disponible pour cette espèce (Bourguiba et al. 2012; Bourguiba et al. 2020).

1.4.5.3. Caractères sélectionnés

Les caractères agronomiques d'intérêt sur lesquels se concentrent les efforts de sélection et de création variétale chez l'abricotier portent sur la résistance des cultivars aux bioagresseurs et l'adaptation aux conditions climatiques. Peuvent s'ajouter à ces caractères, la qualité des fruits et la phénologie. Ces critères de sélection sont pris en compte par les expérimentateurs impliqués dans l'évaluation variétale au sein de la charte nationale d'expérimentation fruitière.

1.4.5.3.1. Qualité des fruits

La qualité des fruits est un ensemble complexe intégrant des composantes perçues par les consommateurs comme l'attractivité à travers la couleur, la saveur faisant référence à l'équilibre sucres – acides, la texture et les arômes. Du point de vue des producteurs et distributeurs, la qualité est perçue à travers la fermeté et la vitesse d'évolution vers la maturité. L'évaluation des composantes de la qualité est un outil d'aide à la détermination du stade optimal de récolte pour une variété donnée afin de limiter l'altération des fruits et optimiser leur durée de vie en post-récolte. Plusieurs critères interviennent dans l'élaboration de la qualité de l'abricot, parmi lesquels figurent :

- La fermeté qui représente une caractéristique primordiale déterminant la fenêtre de la maturité commerciale. En effet, la maturation induit une perte de fermeté, synonyme d'un relâchement de la paroi dû à la dégradation des pectines et des hémicelluloses (Brummell et al. 2004).
- L'aptitude des fruits à la conservation après la récolte y compris la résistance aux manipulations et aux contraintes physiques subies par les fruits lors de leur acheminement dans les circuits de transport et de distribution. Elle doit être suffisante afin de satisfaire les attentes des acteurs en aval de la filière.
- La coloration des fruits est un critère esthétique définissant son attractivité. Son importance réside également dans sa capacité à appréhender le stade de maturité et donc la date optimale de récolte des fruits. La coloration de fond de l'épiderme et de la chair est associée à l'accumulation de pigments colorés tels que les caroténoïdes et dans une moindre mesure les polyphénols. Les caroténoïdes sont des pigments solubles dans les lipides qui s'accumulent au sein des chromoplastes des cellules durant la phase climactérique et jusqu'à la récolte. Ils confèrent aux fruits la couleur orangée. Les caroténoïdes sont classés en deux groupes distincts : les carotènes, incluant le β-carotène et le lycopène, et les xanthophylles telle la lutéine. Les variétés dotées de fruits orangés accumulent préférentiellement le β-carotène, ce qui est le cas de la variété Goldrich, alors que celles caractérisées par des fruits de couleur crème, comme Moniqui, accumulent des pigments incolores tels le phytoène et le phytofluène. La maturation des fruits s'accompagne d'un changement de la coloration suite à la dégradation des chlorophylles et à la biosynthèse des caroténoïdes. Quant aux polyphénols, ils sont représentés par deux classes : les acides phénoliques et les flavonoïdes. Ces derniers sont répartis en 3 sous-classes : les flavan-3-ols, les flavonols et les anthocyanes. La

surimpression rouge des fruits, fréquente chez de nombreuses variétés récentes, résulte de l'accumulation d'anthocyanes dans les couches superficielles de l'épiderme. Par opposition aux caroténoïdes, les anthocyanes sont hydrosolubles et peuvent diffuser dans la chair du fruit pour donner un caractère sanguin qui pourrait être à la base d'une segmentation variétale).

- La production éthylénique renseigne sur la vitesse d'évolution du fruit dès que la maturation est engagée et après la récolte. En effet, le dégagement d'éthylène est prononcé chez les variétés dont les fruits évoluent très rapidement vers la maturité alors que les variétés à production éthylénique faible se caractérisent par des fruits dont l'évolution vers la maturité est plus lente. Chez les premières, la capacité de conservation est très limitée, par suite d'une activité métabolique intense qui écourte leur durée de vie, au contraire des variétés à faible production éthylénique dont l'aptitude à l'entreposage est meilleure.
- La saveur des fruits repose en grande partie sur l'équilibre sucres / acides donc sur l'accumulation de métabolites primaires dans le fruit jusqu'au moment de la récolte. La teneur en sucres solubles, renseignée par l'indice de réfraction (IR), a tendance à augmenter à l'approche de la maturité. Les variétés tardives d'abricot sont caractérisées par un indice de réfraction plus élevé que celui des variétés précoces. Quant à l'acidité, elle est appréciée par la mesure de l'acidité titrable ou par le dosage des teneurs en acides organiques, notamment l'acide citrique et l'acide malique issus essentiellement du métabolisme des sucres dans le fruit. Contrairement à la teneur en sucres, l'acidité est stable quelle que soit la charge des arbres en fruits. Cependant, après la récolte, la teneur en sucres reste stable alors que celle des acides organiques a tendance à diminuer. La diminution des acides est d'autant plus importante que la durée d'entreposage et la température de stockage sont importantes. Le rapport entre la teneur en sucres et la teneur en acides définit la qualité gustative du fruit. L'Indice de Réfraction est l'indicateur prépondérant de la qualité des fruits telle qu'appréciée dans le cadre des études consommateurs conduites par le CTIFL (Scandella and Vernin 2019).

La qualité des fruits résulte de la contribution de trois composantes majeures : la variété qui conditionne le potentiel qualitatif, les conditions de culture et de récolte qui déterminent le potentiel à la récolte et la maitrise des fruits après récolte.

A titre d'exemple, la charge des arbres conditionne la relation source-puits. Par conséquent, l'éclaircissage peut se traduire par une augmentation et une homogénéisation du calibre des fruits compensant une éventuelle diminution du rendement (Lichou and Jay 2012). Ainsi, la gestion des itinéraires techniques (irrigation, fertilisation, taille, éclaircissage) et la maîtrise des modes de conduite en verger et après-récolte (température de conservation et durée de conservation) représentent autant de paramètres déterminant la qualité des fruits.

1.4.5.3.2. Résistance aux maladies

Outre l'amélioration de la qualité agronomique et technologique, la cible de la sélection chez l'abricotier intègre la recherche de facteurs de résistance s'inscrivant dans une ambition visant à préserver la viabilité du verger et à limiter le recours aux produits phytosanitaires en vue de la mise en place de vergers économes en intrants. En effet, l'abricotier est caractérisé par sa sensibilité à un large spectre de bioagresseurs responsables de maladies ayant une incidence majeure sur la pérennité du verger. Dans ce cadre, la protection du verger s'avère cruciale pour garantir la durabilité de la production. Elle implique l'étude du pathosystème abricotier - agent pathogène dans la perspective de l'identification des stratégies de lutte et la mise en place des mesures prophylactiques adéquates.

Dans ce contexte, le focus a porté dans un premier temps sur les trois maladies responsables de mortalité d'arbres en verger : l'enroulement chlorotique (ECA), le chancre bactérien et la sharka. Il s'est ensuite étendu aux maladies préjudiciables au rendement et à la durabilité de la production telles que la **moniliose** sur fleurs, l'**oïdium** et la **rouille**. La survenue d'aléas climatiques durant la phase de floraison engendre des pertes importantes de récolte associées à la moniliose sur fleurs (*Monilia laxa*) dont les dégâts sont observés sur fleurs et rameaux (Figure 14). Quant à l'oïdium, il s'agit d'une maladie qui occasionne des dégâts autant sur fruits que sur feuilles mais dont l'impact en dehors des variétés à épiderme rouge reste limité (Figure 15A). Concernant la rouille, c'est une maladie d'importance économique secondaire qui induit dans le cas d'attaques précoces la chute des feuilles et par voie de conséquence des alternances de production, inacceptable pour les producteurs (Figure 15B).

Figure 14: Symptômes liés à la moniliose

Premières traces de monilia sur fleurs (1), dessèchement des fleurs dû au monilia (2 et 3) et symptômes de monilia sur fruits (4 et 5).

Le **chancre bactérien**, encore appelé dépérissement bactérien ou bactériose est une maladie redoutable causée par *Pseudomonas syringae*, dont les dégâts se manifestent par des nécroses de bourgeons floraux, des criblures sur feuilles et fruits et une exsudation de gomme sur rameaux et charpentières, pouvant engendrer la mortalité le dépérissement complet des arbres (Figure 15C). Les plaies pétiolaires, qui représentent la voie de pénétration de *P. syringae* chez le pêcher et le cerisier, n'ont pas d'impact chez l'abricotier pour lequel les bourgeons constituent la voie d'entrée des bactéries. La pathogénicité des bactéries responsables du chancre est associée à leur pouvoir glaçogène qui permet la cristallisation de l'eau dans les tissus à des températures supérieures à la normale (Prunier et al. 2005). L'étude du pathosystème Abricotier - *Pseudomonas syringae* pv *syringae* a révélé l'existence de deux mécanismes de résistance opérant conjointement en réponse à l'agent pathogène. Il s'agit d'une réponse systémique consistant à bloquer l'infection au niveau des bourgeons couplée à une réponse localisée permettant de limiter la migration de la bactérie dans les tissus (Omrani 2018). La base génétique sous-jacente est actuellement étudiée, et des travaux d'introgression de la résistance sont en cours.

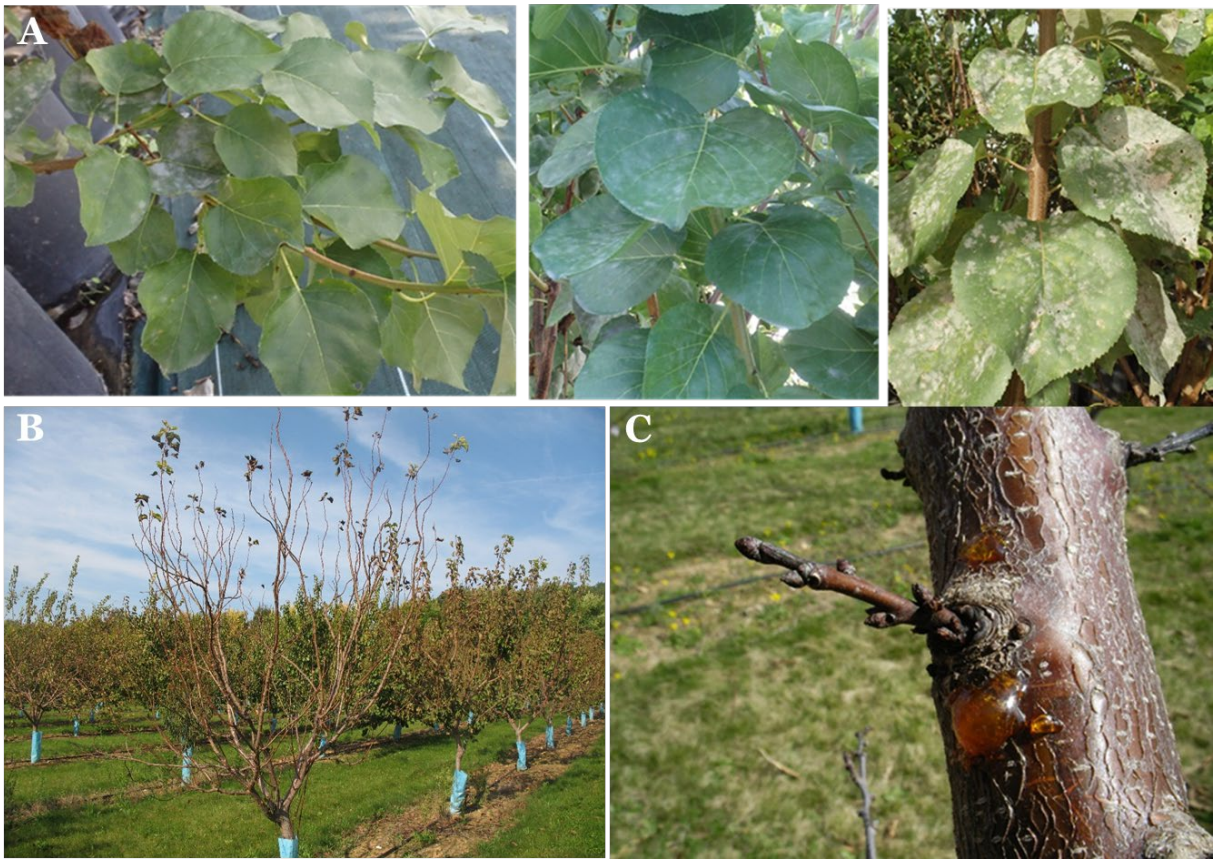


Figure 15: Illustration des symptômes associés à l'oïdium, à la rouille et au chancre bactérien

A : oïdium **(***Sphaerotheca pannosa*) sur feuilles avec une échelle de notation croissante, B : observation des dégâts de rouille (*Tranzschelia discolor*) (chute des feuilles) sur un arbre et C : observation de gomme dû au chancre bactérien (*Pseudomonas syringae*)

L'enroulement chlorotique de l'Abricotier (ECA) est la maladie la plus préoccupante à l'échelle nationale. Elle est responsable de 2 à 5% de mortalité des arbres par an en verger. L'agent pathogène responsable de l'ECA est *Candidatus phytoplasma prunorum*, un phytoplasme (bactérie sans paroi) qui se multiplie dans les vaisseaux conducteurs de la sève élaborée (Audergon et al. 1989; Lichou and Jay 2012). La transmission de l'ECA au verger s'effectue par un psylle *Cacopsylla pruni* sur un mode persistant (le vecteur assure la multiplication du phytoplasme dans son hémolymphe). La stratégie de lutte consiste à appliquer un traitement chimique ciblé contre les vecteurs infectieux au moment de leur émergence printanière. L'assainissement des vergers (prophylaxie) s'avère crucial pour l'élimination de sources d'inoculum que représente le matériel végétal infecté.

La **sharka** est une maladie règlementée causée par le Plum Pox Potyvirus (PPV). Elle affecte les espèces fruitières appartenant au genre *Prunus*. La lutte contre le PPV repose essentiellement sur des mesures de prophylaxie et de prévention qui consistent à l'identification précoce et l'éradication des arbres contaminés afin de limiter la propagation de la maladie. Il n'existe aucun traitement curatif. La maladie est transmise par greffage et par pucerons sur le mode non-persistant. Comme ce mode de dissémination est peu impacté par les traitements phytosanitaires ciblés sur les populations de pucerons, il n'existe pas de lutte préventive. L'existence de sources de résistance au sein de la diversité génétique permet aujourd'hui de développer des variétés résistantes. Le virus responsable de la sharka présente une diversité génétique très structurée avec des niveaux de prévalence clairement identifiés entre les souches M (pêcher, abricotier), D (abricotier, prunier) et REC (prunier, abricotier) présentes en France. Les symptômes observés dépendent des variétés. Ils affectent peu le rendement mais modifient l'apparence et la qualité des fruits (Lichou and Jay 2012) (Figure 16).

Figure 16: Symptômes liés à la Sharka

Intensité croissante des symptômes de sharka sur feuilles (de 1 à 3) ; Symptômes de sharka sur fruits (4 et 5)

1.4.5.3.3. Phénologie

Dans le contexte du réchauffement climatique, le suivi phénologique des stades clés de la croissance des organes végétatifs et floraux chez l'abricotier s'avère indispensable afin d'étudier les mécanismes sous-jacents à l'adaptation climatique des variétés. En effet, l'étude des réponses phénologiques aux changements climatiques et leur intégration dans les schémas de sélection permettrait d'optimiser la capacité des arbres à exprimer leur potentiel de production dans des différentes conditions pédoclimatiques. Dans ce sens, face à la complexité environnementale pouvant se manifester durant leur cycle de développement, appréhender le concept de plasticité phénotypique devrait offrir l'opportunité de prédire la performance agronomique des cultivars. Ceci pourrait garantir la durabilité du potentiel de production et la stabilité des caractères d'intérêt. Cette démarche ne peut se concrétiser qu'en se fondant sur une connaissance approfondie des stades phénologiques repères de l'espèce. L'amélioration de la plasticité phénotypique accompagnera l'amélioration du progrès génétique dans la mesure où des caractères héritables seront intégrés dans les schémas de sélection. L'objectif sera donc d'éliminer les écarts de performances dus aux interactions génotype × environnement. En effet, les changements phénologiques associés aux fluctuations climatiques se traduisent souvent par une précocité de floraison qui expose les cultivars aux risques de gelées printanières entraînant

ainsi des irrégularités de production. La relation entre la phénologie et la sensibilité aux bioagresseurs fait aussi partie des éléments qui sont pris en compte notamment dans des stratégies d'estimation de risque qui permettent de mieux cerner l'impact des effets génétiques.

1.4.5.3.4. Auto-fertilité

A l'instar des espèces fruitières appartenant à la famille des Rosacées, l'abricotier est porteur d'un système d'auto-incompatibilité gamétophytique qui empêche les fécondations entre parents porteurs des mêmes allèles au locus d'incompatibilité. En effet, le mécanisme de l'auto-incompatibilité repose sur le rejet du pollen dans la mesure où le locus multiallélique *S* s'exprime à la fois dans le pollen et dans le pistil. La réponse d'auto-incompatibilité gamétophytique est régie par l'expression d'enzymes ribonucléases (S-RNases) dans le style dégradant ainsi l'acide ribonucléique (ARN) du tube pollinique (Vilanova et al. 2006). Seules les variétés cultivées du groupe européen continental sont auto-fertiles (Kostina 1969). Le locus *S* conférant l'auto-incompatibilité est caractérisée par son hétérozygotie entravant le cumul de caractères d'intérêt chez les géniteurs possédant les mêmes allèles S.

1.4.6. Ressources génétiques et génomiques

Le génome de l'abricotier ($294\,Mb/n$) est réparti sur 8 paires de chromosomes homologues ($2n = 16$). Des avancées scientifiques significatives ont révolutionné l'étude de la génétique des espèces appartenant au genre *Prunus* depuis la mise à disposition d'une première version de la séquence du génome du pêcher en 2010, qui constitue le génome de référence de *Prunus* (Verde et al. 2013). La première version correspond au génotype « Lovell », variété homozygote de *Prunus persica* avec un petit génome ($265\,Mb/n$). Elle a été obtenue par séquençage bas débit avec la technique Sanger. En s'appuyant sur l'évolution des technologies de biologie moléculaire, le séquençage du génome complet (Whole Genome Sequencing WGS) du pêcher a donné lieu à une deuxième version qui a été publiée en 2015 (Verde et al. 2017). Le décryptage du génome de référence des *Prunus* a permis d'éclairer la compréhension de l'histoire évolutive de ces espèces et la structure de leur diversité génétique. C'est ainsi que l'analyse comparée des génomes à l'aide de marqueurs moléculaires a permis de mettre en évidence de fortes relations de synténie et colinéarité (conservation de l'ordre des gènes) entre les espèces du genre *Prunus* (Dirlewanger et al. 2004). Par ailleurs, le développement de puces à ADN a fourni des outils d'intérêt majeurs pour l'étude de la diversité génétique et du polymorphisme des traits d'intérêt chez le pommier (Chagné et al. 2012) et le pêcher (Verde et al. 2012). Outre le génome de référence du pêcher, les espèces appartenant au

genre *Prunus* disposent aujourd'hui des séquences génomiques du cerisier *Prunus avium* (Shirasawa et al. 2017) et de l'abricotier du Japon *Prunus mume* (Zhang et al. 2012) et de *Prunus armeniaca* (Jiang et al. 2019a).

Chez l'abricotier, les études de liaison et d'association génétique, l'inférence de la structure des populations mobilisées en sélection ainsi que l'analyse de l'étendue du DL à l'échelle du génome ont été dans un premier temps, adossés à la séquence du génome du pêcher. L'existence du génome de *Prunus mume*, puis celle d'un abricotier chinois et enfin celle de la variété Marouch (développé au sein de l'UMR de Biologie du Fruit et Pathologie (BFP) à l'INRAE de Bordeaux) (Groppi et al. 2021) dotent l'espèce abricotier des outils de génomique indispensables à la mise en œuvre des approches d'amélioration moderne.

Les ressources génomiques de diverses espèces de la famille de Rosaceae sont accessibles sur la plateforme GDR (Genome Database for Rosaceae, http://www.rosaceae.org/).

1.4.7. Etude du déséquilibre de liaison (DL)

L'analyse de la structure du DL à l'échelle génomique chez le genre *Prunus* a révélé une décroissance rapide du DL à courte distance ($< 100\,pb$), de même chez *P. mume* ($r <$ 0.2 à 50 Kb) et le cerisier ($r < 0.2$ à 100 Kb). En corollaire, une forte densité de marqueurs moléculaires est indispensable pour la détection de polymorphismes causaux pour ces espèces. En revanche, chez le pêcher une densité moindre de marqueurs est suffisante en raison d'une étendue de DL plus importante ($r < 0.8$ à 1.4 Mb). Toutefois, les marqueurs identifiés chez les espèces de *Prunus*, hormis le pêcher, sont potentiellement étroitement liés aux polymorphismes causaux ouvrant ainsi la voie à la cartographie fine des régions génomiques d'intérêt (Aranzana et al. 2019b). Particulièrement chez l'abricotier, l'allogamie préférentielle se traduit par une augmentation considérable d'évènements de recombinaison entraînant une décroissance rapide du DL à courte distance. Dans ce sens, Omrani et al. (2019) ont démontré que le DL intra-chromosomique dans une core-collection formée de 73 accessions d'abricotiers décroit rapidement à une distance physique allant de 100 à 200 pb en fonction des segments chromosomiques.

1.4.8. Structure de la diversité génétique

L'analyse de la structure de la diversité génétique est une étape cruciale dans le processus de sélection variétale. Elle permet de retracer l'histoire évolutive de l'espèce et d'identifier l'étendue de la variation génétique afin de l'intégrer dans les programmes d'amélioration

génétique. Dans ce contexte, l'abricot constitue une espèce modèle intéressante pour l'étude de l'évolution des espèces fruitières en raison des populations naturelles présentant une large gamme de diversité phénotypique et génétique (Liu et al. 2019a).

La caractérisation de la structure génétique du germplasm cultivé de l'espèce *Prunus armeniaca* a été conduite sur une collection de 890 accessions représentatives de la diversité mondiale à l'aide de 25 marqueurs microsatellites ; elle a permis d'identifier cinq groupes phylogénétiques en fonction de l'origine géographique des accessions (Bourguiba et al. 2020) (Figure 17).

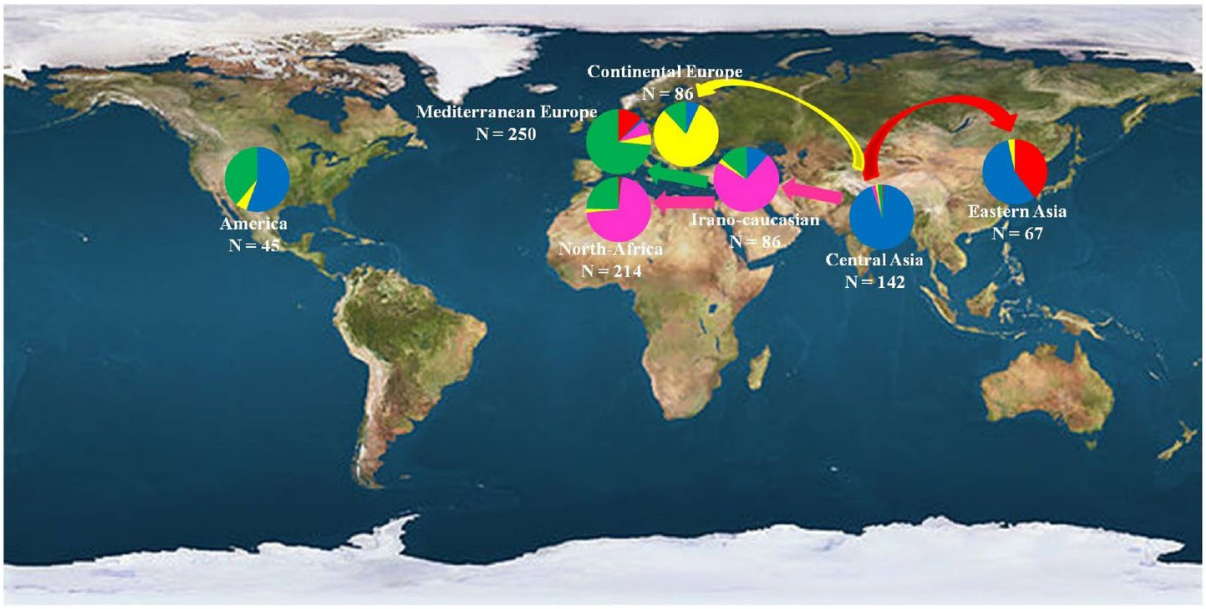


Figure 17: Distribution géographique des 890 accessions d'abricots classées suivant leur origine géographique (Bourguiba et al. 2020)

L'étude a porté sur 7 groupes géographiques répartis de l'Est à l'Ouest comme suit : Groupe 1 : "Asie de l'Est", Groupe 2 : "Asie centrale", Groupe 3 : "Irano-Caucasien", Groupe 4 : "Europe continentale", Groupe 5 : "Europe méditerranéenne", Groupe 6 : "Afrique du Nord" et Groupe 7 : "Amérique".

L'analyse de la diversité de ce panel a révélé l'existence d'un phylogroupe renfermant des accessions en provenance de Chine et d'Asie centrale, caractérisées par un niveau élevé de diversité génétique et de richesse allélique, ayant permis de confirmer l'hypothèse sur le centre d'origine géographique de l'abricot (Faust et al. 1998). Ce phylogroupe constitue un réservoir de gènes potentiellement intéressants pour la création variétale (Bourguiba et al. 2020). Le deuxième phylogroupe est représenté par la région Irano-caucasienne qui constitue un centre secondaire de diversification de l'abricotier à partir de son centre d'origine. Par ailleurs, les accessions appartenant au phylum de l'Europe continentale présentent le plus faible niveau de variabilité phénotypique et génétique. Elles ont été utilisées dans des programmes d'hybridation

et ont donné naissance à de cultivars importants (Romero et al. 2003; Pedryc et al. 2009). Le quatrième phylogroupe comprend des accessions de l'Asie orientale et représente le groupe le plus éloigné. Le cinquième phylogroupe représente la méditerranée avec l'Europe méditerranéenne et l'Afrique du Nord comportant des accessions dotées d'une base génétique étroite mais fortement ancrée sur le centre de diversification secondaire Irano-caucasien.

Au regard de la diversité mondiale, trois voies de diffusion correspondant à trois évènements potentiels de domestication ont été empruntées par l'espèce lors de sa domestication à partir du centre secondaire de diversification : Irano-Caucasien, Nord du bassin méditerranéen et Sud du bassin méditerranéen. Ceci est illustré par l'existence de flux de de gènes entre les 3 pools génétiques (Bourguiba et al. 2012). Une perte de diversité a ainsi été observée du centre d'origine jusqu'aux zones de culture. Cependant, aucune signature génétique n'a été identifiée pour le goulot d'étranglement se traduisant par l'absence d'excès d'hétérozygotie (Bourguiba et al. 2020).

En définitive, la valorisation de la connaissance de l'information sur la structure génétique est indispensable lors de la conception d'une population de référence dans le cadre de la sélection et la création variétale ainsi que pour l'identification des régions génomiques intéressantes par le biais d'études d'association à l'échelle du génome.

Objectifs de la thèse

Mes travaux de thèse s'inscrivent dans le cadre de l'évaluation de stratégies de sélection fondées sur l'information haut-débit issue de marqueurs moléculaires (sélection génomique) et de spectroscopie infrarouge (sélection phénomique). Mes objectifs s'articulent autour de :

- L'étude du déterminisme génétique des caractères cibles par le biais d'analyses de liaison et d'association génétiques dans la population biparentale Go×Mo et un panel de diversité incluant 93 accessions, respectivement.
- L'évaluation de la capacité prédictive des modèles de sélection génomique et phénomique pour des caractères d'intérêt liés à la qualité des fruits, à la phénologie de l'arbre et à la résistance à quelques maladies.
- L'optimisation de la capacité prédictive des modèles génomiques et phénomiques en valorisant l'information apportée par leurs déterminismes génétiques.

Chapitre 2 :
Matériel & méthodes

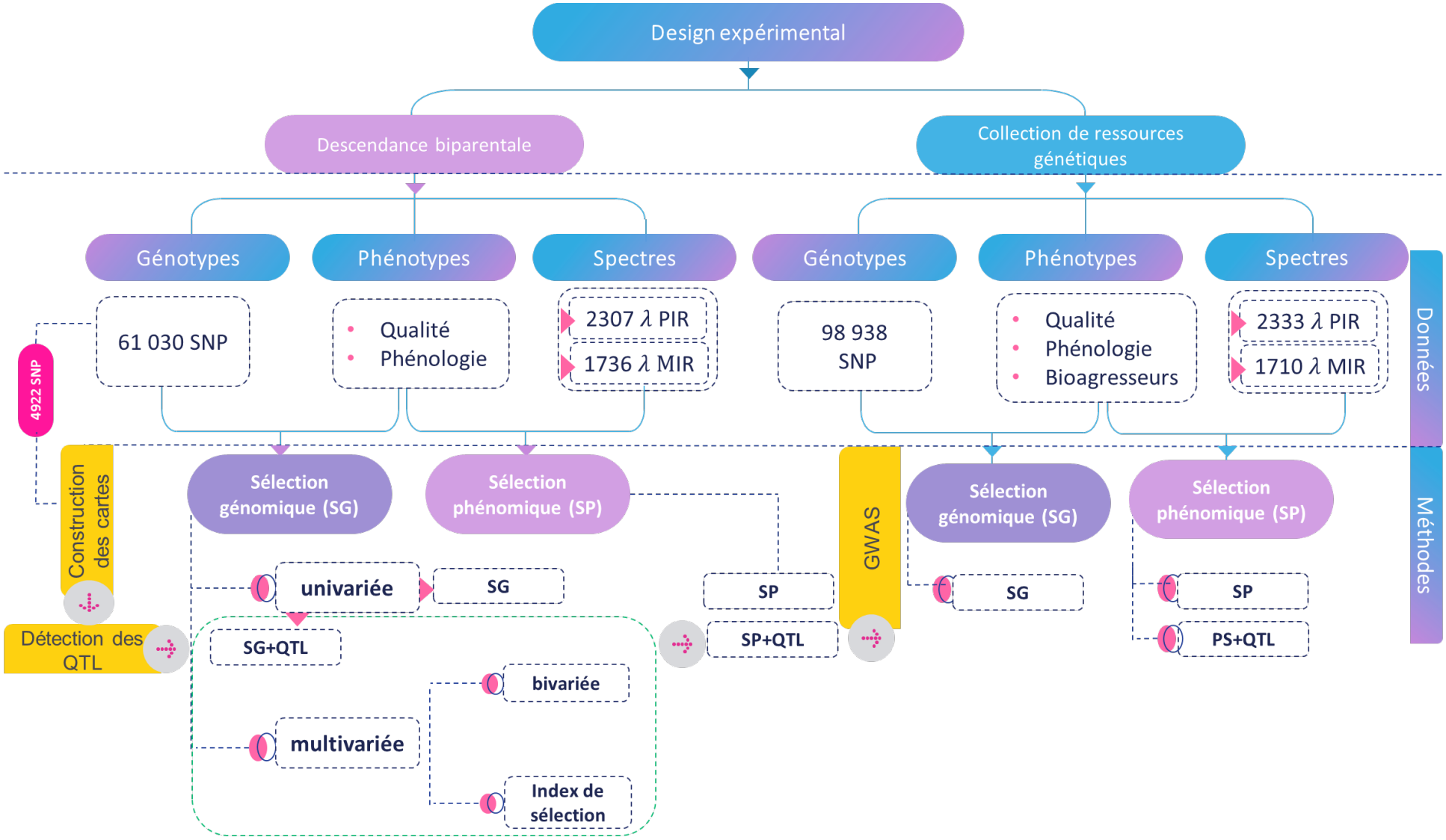


Figure 18: Synthèse du matériel végétal, données phénotypiques, génotypiques et spectrales et méthodes mobilisées dans le cadre de la thèse

Chapitre 2 : Matériel et méthodes

Dans cette section, nous décrivons le matériel végétal utilisé, ainsi que les méthodes employées afin de répondre aux objectifs de la thèse (Figure 18).

2.1. Matériel végétal

Le dispositif expérimental choisi pour répondre aux questions scientifiques sur lesquelles s'appuie cette thèse comprend une population biparentale (Goldrich x Moniqui ou Go×Mo) de 153 individus installée sur le domaine expérimental INRAE de l'Amarine, situé à Bellegarde dans le département du Gard (30), ainsi qu'une collection de ressources génétiques composée de 93 accessions, implantée sur le domaine de Gotheron relevant de l'unité expérimentale de recherches intégrées (UERI INRAE) située à Saint-Marcel-lès-Valence dans de département de la Drôme (26).

2.1.1. Population biparentale en pseudo-F1

La descendance hybride étudiée est une population pseudo-F1 issue du croisement entre deux variétés présentant des performances agronomiques contrastées pour la qualité des fruits. Le parent mâle Moniqui est une variété espagnole de saison, auto-incompatible, caractérisée par des fruits blancs, de gros calibre, présentant une fermeté faible. Les fruits de cette variété présentent une excellente qualité gustative au regard de leur texture juteuse et de leur saveur sucrée, avec également une production éthylénique élevée se traduisant par une évolution rapide de l'activité métabolique à maturité et après la récolte. Quant au parent femelle Goldrich, il s'agit d'une variété de saison originaire d'Amérique du Nord, auto-incompatible, et caractérisée par des fruits fermes, de couleur orange et à faible production éthylénique. Le design expérimental conçu pour cette étude repose sur la distribution aléatoire des hybrides plantés sur leurs propres racines (non greffés), avec des blocs dédiés aux parents disposés sur la diagonale du plan expérimental installé à l'Amarine (2005).

2.1.1.1. Phénotypage pour la qualité des fruits

L'évaluation de la qualité des fruits de la descendance Go×Mo a porté sur deux années consécutives, 2006 et 2007, antérieurement à la présente étude. Des mesures physiques portant sur le poids et la couleur des fruits ont été effectuées conjointement à des mesures biochimiques afin de caractériser la composition des fruits en sucres et acides organiques. Des échantillons de 40 fruits par génotype ont été récoltés à proximité de leur maturité physiologique. Les fruits ont été triés en fonction de leur fermeté, déterminée par la pression (kPa) requise pour déformer

le fruit de 3% de sa hauteur avec un analyseur de texture polyvalent (Pénélaup, Serisud, Montpellier, France). Les fruits ont été subdivisés en trois lots homogènes de quatre fruits par génotype, correspondant à des stades de maturité contrastés : un stade de maturité commerciale avec une pression de 130 à 80 kPa, un stade de demi-maturité de 80 à 50 kPa et un stade de fruits mûrs avec une fermeté inférieure à 50 kPa. Les caractéristiques physiques, physiologiques et biochimiques ont été mesurées sur ces trois lots représentatifs pour l'ensemble des génotypes. La couleur de fond de la face non exposée au soleil et la couleur de surimpression (face exposée au soleil) ont été mesurées à l'aide d'un chromamètre CR-400 (Minolta, Osaka, Japon) et exprimées dans l'espace chromatique $L^*a^*b^*$ CIE (1976). L'angle de teinte, a été calculé en utilisant les coordonnées de chromaticité a et b comme suit :

$$Hue = tg^{-1}({}^{b^*}/_{a^*})$$

La production éthylénique a été appréciée par la mesure du dégagement d'éthylène par chromatographie en phase gazeuse suite au confinement individuel des fruits pendant 1 heure dans des pots hermétiquement fermés. Le protocole expérimental qui a été mis au point pour quantifier le taux de production d'éthylène, exprimé en $nmolkg^{-1}h^{-1}$ est décrit dans Chambroy et al. (1995). Concernant les analyses biochimiques, les teneurs en sucres (glucose, fructose et saccharose) et en acides organiques (acide citrique et acide malique) ont été mesurées par dosages enzymatiques et exprimées en g 100 g^{-1} de matière fraiche pour les sucres et en $milliéquivalents$ 100 g^{-1} pour les acides.

La mesure de l'indice de réfraction, qui est corrélée à la teneur en matière sèche soluble, a été effectuée à l'aide d'un réfractomètre digital et exprimée en % $Brix$ à une température égale à $20^{\circ}C$. L'acidité titrable a été déterminée sur une fraction aliquote de mésocarpe broyé, neutralisée par une solution de soude de concentration égale à 0.1 N jusqu'à atteindre un pH égal à 8.1 et exprimée en milliéquivalents par gramme de matière fraîche (meq 100 g^{-1}).

2.1.1.2. Phénotypage pour la phénologie

La phénologie a été évaluée pendant trois années (2006, 2007 et 2009) en référence à deux caractères phénologiques clé : la date de floraison et la date de maturité, exprimées en jours juliens. La date de floraison a été déterminée lorsque 50% des boutons floraux avaient atteint le stade de la floraison (BBCH60, 65, 69). Quant à la date de maturité (BBCH87), elle a été notée à l'approche de la maturité physiologique des fruits (Dirlewanger et al. 2012).

2.1.1.3. Génotypage

La descendance Go×Mo a fait l'objet d'un génotypage par séquençage (GBS), effectué suivant le protocole décrit par (Elshire et al. 2011) dans le cadre du projet de recherche FruitSelGen (2015). Le génotypage utilisant la technologie NGS Illumina HiSeq2000 a été réalisé au sein de la plateforme génomique GeT-PlaGe (INRAE Occitanie-Toulouse) et la préparation des banques a été effectuée au sein de la plateforme de génotypage hébergée à l'unité de recherche 'Amélioration génétique et adaptation des plantes méditerranéenne et tropicales' (AGAP ; Cirad, INRAE, Institut Agro). L'extraction de l'ADN a été effectuée à partir des feuilles à l'aide du kit de réactifs Qiagen. La digestion des fragments d'ADN et l'ajout des adaptateurs encadrés par des codes-barres aux extrémités ont été réalisés par le biais de l'endonucléase de restriction ApeKI. Les banques d'ADN générées ont été amplifiées par PCR et séquencées. Les signaux fluorescents ont été convertis en séquences nucléotidiques de 100 paires de bases qui ont été, par la suite, stockées dans des fichiers sous format Fastq. La séquence complète du génome de l'abricotier n'étant pas disponible au début de la thèse et afin de valoriser la synténie entre l'abricotier et le pêcher, la détection des variants dans les données de séquençage a été effectuée en mobilisant la première version du génome du pêcher utilisé comme espèce de référence du genre *Prunus* (Verde et al. 2013). Les lectures ont été alignées sur la séquence de référence par le biais du logiciel d'alignement Burrows-Wheeler (Burrows-Wheeler Aligner BWA) (Li and Durbin 2009). A l'issue de la détection des variants, un ensemble de 2 565 573 SNP a été obtenu, présentant une profondeur de couverture moyenne égale à 19,33×. Les lectures ont été filtrées via les algorithmes des logiciels GATK, bcftools et le package VariantAnnotation du logiciel R (Obenchain et al. 2014a). L'analyse du score de qualité des séquences (genotype quality GQ) a donné lieu à un panel renfermant 61 030 SNP avec un score de GQ>20. Les marqueurs moléculaires ont été codés en dosage allélique 0, 1 et 2 en fonction du nombre de copies de l'allèle alternatif.

2.1.1.4. Acquisitions spectrales

Les données spectroscopiques ont été acquises dans le domaine spectral du proche infrarouge (PIR) sur des fruits intacts et dans le moyen infrarouge (MIR) sur des homogénats de fruits à raison de 4 fruits par génotype pour chacun des trois lots de maturité. Dans le PIR, la réflectance diffuse correspondant à chaque spectre a été enregistrée sur une moyenne de 32 scans et réalisée à l'aide d'un spectromètre analyseur polyvalent (Bruker Optics®, Wissembourg, France) (Bureau et al. 2009c). Dans le MIR, des spectres de réflectance totale atténuée (ATR) ont été collectés à température ambiante à l'aide d'un spectromètre à transformée de Fourier équipé

d'un cristal de séléniure de zinc ATR et d'un détecteur de sulfate de triglycine deutéré. Les homogénats de broyats de fruits (un échantillon par génotype pour chaque lot de maturité) ont été placés sur le cristal pour l'acquisition des spectres. Les spectres PIR ont été acquis pour des nombres d'onde allant de 12 500 à 4 000, tandis que les spectres MIR ont été enregistrés pour une plage de nombres d'onde variant de 4 000 à 700 cm^{-1}.

2.1.2. Collection de ressources génétiques

La collection de ressources génétiques est constituée de 93 accessions d'abricots cultivées dans un système de culture sous très faibles intrants phytosanitaires avec une protection ciblant uniquement les insectes vecteurs de l'enroulement chlorotique (ECA). Les accessions ont été sélectionnées parmi les collections de l'INRAE sur la base des résultats d'études de la diversité génétique effectués par Bourguiba et al. (2012) et Bourguiba et al. (2020) en utilisant des marqueurs SSR.

Les cultivars sont greffés sur des porte-greffes de pêcher Montclar ® Chanturgue. Le panel d'abricotiers a été planté en verger en 2018 sur le site expérimental de l'INRAE situé à Gotheron (France) dans un dispositif randomisé comptant 5 blocs avec une répétition par bloc pour chaque accession (dispositif en Bloc Randomisé mono-arbre). Le panel d'étude fait partie d'un dispositif expérimental multisites situé dans différentes zones de production d'abricots en France (Gotheron, L'Amarine et Torreilles) et en Suisse (Conthey), dédié à l'évaluation de la sensibilité des collections à un cortège d'agents pathogènes dans le cadre de conditions pédoclimatiques contrastées.

2.1.2.1. Phénotypage pour la qualité des fruits

L'évaluation de la qualité des fruits du panel de diversité a porté sur deux années de mesure (2019 et 2020) incluant le poids individuel du fruit, l'indice de réfraction et l'acidité titrable ont été déterminés en suivant les protocoles expérimentaux mis au point pour la population biparentale Go×Mo.

2.1.2.2. Phénotypage pour la sensibilité aux maladies

La caractérisation de la sensibilité des ressources génétiques aux maladies bactériennes et cryptogamiques s'inscrit dans une optique de limitation du recours aux produits phytosanitaires afin de répondre à des enjeux sanitaires et environnementaux. Dans cette perspective, la collection a été phénotypée pour la sensibilité à des maladies pouvant être préjudiciables pour la production et la pérennité du verger. Parmi ces maladies figurent l'oïdium (*Podosphaera sp.*),

la moniliose sur fleurs (*Monilinia laxa*), la rouille (*Tranzchelia pruni-spinosae*) et la bactériose (*Pseudomonas syringae pv. syringae*). L'évaluation de la sensibilité des arbres aux maladies a porté sur la seule année 2020.

Sensibilité à l'oïdium

Le phénotypage pour la sensibilité à l'oïdium, causé par le champignon pathogène *Sphaerotheca pannosa*, a été effectué sur 50 fruits par arbre prélevés à l'approche de la maturité des fruits. Il a porté sur la notation du pourcentage de fruits oïdiés.

Sensibilité à la moniliose sur fleurs

Il s'agit d'une maladie cryptogamique causée par divers champignons pathogènes dont les plus redoutables en termes de dégâts occasionnés sont : *Monilia laxa*, *Monilia fructicola* et *Monilia fructigena*. La caractérisation de la sensibilité à la moniliose a été effectuée 30 jours après la floraison afin que les symptômes sur rameaux s'expriment et le développement de la végétation ne gêne pas ni l'observation ni la notation. Elle a porté sur la notation du pourcentage de rameaux présentant des fleurs desséchées suite à l'attaque du champignon par rapport à l'ensemble des rameaux à fleurs de l'arbre.

Sensibilité à la rouille

L'évaluation de la sensibilité à la rouille (*Tranzschelia discolor*) s'est basée sur l'estimation visuelle de la sévérité des symptômes en termes de pourcentage de feuilles présentant des pustules selon une échelle rapportée dans le tableau 1. L'échelle de notation de l'intensité de la rouille varie de 0 pour les arbres dont les feuilles sont exemptes de la maladie jusqu'à 5 pour 80% de feuilles manifestant les symptômes de la rouille. Quant aux symptômes associés à la chute de feuilles, une échelle de notation de 0 correspondant à 0% de feuilles chutées à 5 correspondant à plus de 80% de feuilles chutées a été mobilisée.

Tableau 1: Echelle de notation de la rouille

Echelle de notation	Pourcentage de feuilles présentant des pustules	Pourcentage de feuilles chutées
0	Aucun symptôme de rouille observé	Aucune chute
1	≤ à 10%	≤ à 10%
2	>10% et ≤30%	>10% et ≤30%
3	>30% et ≤60%	>30% et ≤60%
4	>60% et ≤80%	>60% et ≤80%
5	>80%	>80%

Sensibilité à la bactériose

Le phénotypage pour la sensibilité au chancre bactérien causé par *Pseudomonas syringae* a été réalisé par observation des points de gomme par charpentière en conditions d'inoculation naturelle n'ayant pas été favorables à l'infestation et à l'expression des symptômes. Les symptômes ont été quantifiés vers la fin de la floraison avant que les desséchements de rameaux liés aux attaques de monilioses à la fleur n'apparaissent. Le pourcentage de charpentières présentant des symptômes (bourgeons ou coursonnes nécrosés avec ou sans présence de gomme et avec écorce rouge-brun en surface et tissus marrons-nécrosés dessous et jeunes branches ne débourrant pas avec tissus marrons-nécrosés sous l'écorce et/ou présence de gomme) a été estimé suivant une échelle de 0, 25, 50, 75, 100%.

2.1.2.3. Phénotypage pour la phénologie

A l'instar des notations réalisées sur la descendance Go×Mo, le phénotypage pour la phénologie sur le panel de diversité a porté sur la date de floraison ayant eu lieu lorsque 50% des boutons floraux avaient atteint le stade de la floraison (BBCH60, 65, 69) et la date de maturité (BBCH87) à l'approche de la maturité physiologique appréciée par le ramollissement des fruits dû à la perte de fermeté associée à la perte d'adhésion des composants de la paroi cellulaire.

2.1.2.4. Génotypage

Le panel de l'étude a été génotypé à l'aide de la technologie Illumina HiSeq2000 NGS par la plateforme génomique GeT-PlaGe (INRAE Toulouse) suivant le protocole d'Elshire et al. (2011). La préparation des bibliothèques a été réalisée à l'unité de recherche AGAP (INRAE Montpellier). Les lectures brutes ont été filtrées en fonction de leur score de qualité (QS>20). Les séquences d'ADN ont été alignées sur le génome d'un cultivar d'abricot marocain (Marouch) récemment publié (Groppi et al. 2021) via le logiciel Burrows-Wheeler (Burrows-Wheeler Aligner BWA). Les séquences dupliquées ont été supprimées à l'aide de Samtools (samtools v1.9, sambamba v0.7.1) et la qualité de l'alignement a été vérifiée en utilisant multiQCs. Les SNPs et les indels ont été appelés avec GATK (gatk4 v4.1.4.1) et filtrés. Un total de 98 938 SNPs bialéliques avec un seuil MAF de 0.05 et un maximum de 20 % de données manquantes a ensuite été sélectionné pour couvrir tout le génome à des intervalles homogènes fournissant une densité de couverture d'environ un marqueur tous les 1958 pb.

2.1.2.5. Acquisition des spectres

L'acquisition spectrale a été effectuée dans deux domaines spectraux de l'infrarouge. Les spectres ont été obtenus en mode réflectance diffuse couvrant une gamme de variation du nombre d'ondes de 12 493 à 3 498 cm^{-1} dans le PIR et comprise entre 3 996 et 700 cm^{-1} dans le MIR. Les spectres PIR ont été acquis en 2020 à partir d'échantillons de feuilles collectées aléatoirement sur les rameaux longs à raison de 5 répétitions par génotype, la réflectance étant mesurée sur la face supérieure des feuilles. Quant aux spectres MIR, ils ont été obtenus à partir d'homogénats de fruits broyés, comme dans le cas de la descendance Go×Mo. La caractérisation spectrale a été réalisée sur les deux années 2019 et 2020 pour le MIR.

2.2. Méthodes statistiques

La modélisation statistique des données phénotypiques, moléculaires et spectrales a été effectuée à l'aide du logiciel R (R Core Team 2018).

2.2.1. Modélisation statistique des données phénotypiques

Le partitionnement de la variance phénotypique ainsi que la détermination des facteurs significatifs a été effectuée par le biais de l'analyse de variance (ANOVA). Les données phénotypiques ont été ajustées à l'aide de la fonction 'lmer' implémentée dans le package 'lme4' (Bates et al. 2014b). Dans le cadre du modèle mixte, la variation phénotypique s'exprime comme étant :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + \alpha\beta_{ij} + e_{ijk} \qquad 2.1$$

où y_{ijk} représente le phénotype, μ est la moyenne, α_i est l'effet aléatoire du génotype i, β_j est l'effet de l'année j, δ_k est l'effet du stade de maturité k, $\alpha\beta_{ij}$ est l'effet de l'interaction entre le génotype et l'année et e_{ijk} est l'erreur.

2.2.2. Construction des cartes génétiques

L'élaboration des cartes génétiques a été fondée sur une stratégie de double-pseudo testcross étant donné que les parents sont fortement hétérozygotes en raison de leur régime de reproduction préférentiellement allogame (Grattapaglia and Sederoff 1994). Préalablement à la construction des cartes, les marqueurs qui s'écartent de la ségrégation mendélienne et de l'équilibre de H-W ont été éliminés à l'aide du package 'rutilstimflutre' (Flutre 2019). Deux cartes génétiques parentales ont été obtenues en utilisant le package 'ASMap' (Taylor and Butler 2017). L'algorithme mis au point afin de déterminer l'ordonnancement optimal des

marqueurs est celui de l'arbre couvrant de poids minimal (ACM) où les marqueurs représentent les sommets et les arêtes représentent la distance de Hamming (Cheriton and Tarjan 1976).

Soit P_{ij} la probabilité qu'un évènement de recombinaison se produise entre deux marqueurs m_i et m_j avec $0 \leq P_{ij} \leq 0.5$. La probabilité que deux marqueurs appartiennent au même groupe de liaison repose sur la résolution de l'équation :

$$P(d_{ij} < \delta) \leq exp(-\frac{2\left(\frac{n}{2} - \delta\right)^2}{n}) \qquad 2.2$$

où d_{ij} est la distance de Hamming entre les deux marqueurs m_i et m_j, n est le nombre de génotypes et $\hat{\delta}$ est le seuil associé à la distance de Hamming avec $\delta < n/2$. Si $P_{ij} = 0.5$ alors m_i et m_j appartiennent à deux groupes de liaison différents. Le maximum de vraisemblance de la fonction P_{ij} est représenté par la distance de Hamming divisée par le nombre de génotypes d_{ij}/n. L'ordre des marqueurs a été estimé à l'aide de la fonction de distance génétique de Kosambi avec un LOD seuil égal à 3.

La représentation graphique des deux cartes génétiques a été réalisée à l'aide du logiciel MapChart 2.3 (Voorrips 2002b) (annexe I).

2.2.3. Analyse de l'architecture génétique des caractères cibles

Afin d'étudier le déterminisme génétique des caractères cibles, nous avons eu recours à la cartographie d'intervalle composite (CIM) en utilisant le package R 'qtl' dans le but d'identifier les régions génomiques qui sous-tendent les caractères étudiés (Broman et al. 2003b). La détection de QTL a été effectuée en mobilisant les deux cartes génétiques parentales. La significativité du LOD sore a été déterminée par des tests de permutations (1 000 permutations) avec un seuil égal à 0.01. Le pourcentage de variation expliquée par chaque QTL identifié a été déterminé par analyse de variance.

La détection de QTLs a été effectuée en interannuel en utilisant les valeurs phénotypiques ajustées et en annuel sur deux ensembles de données indépendants enregistrés en 2006 et 2007 dans le but d'évaluer la stabilité des QTL.

2.2.4. Prédiction génomique univariée

Pour l'évaluation de la performance des modèles de sélection génomique en termes de précision de prédiction, nous avons utilisé 6 modèles de régression linéaire paramétrique. Ils sont fondés sur des hypothèses qui divergent par rapport aux lois régissant la distribution des effets attribués

aux marqueurs. Il s'agit des modèles RR-BLUP, Bayes A, Bayes B, Bayes C, Lasso Bayésien (BL) et la régression Ridge bayésienne (BRR). L'objectif étant de prédire les phénotypes, supposés être inconnus, dans la partition de validation.

Le modèle de référence lie la variation phénotypique à la variation issue de la contribution de tous les marqueurs supposés être en déséquilibre de liaison avec les variants causaux des caractères cibles. L'équation du modèle est fournie dans :

$$y = X\beta + Zu + e \qquad 2.3$$

où y est le vecteur des phénotypes, X est une matrice d'incidence des effets fixes reliant les phénotypes au vecteur des effets fixes β, β est le vecteur des effets fixes, Z est une matrice d'incidence des effets aléatoires reliant les phénotypes au vecteur des effets génétiques additifs aléatoires u et e désigne le vecteur des erreurs aléatoires.

Sous l'hypothèse que les paramètres aléatoires suivent une loi normale d'espérance nulle avec $u \sim N(0, I\sigma_u^2)$ et $e \sim N(0, I\sigma_e^2)$, la valeur génomique estimée correspond à la somme des effets associés aux marqueurs multipliée par leurs doses alléliques respectives, comme illustré dans l'équation :

$$GEBV_i = \sum_{j=1}^{n} Z'_{ij}\, \hat{u}_j \qquad 2.4$$

où Z'_{ij} représente la dose allélique du $i^{ème}$ individu au $j^{ème}$ locus marqueur et $\hat{u}_j$ représente l'effet estimé du $j^{ème}$ locus.

L'évaluation de la performance des modèles de prédiction a été effectuée par le biais d'une stratégie de validation croisée reposant sur le partage de la population d'étude en une partition d'entraînement renfermant 75% de la population totale et une partition de validation renfermant 25% des individus dont les données phénotypiques sont supposées être inconnues. L'échantillonnage aléatoire dans la population d'étude en vue de la validation croisée a été itéré 100 fois et la précision de la prédiction a été estimée par le biais du coefficient de corrélation de Pearson entre les phénotypes observés et les phénotypes prédits.

2.2.5. Optimisation des modèles de sélection génomique

2.2.5.1. Pondération des modèles de sélection génomique

Afin d'optimiser la précision de la prédiction des modèles de sélection génomique, nous avons évalué l'intégration de l'information apportée par l'architecture génétique dans les modèles. Cette stratégie d'optimisation se base sur la ségrégation des marqueurs en fonction de leur

contribution à la variation phénotypique. Elle consiste à séparer le panel des marqueurs moléculaires en deux sous-ensembles où les SNP significativement liés aux caractères cibles sont traités comme étant des covariables et pondérés par leurs contributions respectives à la variation phénotypique. Ce set de marqueurs qui sous-tendent l'architecture génétique des caractères évalués est soustrait de la matrice des marqueurs présentant des effets ne dépassant pas le seuil de significativité statistique qui, quant à eux, sont traités comme des paramètres aléatoires.

La pondération des modèles de sélection génomique consiste à accorder un poids plus important aux marqueurs étroitement liés aux caractères. En corollaire, la procédure se déroule en deux étapes. La première étape consiste à identifier les QTL par analyse de liaison génétique dans la population d'entraînement définie pour chaque itération, et la deuxième repose sur l'inclusion des QTLs dans le modèle de prédiction en tant que covariables.

Le modèle de prédiction génomique qui tient compte de l'information a priori sur l'architecture du caractère est défini comme suit :

$$y = X'\beta' + Z'u' + e \qquad 2.5$$

où y est le vecteur des observations phénotypiques, X' est la matrice d'incidence des effets fixes reliant les phénotypes aux doses alléliques des QTL, β' est le vecteur des doses alléliques des QTL, Z' est la matrice d'incidence des effets aléatoires reliant les phénotypes au panel de marqueurs ne dépassant pas le seuil de significativité statistique, u' est le vecteur des doses alléliques des marqueurs restants et e est le vecteur des résidus.

2.2.5.2. Prédiction génomique multivariée

Dans l'optique de l'optimisation des modèles de sélection génomique, nous avons évalué le cadre de prédiction multivariée consistant à s'appuyer sur des caractères proxies dont le phénotypage est moins coûteux et facile à mettre en œuvre pour prédire des caractères focaux. Dans un souci de simplification, nous nous sommes penchés sur la prédiction bivariée intégrant un seul caractère proxy pour prédire le caractère cible. Afin de mettre en œuvre la prédiction bivariée, nous avons utilisé le modèle GBLUP fourni dans le package R "sommer" (Covarrubias-Pazaran 2018).

Sous l'hypothèse de normalité multivariée des effets des aléatoires, l'équation du modèle bivarié est fournie par l'équation :

$$y_i = X_i\beta_i + Z_i u_i + e_i \qquad 2.6$$

où y_i est le vecteur des données phénotypiques pour le caractère *i* avec *i* = 1, . . ., t, X_i est la matrice d'incidence des effets fixes, β_i est le vecteur des effets fixes, Z_i est la matrice d'incidence des effets aléatoires, u_i est le vecteur des effets aléatoires et e_i désigne le vecteur des erreurs aléatoires.

Le cadre de mise en œuvre de la prédiction bivariée suit celui de (Maier *et al*. 2015) et (Covarrubias-Pazaran *et al*. 2018) :

$$y_1 = X_1\beta_1 + Z_1 u_1 + e_1 \text{ pour le caractère 1} \qquad 2.7$$
$$y_2 = X_2\beta_2 + Z_2 u_2 + e_2 \text{ pour le caractère 2}$$

Le caractère 1 représente le caractère focal à prédire en valorisant les données phénotypiques associées au caractère 2 qui représente le proxy.

2.2.6. Précision de la sélection phénomique

Afin d'évaluer la capacité prédictive des modèles de sélection phénomique, nous avons utilisé le modèle RR-BLUP qui a été calibré en utilisant les spectres PIR et MIR. Différents algorithmes de prétraitement ont été appliqués aux données spectroscopiques et l'incidence du prétraitement sur la précision de prédiction du modèle a été évaluée. Le prétraitement qui maximise la capacité prédictive moyenne et réduit l'étendue de la variabilité de la précision (Annexes II et III) est utilisé par la suite dans le cadre de l'étude comparative par rapport à la SG et de l'optimisation par les QTLs détectés par analyse de liaison génétique dans la population biparentale ou par étude d'association dans le panel de diversité.

Chapitre 3 : Adoption et optimisation de la sélection génomique dans un dispositif biparental chez l'abricotier

3.1. Présentation du chapitre

Dans ce chapitre, nous répondons à la première question de cette de thèse portant sur la performance de la sélection génomique chez l'abricotier. Dans ce contexte, notre travail représente la première évaluation de la sélection génomique chez l'abricotier en mettant l'accent sur les principaux attributs de qualité du fruit afin d'aider les décisions de sélection dans le cadre de programmes de sélection axés sur la qualité.

Nous nous sommes intéressés dans un premier temps à une descendance biparentale caractérisée par une variabilité phénotypique importante étant donné qu'elle dérive de deux parents de fonds génétiques très éloignés et présentant des phénotypes contrastés pour la qualité des fruits, comme illustré dans la figure 19. En effet, Goldrich est une variété nord-américaine à chair orangée, caractérisée par une teneur élevée en acides et faible en sucres alors que Moniqui est une variété espagnole à chair blanche et à évolution rapide vers la maturité et donc ayant une production éthylénique considérablement plus importante que celle de Goldrich. A cela s'ajoute une variabilité génétique importante due à la forte hétérozygotie de l'espèce, qui s'est traduite par des valeurs d'héritabilité au sens large variant de 0.56 pour le saccharose à 0.92 pour le glucose.

3.2. Construction des cartes génétiques

Deux cartes génétiques parentales ont été construites par le biais d'une stratégie de double pseudo-testcross, ayant été mise en œuvre pour la première fois chez les arbres forestiers et particulièrement l'eucalyptus (Grattapaglia and Sederoff, 1994) et par la suite largement adoptée chez les arbres fruitiers en raison de leur forte hétérozygotie. Etant donné que dans la descendance en pseudo F1 issue du croisement entre deux parents hétérozygotes, la ségrégation des marqueurs empêche la distinction entre les formes alléliques hétérozygotes et homozygotes, les deux cartes génétiques renfermant des marqueurs issus de la ségrégation lors de la méiose pour chacun des deux parents, ont été élaborées séparément avec des configurations de type 1:1 de telle sorte que, pour une carte donnée, les marqueurs sont hétérozygotes chez un parent et absents chez l'autre (parent homozygote nul).

Préalablement à l'élaboration des cartes, les marqueurs moléculaires, répondant à la loi de ségrégation mendélienne et présentant moins de 1% d'erreurs de génotypage et de données

manquantes, ont été répartis en deux groupes. Le premier groupe renfermant 366 SNPs hétérozygotes pour Goldrich et le deuxième 250 SNPs pour Moniqui ont servi pour l'élaboration de la carte du parent femelle et du parent mâle, respectivement.

Les SNPs cartographiés pour Goldrich ont été répartis sur huit groupes de liaison (GL), présentant une longueur totale de 562 cM et une distance moyenne de 1.6 cM entre les SNPs adjacents. Pour le parent mâle Moniqui, la carte génétique s'étendait sur une longueur totale de 842.3 cM avec un espacement moyen de 3.5 cM entre les SNPs. Les 250 SNPs cartographiés pour Moniqui ont été positionnés sur 10 GLs. Les chromosomes 1 et 7 ont été divisés en deux GLs chacun. La répartition des chromosomes sur les GLs est représentée dans la figure 20.

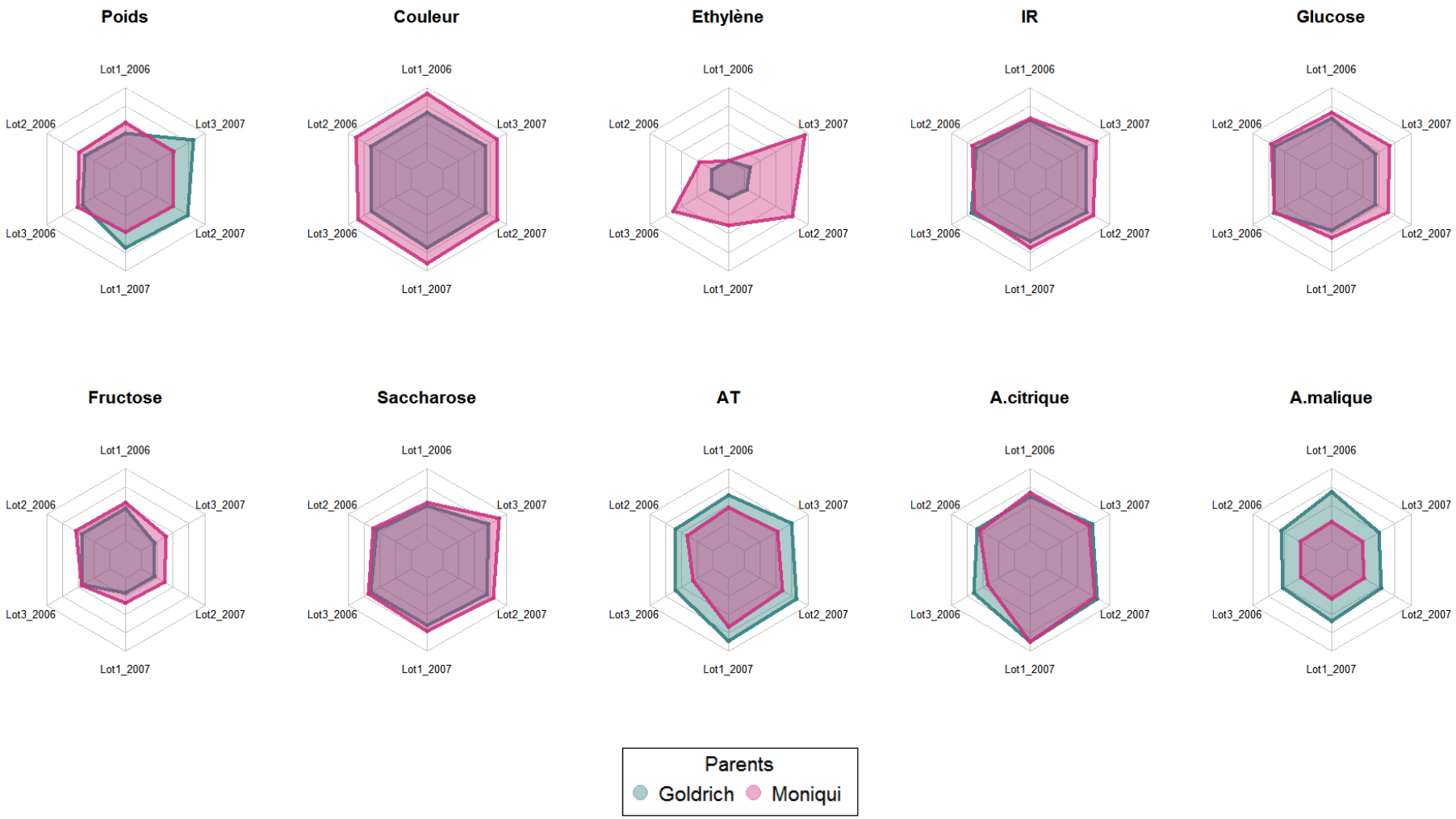


Figure 19: Représentations graphiques des données phénotypiques liées à la qualité des fruits pour les deux parents Goldrich et Moniqui

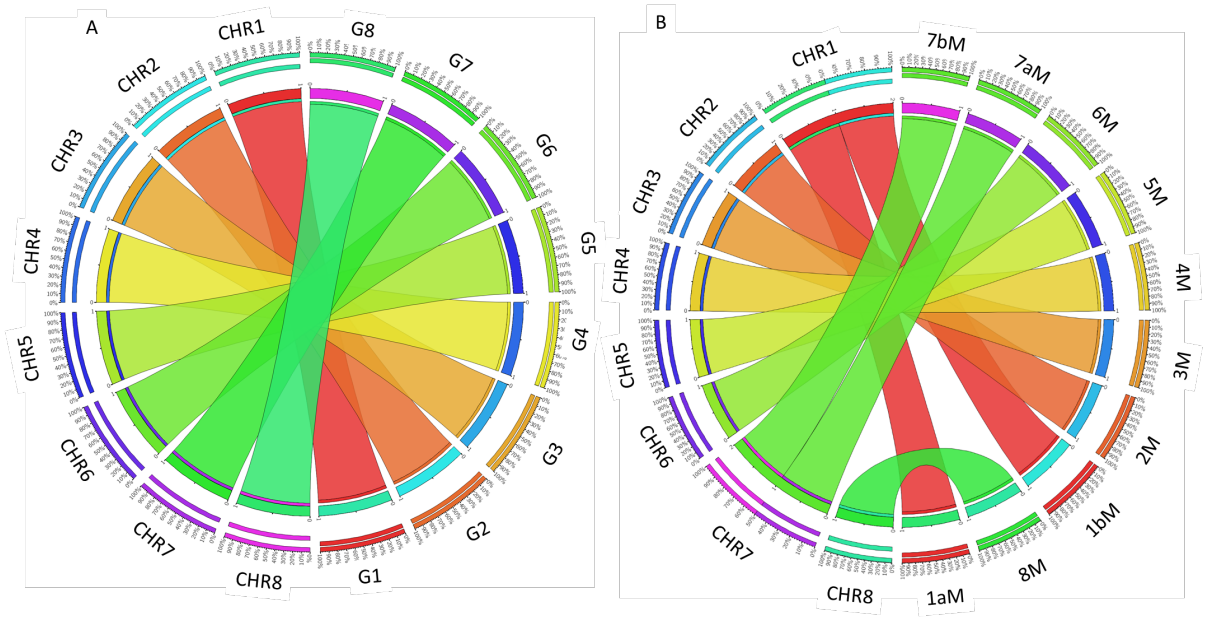


Figure 20: Correspondances entre les chromosomes et les groupes de liaisons des deux cartes génétiques parentales, du parent femelle Goldrich (A) et du parent mâle Moniqui (B)

3.3. Détection de QTLs liés à la qualité des fruits

L'architecture génétique qui sous-tend les attributs de qualité des fruits a été évaluée afin d'approfondir notre connaissance sur les régions chromosomiques abritant les QTLs d'intérêt. Les analyses de cartographie portant sur les deux années de mesure ainsi qu'en interannuel, ont été effectuées séparémmment sur les deux cartes parentales.

Dans cette optique, la cartographie par intervalle composite a révélé l'existence de 20 QTLs significativement liés aux 10 caractères de qualité expliquant une proportion de la variance phénotypique variant de 7.6% pour l'AT jusqu'à 51.2% pour la couleur du fond (Figure 21). En corollaire, les caractères ciblés par cette étude peuvent se décliner en deux catégories : Des caractères à architecture génétique simple dont le contrôle génétique est principalement modulé par des QTLs majeurs tels que la production éthylénique et la couleur et des caractères complexes dont les QTLs identifiés ne représentent qu'une part faible de la variance. Tel est le cas de la teneur en glucose avec un QTL n'expliquant que 10% des différences phénotypiques pour ce caractère.

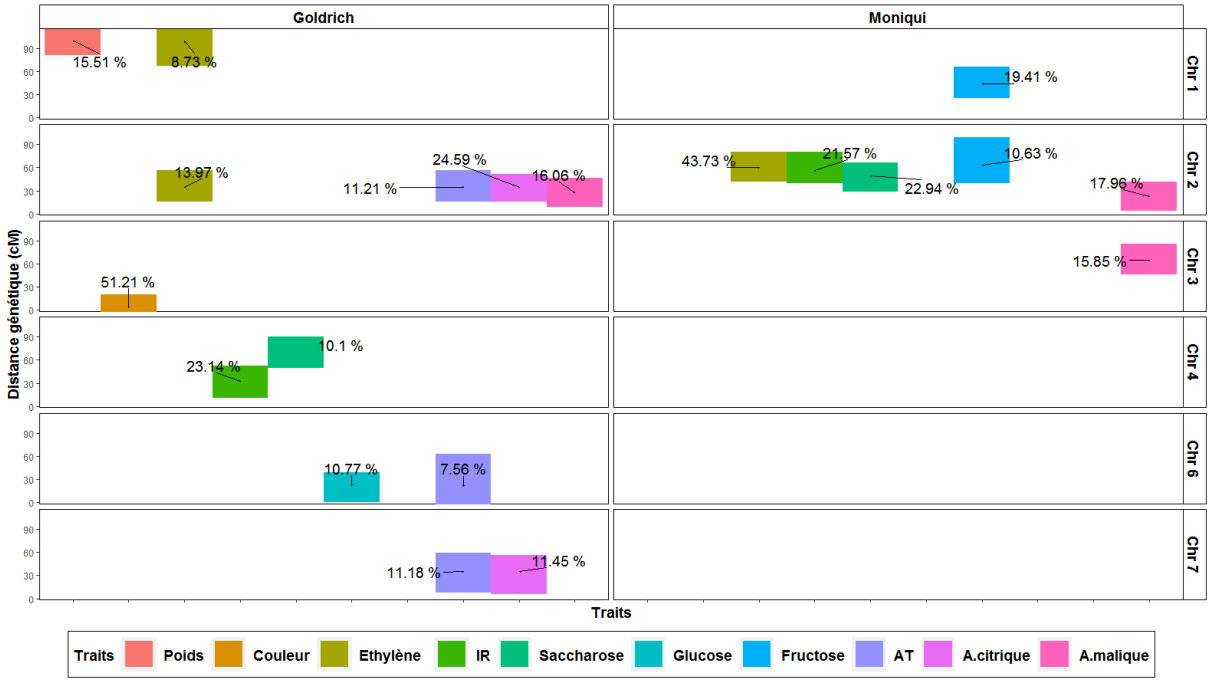


Figure 21: Représentation graphique des QTLs détectés pour les populations en pseudo-testcross

En concordance avec l'analyse de liaison génétique interannuelle, les analyses annuelles ont révélé la colocalisation entre plusieurs QTLs notamment ceux du saccharose et l'IR sur le GL4 de Goldrich ainsi que l'acide citrique et l'AT sur le GL 7 de Goldrich. Le glucose et l'AT ont présenté également des QTLs colocalisant sur le GL 6 de la carte de Goldrich. Un cluster de QTLs a été identifié pour l'éthylène, l'IR, le saccharose et l'acide malique sur le GL2 de Moniqui. Les clusters de QTLs suggèrent la présence d'interactions de nature pléiotropique entre les QTLs détectés ou une liaison génétique entre eux.

La colocalisation des QTLs liés à la production éthylénique avec ceux liés aux différents attributs de la qualité des fruits notamment les sucres, les acides et la couleur soulignent le rôle de l'éthylène comme étant une phytohormone qui interfère dans plusieurs voies métaboliques et dont la biosynthèse est couplée aux changements physiques et biochimiques se produisant lors de la maturation des fruits, tels que la dégradation de la chlorophylle, l'accumulation de caroténoïdes et la modulation de la teneur en sucres, ainsi que des changements dans les profils d'acides organiques (Paul et al. 2012).

3.4. Prédictions génomiques univariées

Nous avons étudié les performances prédictives de six modèles (RR-BLUP, Bayes A, Bayes B, Bayes C, BL et BRR) par le biais d'une stratégie de validation croisée 75% entraînement et

25% validation, avec 100 itérations. La précision de prédiction moyenne variant de 0.31 avec BL (glucose) à 0.78 avec RR-BLUP (éthylène) reflète la supériorité de la régression Ridge pour la majorité des caractères tandis que Bayes B a montré une performance de prédiction meilleure pour la couleur du fond, étant un caractère simple contrôlé par un QTL majeur (Figure 22). Outre une capacité prédictive moyenne plus performante pour la quasi-totalité des caractères évalués, RR-BLUP a présenté la plus faible variation en termes de précision de prédiction notamment pour l'éthylène et l'acide citrique. Quant aux facteurs contrôlant la performance des modèles, l'augmentation de la taille de la population d'entraînement s'est traduite par une amélioration de la précision. En revanche, le nombre de marqueurs optimal est conditionné par une couverture suffisante du génome au-delà de laquelle la précision diminue après avoir atteint un plateau correspondant à une capacité prédictive maximale.

Au-delà de la précision des prédictions, la population d'étude étant constituée de plein-frères présente une proportion considérable de partage d'allèles et un DL fort entre marqueurs et QTLs. Outre l'identité par descendance et l'étendue du DL, les prédictions génomiques intra-famille reflètent une distance génétique étroite entre les deux partitions d'entraînement et de validation.

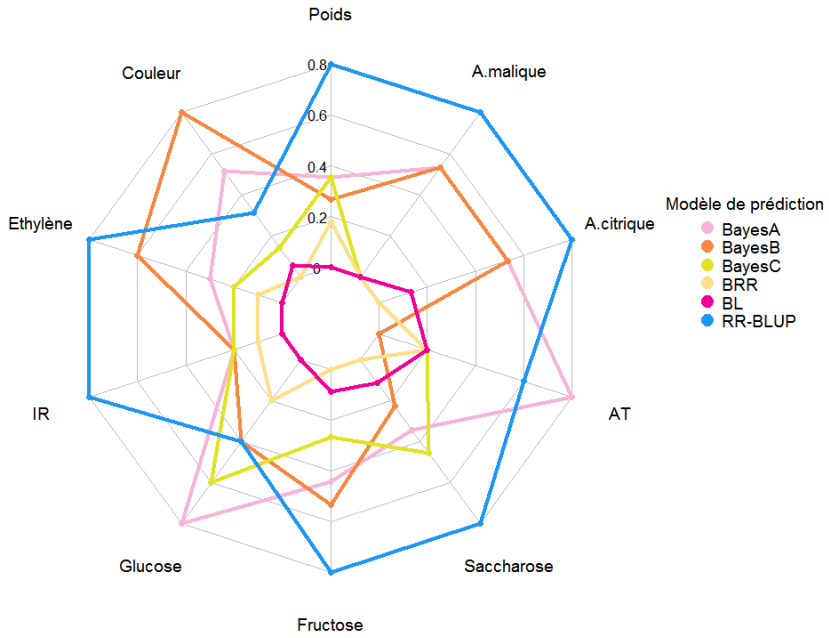


Figure 22: Précision de la prédiction génomique

La mise en place de la SG au sein des populations biparentales fournit un cadre robuste et stable au regard de l'interférence de l'apparentement et du patrimoine de DL en vue d'assurer la fiabilité et la précision des prédictions. En revanche, l'un des principaux écueils se présentant aux prédictions génomiques dans une même population de plein-frères est le manque de flexibilité et par conséquent la difficulté de l'extrapolation de ces prédictions dans des populations intégrant de la diversité ou des populations biparentales dérivant de fonds génétiques très éloignés de la population dans laquelle a été construit le modèle.

3.5. Prédictions généomiques informées par l'architecture génétiques des caractères

Ce scénario d'optimisation s'appuie sur la valorisation des résultats de l'analyse de liaison entre variation génétique et variation phénotypique réalisée dans les différentes partitions d'entraînement. Les modèles de SG incluant les QTLs préalablement identifiés ont présenté une meilleure capacité prédictive en comparaison avec les modèles de référence pour la quasi-totalité des caractères. La réponse des modèles à l'inclusion de cette information est variable en fonction des caractères et le gain en termes de précision de prédiction est tributaire de la représentativité des QTLs par rapport à la variance totale. Prenons le cas du caractère lié à la couleur pour lequel l'inclusion de deux QTLs expliquant 58% de la variation phénotypique a résulté en une amélioration de la précision pour les six modèles évalués allant de 7.7% pour Bayes B jusqu'à 36% pour BRR. En revanche, pour un caractère plus complexe tel que la teneur en glucose, l'ajout des QTLs à effet faible en covariables a entraîné soit une légère optimisation ou une diminution de la précision de prédiction (Figure 23).

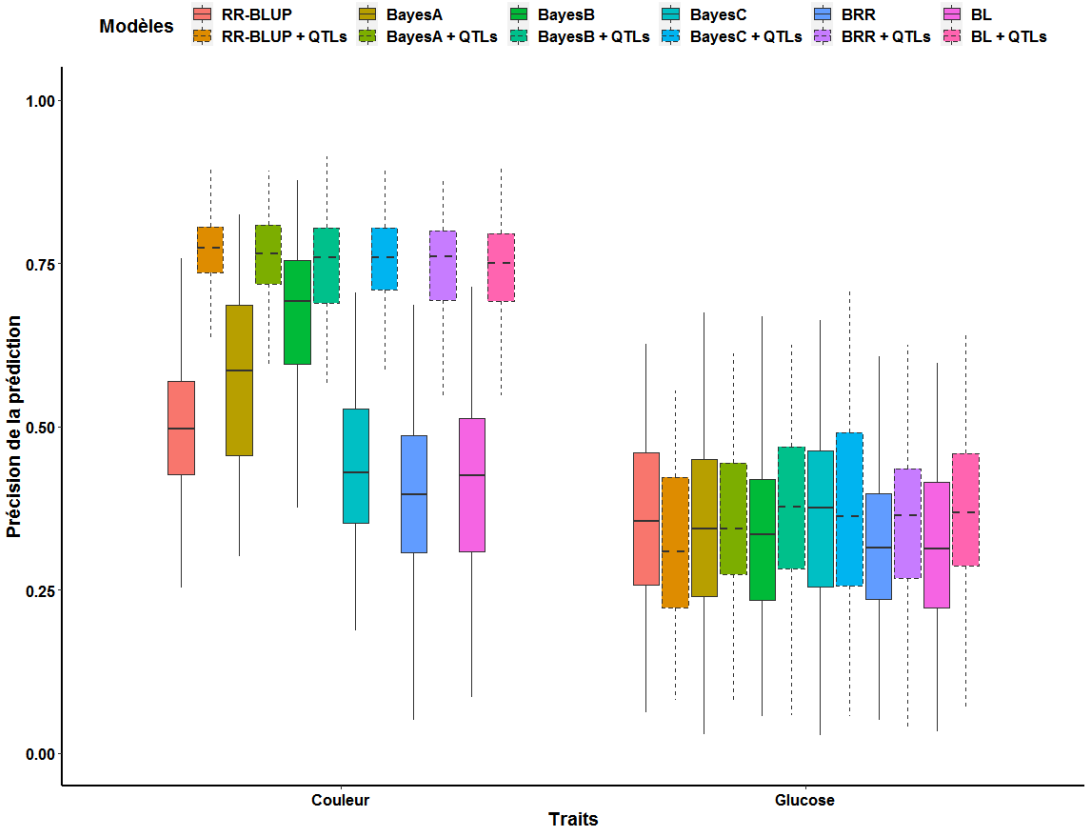


Figure 23: Comparaison entre les modèles optimisés avec l'information a priori sur le déterminisme génétique des caractères et les modèles de référence sans a priori

3.6. Prédictions génomiques bivariées

Dans l'optique d'optimisation de capacité prédictive des modèles de SG, nous avons exploré le cadre de prédiction bivariée consistant à mobiliser l'information phénotypique apportée par un caractère facile à mesurer pour prédire un caractère focal difficile à phénotyper. Ce cadre de prédiction mime un scénario de sélection dans des conditions opérationnelles où le focus est porté sur plusieurs caractères simultanément et la disponibilité des ressources est limitée à des caractères faciles à mesurer tels que la couleur, l'IR, l'AT ou le poids du fruit. Ainsi, dans tous les programmes de sélection, il existe des caractères dont le phénotypage est laborieux et difficile à mettre en œuvre et le recours à la sélection multi-trait pourrait être une opportunité prometteuse en vue de l'accélération du progrès génétique.

Les modèles bivariés se sont basés soit sur des valeurs phénotypiques (modèle bivarié) ou sur un index incluant les GEBVs correspondants aux proxies (modèle basé sur index). Dans le cas particulier du modèle basé sur un index de sélection où la GEBV du caractère focal est calculée comme indiqué dans l'équation :

$$GEBV_{focal} = GEBV_{proxy}\beta_{proxy} + u_i \quad 3.1$$

$GEBV_{focal}$ est la valeur génétique du caractère principal ou focal, $GEBV_{proxy}$ est la valeur génétique du caractère secondaire ou proxy et u_i est l'effet génétique de l'individu i.

Le paramètre β_{proxy} détermine l'effet du caractère proxy sur le caractère focal, ce qui est étroitement lié à la corrélation génétique entre les deux caractères. Par conséquent, la performance des modèles bivariés est tributaire de la corrélation entre caractères proxy et focal. Donc, le gain en précision est d'autant plus important que la corrélation génétique est forte. Pour les caractères fortement génétiquement corrélés notamment l'IR et les teneurs en sucres ainsi que l'AT et l'acide citrique, la précision de prédiction des modèles bivariés est systématiquement supérieure par rapport aux modèles univariés, comme indiqué dans la Figure 24.

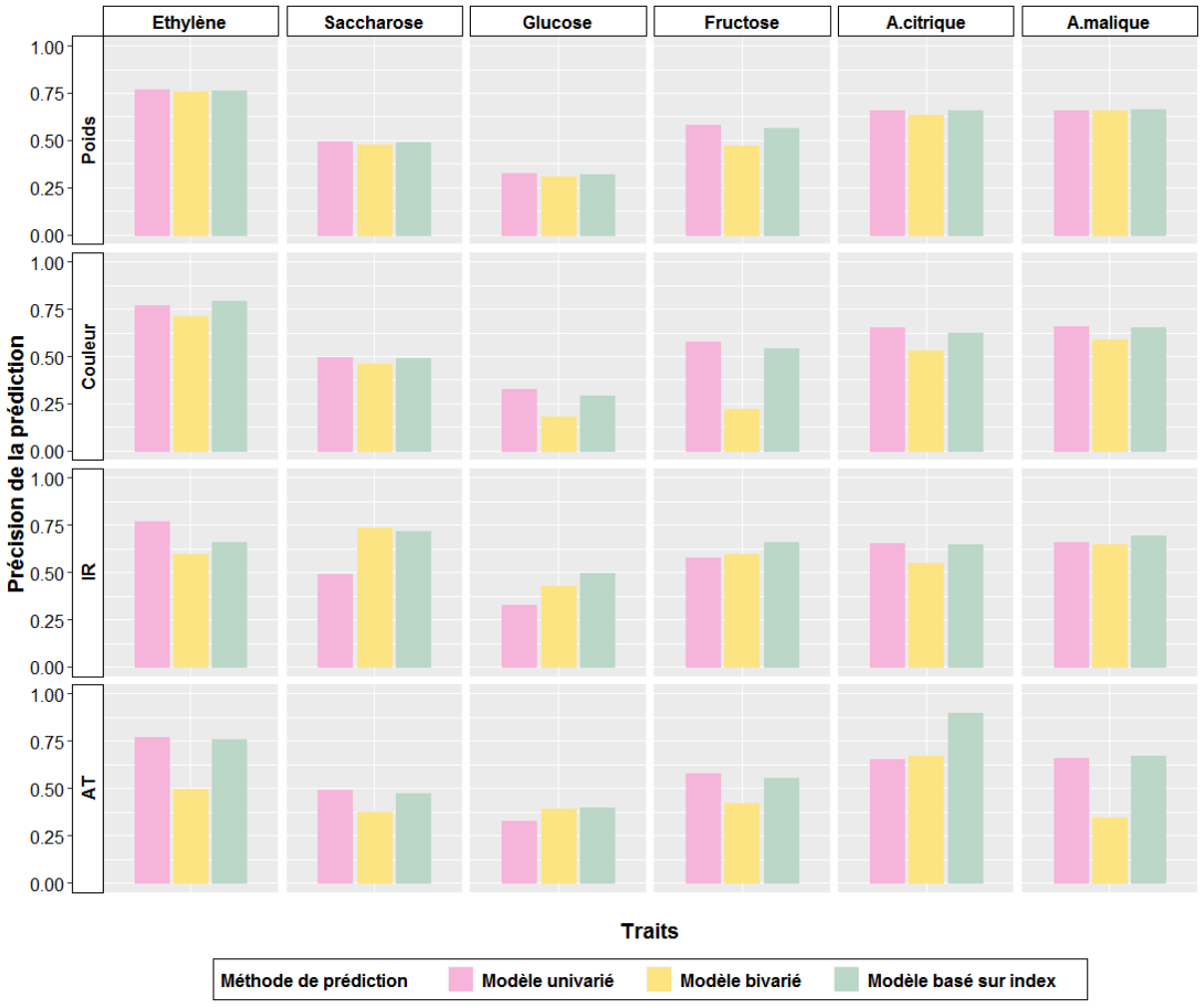


Figure 24: Précision de la prédiction génomique bivariée comparée à la précision de prédiction univariée

3.7. Conclusion

Nous avons présenté dans ce chapitre, l'efficacité de la sélection génomique au regard de la précision de prédiction de caractères supposés masqués dans un dispositif biparental présentant des caractéristiques contrastées par rapport à la qualité des fruits.

Nous avons exploré des scénarios d'optimisation de la précision de prédiction. D'une part, l'augmentation de la taille de la population d'entraînement et le nombre de marqueurs ainsi que le choix du modèle de prédiction en fonction de l'architecture génétique des caractères cibles permettent d'améliorer la précision de prédiction.

D'autre part, la prise en compte de l'information apportée par la régression de la variation génétique par rapport à la variation phénotypique offre un cadre d'optimisation d'intérêt notamment pour les caractères régis par des QTLs majeurs. Par ailleurs, la prédiction bivariée offre l'opportunité de s'appuyer sur des caractères dont le phénotypage est facile à mettre en œuvre pour prédire des caractères d'intérêt qui leur sont fortement corrélés.

Adoption and optimization of genomic selection to sustain breeding for apricot fruit quality

Mariem Nsibi*, Barbara Gouble †, Sylvie Bureau†, Timothée Flutre‡, Christopher Sauvage*,[2], Jean-Marc Audergon*,[1], Jean-Luc Regnard§

* INRAE, Génétique et Amélioration des Fruits et Légumes, 84143 Montfavet Cedex, France,

†INRAE, Avignon University, UMR SQPOV, 84914 Avignon, France,

‡ INRAE, CNRS, AgroParisTech, Univ. Paris-Saclay, GQE-Le Moulon, 91190 Gif-sur-Yvette, France

§ Univ Montpellier, CIRAD, INRAE, Institut Agro, UMR AGAP, 34398 Montpellier Cedex 5, France ORCID IDs: 0000-0002-9335-5091 (S.B.); 0000-0003-4489-4782 (T.F.); 0000-0001-5466-9955 (C.S.) ; 0000-0002-3132-5815 (J.-M.A.); 0000-0001-8614-0618 (J.-L.R.)

ABSTRACT Genomic selection (GS) is a breeding approach which exploits genome-wide information and whose unprecedented success has shaped several animal and plant breeding schemes through delivering their genetic progress. This is the first study assessing the potential of GS in apricot (Prunus armeniaca) to enhance postharvest fruit quality attributes. Genomic predictions were based on a F1 pseudo-testcross population, comprising 153 individuals with contrasting fruit quality traits. They were phenotyped for physical and biochemical fruit metrics in contrasting climatic conditions over two years. Prediction accuracy (PA) varied from 0.31 for glucose content with the Bayesian LASSO (BL) to 0.78 for ethylene production with RR-BLUP, which yielded the most accurate predictions in comparison to Bayesian models and only 10% out of 61,030 SNPs were sufficient to reach accurate predictions. Useful insights were provided on the genetic architecture of apricot fruit quality whose integration in prediction models improved their performance, notably for traits governed by major QTLs. Furthermore, multivariate modeling yielded promising outcomes in terms of PA for highly genetically correlated traits. This provides a useful framework for the implementation of indirect selection based on easy-to-measure traits. Thus, we highlighted the main levers to take into account for the implementation of GS for fruit quality in apricot, but also to improve the genetic gain in perennial species.

KEYWORDS: *Prunus armeniaca*, Genomic, Prediction, Cross validation, Accuracy, Genetic architecture, Multivariate modeling

[1]Corresponding author: INRAE, Génétique et Amélioration des Fruits et Légumes, 84143 Montfavet Cedex, France, E-mail: jean-marc.audergon@inrae.fr

[2]present address : Syngenta SAS France, 1228 Chemin de l'Hobit, 31790 Saint Sauveur, France

INTRODUCTION

Apricot (*Prunus armeniaca*) is a perennial fruit crop pertaining to Rosaceae family and *Prunus* genus, which encompasses several economically important species such as peach, almond, cherry and plum. It is one of the leading stone fruit species due to its economic contribution to the fruit industry. From a biological standpoint, apricot is characterized by its diploid genome (2n = 2x = 16) of 294 Mb/1n and its high heterozygosity (Arumuganathan and Earle 1991). The availability of a high-quality genome sequence in peach, defined as a reference *Prunus* species highly genetically characterized (Infante *et al.* 2008; Verde *et al.* 2013), as well as the high level of synteny between the *Prunus* species, have paved the way for elucidating the genetics of key commercial traits in *Prunus* species (Aranzana et al. 2019a). They provide both

a powerful framework for apricot genetic improvement and valuable tools to elucidate the genetic architecture of traits of interest.

Since the sixties, apricot breeding programs have been geared towards conventional breeding based on mass field selection, a time-consuming and labor-intensive process, which might reach 15 to 20 years from pre-breeding to the release of a new variety. Besides the length of apricot breeding cycle, several biological features inherent to this species impede genetic progress such as its wide range heterozygosity and a preferential self-incompatibility regime that induces uneven production according to climatic conditions. Recently, a particular focus has been projected towards fruit quality, a dynamic concept which encompasses a broad range of attributes linked to attractiveness, flavor, taste and texture with reference to fruit color, balance of sugars and acids and shelf-life. The burgeoning interest in fruit quality aims at shaping a sustainable fruit industry taking into account consumer preference trends that are expressed in a competitive landscape faced with climate change. Furthermore, commercial depreciation due to ripeness deficiencies resulting from early harvest and susceptibility to flesh mealiness incented the breeders to circumvent the issues linked to postharvest quality and thus contribute to the enhancement of fruit quality metrics (Gatti *et al.* 2009). In the scope of intrinsic challenges of apricot, fruit quality-oriented selection schemes aim to meet consumers' needs for improved quality attributes and address stakeholders' demands in the apricot sector. Therefore, controlled cross-pollination schemes allow recombination of desirable characteristics according to the integrated concept of ideotype, which is likely to guide biological designs of improved varieties through the identification and the integration of causal variants for high-value traits (Ramstein *et al.* 2019).

Here emerges one of the prominent impetuses of marker-based breeding approaches, a breeding strategy whose feasibility strongly tailored by the genetic architecture of target traits. Indeed, marker-assisted selection (MAS), is particularly relevant for monogenic inheritance, whilst genomic selection (GS), a novel breeding approach that has revolutionized animal and plant breeding communities, is favored for oligogenic and polygenic inheritance (Kumar et al. 2012b). GS is likely to capture the missing heritability of complex traits by modeling thousands of single nucleotide polymorphisms concomitantly (Makowsky et al. 2011a; Resende et al. 2012b). Meuwissen et al.'s landmark article (2001) laid the foundation for predicting genetic merit in plant and animal breeding and thus identifying superior genotypes among selection candidates according to their whole-genome sequence information (de los Campos *et al.* 2013). Unlike MAS, which pinpoints putative genes underlying the traits of interest, GS potentially considers all markers' effects without prior selection (Makowsky et al. 2011a; Resende et al. 2012b). This breeding approach is at its outset for crop plants and notably for perennial trees that are characterized by long breeding cycles due to the length of juvenile phase and generation time. Therefore, the recourse to GS for perennial species arises from the need to accelerate the pace of the breeding process. The relevance of this breeding approach has been assessed in forest trees such as eucalyptus (Resende et al. 2012b), black spruce (Lenz *et al.* 2017), white spruce, loblolly pine (Resende et al. 2012d), maritime pine (Bartholomé *et al.* 2016) as well as in perennial fruit crops such as grapevine (Fodor et al. 2014a), apple (Muranty *et al.* 2015), citrus (Minamikawa et al. 2017a), cranberry (Covarrubias-Pazaran et al. 2018b) and kiwifruit (Testolin 2011).

Large-scale genomic information against limited phenotypic records leads to an ascertainment bias due to the number of predictors (p), which is higher than the number of observations (n), resulting in multicollinearity and overfitting and accordingly low prediction performance (Desta and Ortiz 2014). To alleviate this statistical challenge due to dimensionality, a wide range of mathematical models are intended to infer linear combinations

of the original predictors in order to reduce through shrinking regression coefficients back towards zero (Whittaker et al. 2000; Gianola et al. 2003; Solberg et al. 2009). The extent of shrinkage of the marker effects differ across prediction models. For instance, in ridge regression shrinkage is performed equally across markers. However, this assumption is likely to be unreal because some markers are in linkage disequilibrium (LD) with loci with no genetic variance (Goddard and Hayes 2007). Conversely, in models designed under the Bayesian framework, shrinkage of effects is marker-specific (Crossa et al. 2017a). Further, the performance of GS is markedly influenced by several factors including marker density (Grattapaglia and Resende 2011; Lenz *et al.* 2017), training population size (Grattapaglia and Resende 2011), genetic relationship between training population and breeding population (Isidro et al. 2014; Lenz et al. 2017; Rincent et al. 2012), population structure (Zhong et al. 2009a; Rincent et al. 2017), the extent of LD (Daetwyler et al. 2008; Wientjes et al. 2013; Liu et al. 2015b), statistical models (Lorenzana and Bernardo 2009; Heslot et al. 2012; Resende et al. 2012d; Onogi 2020), trait heritability (Calus et al. 2008a) as well as the genetic architecture of target traits (Daetwyler et al. 2010a; Morgante et al. 2018). Along with the ideotype concept, multiple traits of interest can also be considered simultaneously through multivariate models to achieve more accurate predictions in comparison to single-trait models. Several simulation and empirical studies shed light upon the significant potential of multiple trait genomic prediction in optimizing prediction performance (Calus and Veerkamp 2011; Guo et al. 2014; Karaman et al. 2018; Covarrubias-Pazaran et al. 2018b; Michel et al. 2017). In this regard, the selection index strategy permits breeders to obtain genotypes that concomitantly incorporate several desirable characteristics. However, the efficiency of selection for multiple traits simultaneously depends considerably on the genetic correlation between these traits, that reflects the extent to which selection for a focal trait triggers an indirect response to selection for a secondary trait (Akdemir *et al.* 2018; Rana *et al.* 2019). In conjunction with the optimization of prediction model design, the integration of the insights gained by elucidating the genetic architecture of traits under selection might be of great interest. Several studies have emphasized the potential of including genomic information underlying the variation of target traits in prediction models (Spindel et al. 2016; Fang et al. 2017; Lopes et al. 2017; Liu et al. 2019b).

Therefore, the main objectives of our study were to (1) gain further insights into key fruit quality traits which are difficult to access, (2) evaluate the performance of GS prediction model applied to breeding for apricot fruit quality and (3) optimize GS accuracy by accounting for QTL mapping findings in prediction models and performing predictions under a multivariate framework.

MATERIALS AND METHODS

Plant material

The plant material used in this study is a F1 pseudo-testcross progeny of 153 individuals issued from a cross between 'Goldrich' and 'Moniqui' cultivars, which exhibit contrasted fruit quality traits. 'Goldrich' cv., used as female parent, is a North American early-season apricot cultivar. Self-incompatible, it is characterized by large, firm, orange fruit without blush and with a high level of acidity (Munoz-Sanz *et al.* 2017). 'Moniqui' cv., used as male parent, is a Spanish season apricot cultivar. Self-incompatible, it is characterized by large, soft and tasty white flesh fruit. 'Moniqui' is characterized by a high ethylene production, which results in a fast evolution at maturity and post-harvest, while 'Goldrich' presents a lower ethylene production, which results in an average fruit evolution at maturity and post-harvest. The F1 progeny was grown at the INRAE experimental field of Amarine in southern France. Seedlings were randomly planted on their own roots in 2005. Trees were managed under integrated

management system which implies that orchards are geared towards a sustainable production system with a trend towards the reduction of the use of chemical products.

Phenotyping for fruit quality

The phenotypic characterization of the Goldrich × Moniqui (Go×Mo) progeny was carried out for 22 quality traits over two consecutive years 2006 and 2007, which showed contrasted climatic conditions. A total of 40 fruits per genotype were randomly collected close to physiological maturity stage. Fruits were sorted according to their global firmness, determined by the pressure (kPa) required to achieve 3% deformation of fruit height with a multipurpose texture analyzer (Pénélaup, Serisud, Montpellier, France). Fruits were subdivided into three homogenous lots of four fruits per genotype of contrasting maturity: commercial maturity stage with pressure from 130 to 80 kPa, half ripe stage from 80 to 50 kPa and mature fruits with firmness less than 50 kPa. The physical, physiological and biochemical traits were measured on these three representative batches for all genotype. The fruit weight (kg) was measured at the same time as firmness. The skin color of the blushed side (sun-exposed side) and ground color of the non-blushed side (unexposed to sunlight), as well as flesh color, were determined using a CR-400 chromameter (Minolta, Osaka, Japan) and expressed in the CIE 1976 L*a*b*color space (illuminant D65, 0°view angle, illumination area diameter 8 mm). Hue angle, was computed using the chromaticity coordinates a^* and b^* as follows:

$$Hue = tg^{-1}(b^*/a^*)$$

The ethylene production rate was assessed as physiological parameter linked to maturity stage of climacteric fruits. Ethylene production, expressed in nmol $kg^{-1}h^{-1}$, was measured by gas chromatography after 1 h of confinement in a hermetically closed jar (Chambroy et al. 1995; Bureau et al. 2009a). Then, flesh color was measured and fruits were cut and frozen at -20 °C for further biochemical analyses. Fruit stones were weighed individually (kg). Fruit pieces were ground with an Ultra-Turrax T25 equipment (Ika Labortechnik, Staufen, Germany) to obtain a slurry. The refractive index (RI) which stands for the solid soluble content (SSC) was determined with a digital refractometer (PR-101 ATAGO, Norfolk, VA) and expressed in °Brix at 20 °C. Titratable acidity (TA) was determined by neutralization up to pH 8.1 with 0.1 N NaOH and expressed in meq 100 g^{-1} of fresh weight using an autotitrator (Methrom, Herisau, Switzerland). Soluble sugars (glucose, fructose, sucrose) and organic acids (malic acid and citric acid) were quantified using an enzymatic method using kits for food analysis (Boehringer Mannheim Co., Mannheim, Germany) and expressed in g 100 g^{-1} of fresh weight for sugars and meq 100 g^{-1} of fresh weight for acids. These measurements were performed with an automatic analyzer BM-704 (Hitachi, Tokyo, Japan). Hence, twenty-two quality parameters were assessed for apricot fruit characterization. Only ten traits of agronomic interest were selected in order to assess the efficiency of genomic prediction for fruit quality traits. Their choice was motivated by selecting the criteria which underpin consumers' perception of apricot fruit and meet the exigencies of stakeholders in the apricot sector. Among the selected traits, hue of the ground color was chosen as a commercial feature insightful for pigments such as carotenoids and chlorophyll. Ethylene production, plant hormone responsible for fruit ripening and precursor of several biosynthetic pathways linked to quality, organoleptic components provided by organic acids (malic and citric acids and TA) and sugars (sucrose, fructose, glucose and RI) as well as fruit weight figure among the traits used in the downstream analysis.

Statistical modelling of the phenotypic data

Statistical modelling of the fruit quality attributes was performed using R software version 3.6.0 (R Core Team, 2018). Significance assessment of variance components was carried out using ANOVA tests to determine the significant factors contributing to the phenotypic variation intended to be included in the adjustment model. In light of significance tests outcome, phenotypic data were adjusted using 'lmer' function provided in lme4 package (Bates et al. 2014a) within a mixed model framework whose equation can be described as follows:

$$y_{ij} = \mu + \alpha_i + \beta_j + \delta_k + \alpha\beta_{ij} + e_{ijk}$$

where y_{ij} is the phenotypic value of the genotype i for the year j, μ is the overall mean, α_i is the random effect of the genotype i, β_j refers to the fixed effect of the year j, δ_k is the fixed effect of the maturity group k corresponding to the fruit lot, $\alpha\beta_{ij}$ is the interaction effect of the genotype i and the year j and e_{ijk} is the random residual effect.

Heritability computation

Broad-sense heritability H^2 for fruit quality traits, defined as the proportion of phenotypic variance attributed to additive, dominance and epistatic patterns, was computed using the following formula:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{gy}^2}{n_y} + \frac{\sigma_l^2}{n_l} + \frac{\sigma_y^2}{n_y} + \frac{\sigma_e^2}{n_y * n_l}}$$

where σ_g^2 is the genetic variance, σ_{gy}^2 is the variance attributed to the interaction between genotypes and years, n_y refers to the number of years and n_l is the number of fruit lots.

Genotyping data

Genotyping by sequencing, performed according to the protocol described by (Elshire et al. 2011), was carried out within the FruitSelGen project. Regarding the high level of synteny between the apricot and peach genomes, the fastq sequences were aligned to the peach genome. Raw data filtering, sequence alignment and variant calling were performed using GATK software (Genome Analysis Toolkit) (McKenna *et al.* 2010). The outcome of a further filtering process using VariantAnnotation package (Obenchain et al. 2014b) resulted in a set of 61,030 SNPs with a genotype quality score greater than 20 and a missing rate lower than 5%. Out of the 184 individuals, 31 individuals exhibiting spurious genotypic profile were discarded and thus 153 individuals were kept for downstream analysis. SNP markers were coded as 0, 1 and 2, according to the number of copies of the alternative allele and missing marker information was imputed as the mean of the genotypic scores of non-missing data at the level of each maker.

Construction of the linkage maps

The advent of genomic selection to breed for apricot quality traits requires a better understanding of their genetic architecture. However, up to now, the genetic determinism underlying fruit quality in apricot was scarcely investigated (Ruiz *et al.* 2010). Thus, linkage mapping was performed in order to uncover the genetic architecture of the 10 fruit quality traits. Prior to QTL identification, two genetic linkage maps were constructed for the full-sib progeny using a pseudo-test cross mapping strategy (Grattapaglia and Sederoff 1994). The whole set of 61,030 SNPs was filtered according to Mendelian inheritance, and those presenting strong

deviation from Hardy-Weinberg equilibrium (p-value < 1×10^{-6}) were discarded using the function filterSegreg provided by rutilstimflutre package (Flutre et al. 2015). Afterwards, the markers which depicted more than 1% of missing information and more than 1% of genotyping errors were eliminated and linkage group clustering, marker ordering and genetic distance calculations were achieved by means of mstmap.data.frame function under ASMap package (Taylor and Butler 2017). Maps construction was performed using Kosambi's mapping function and a logarithm of the odds ratio (LOD) of 3.

QTL detection

In order to provide insights into the genetic architecture of the target traits, we performed a composite interval mapping strategy using R/qtl package (Broman et al. 2003a) with the aim of identifying the genomic regions underpinning apricot fruit quality. In this respect, 1,000 permutations were undertaken with a significance level set at 0.01 in order to identify putative QTLs and determine the threshold of LOD scores. Then, the part of phenotypic variance explained by SNPs significantly linked to target traits was estimated. Additionally, a joint QTL detection analysis was performed on two independent datasets recorded in 2006 and 2007 with the aim of assessing the stability of QTLs associated to the adjusted means corresponding to the phenotypic records. The graphical representation of the two genetic maps as well as QTL-linked markers was drawn using MapChart 2.3 software (Voorrips 2002a).

Genomic selection modelling

Prediction of the genomic estimated breeding values was performed using a baseline model where the genomic information as well as the phenotypic records were fitted in order to estimate marker effects and thus the breeding values:

$$y = X\beta + Zu + e$$

where y is the vector of the phenotypic records, X is an incidence matrix for fixed effects relating fruit quality to the vector of fixed effects β, β is a vector of fixed effects estimates, Z is an incidence matrix for random effects relating fruit quality to vectors of random additive genetic effect u, u is a vector of random effects and e denotes the vector of random errors.

Cross-validation procedure

The performance of prediction, mirrored in the predictive accuracy for the ten key quality traits, was assessed using a cross-validation strategy where data were randomly partitioned into 2 subsets: 75% of the reference set was assigned to the training set intended to calibrate the prediction model and the remaining 25% was used as the validation set whose phenotypes were assumed to be unknown. This cross-validation scheme was iterated 100 times where samples were drawn with replacement from the reference set. Pearson's correlation between predicted phenotypes and the observed ones was used to determine the accuracy of the predicted phenotypes.

Factors controlling genomic prediction accuracy (PA)

As several parameters control the prediction performance such as statistical models, training population size and marker density, these parameters were investigated using randomly drawn subsets of the reference dataset in order to point out the factors governing the potential variation in PA, to assess their respective effect.

Impact of statistical prediction models

Within the framework of genomic prediction, various statistical methods have been proposed in literature. These models share the same prediction equation for the estimation of the GEBVs whilst they are grounded on different assumptions concerning markers effects. Five Bayesian models were explored: Ridge regression best linear unbiased prediction (RRBLUP) model implemented in the rrBLUP package (Endelman 2011c) as well as Bayes A, Bayes B, Bayes C, Bayesian LASSO (BL) and Bayesian ridge regression (BRR) implemented in the package BGLR (Pérez and de los Campos 2014).

Impact of training population (TP) size

In order to explore the impact of population size on PA, we used three randomly drawn subsets of 43, 76, and 115 individuals corresponding to 25%, 50% and 75% of the respective study population to elaborate the prediction model and thus compute breeding values of the remaining individuals.

Impact of marker density

Furthermore, we assessed the extent to which randomly selected marker subsets of different sizes, from 1% to 100% could affect the accuracy of prediction.

Genomic prediction optimization

Herein, we assessed two prediction strategies with a view to the improvement of PA.

Accounting for genetic architecture

The first optimization scenario made use of the information brought by QTL mapping. Thus, SNPs tightly linked to QTLs with medium to large effects were included in the prediction models as fixed covariates in order to assess the prediction accuracy. The genomic prediction model which accounts for prior information on genetic architecture is defined as:

$$y = X\beta + Zu + e$$

Multivariate genomic prediction

A second scenario dedicated to optimizing genomic selection accuracy is the multi-trait prediction implemented using the R package 'sommer' (Covarrubias-Pazaran 2016), which provides a framework for fitting multivariate prediction models. Hence, this prediction strategy was performed using Genomic BLUP (GBLUP) model whose equation is provided below:

$$y = X_t\beta_t + Z_t u_t + e$$

where y is a N × t vector of the phenotypic records, X is an incidence matrix for fixed effects relating fruit quality to the vector of fixed effects β, which is a vector of fixed effects estimates, Z is an incidence matrix for random effects relating fruit quality to vectors of random additive genetic effect u, u is a vector of random effects and e denotes the vector of random errors and t is the number of traits.

Data availability

Supplemental data are available in Files S1-6. File S1 contains the raw phenotypic data. File S2 contains the estimations of trait heritability. The genotypic data are available in File S3. The results of QTL detection are summarized in File S4 and the two genetic maps are provided in File S5. Pairwise genetic correlation between the traits are available in File S6. R scripts are provided on request.

RESULTS

Exploration of the phenotypic data

The distribution of phenotypic values according to the maturity stage (Figure 1A) reflects the quantitative determinism of most apricot fruit quality traits, except for ethylene production rate which exhibited a skewed distribution and was adjusted with a logarithmic transformation to restore its normal distribution. Continuous distribution of the phenotypic records points to polygenic inheritance of most traits and a potential contribution of several QTLs to the phenotypic variation. The apricot quality traits were positively affected by maturity (Figure 1A). The average phenotypic values for fruit RI and sucrose content increased from 13.7 to 15 °Brix and from 5.5 to 7 g 100 g^{-1} of fresh weight respectively, while the titratable acidity decreased slightly from 25.9 to 21.8 meq 100 g^{-1} fresh weight. A sharp rise in the ethylene production rate was observed throughout ripening, which corresponds to a 94% increase from group 1 to group 3. Likewise, the ground color changed with maturity stage, the average value of which decreased from 73 to 71 degrees. The range of variability of phenotypic values varied from 8.5 % for ground color to 30.4 % for fructose content. The extent of phenotypic variation reflects the diversity within the genetic pattern of the study population issued from two varieties with contrasted fruit characteristics.

The partition of the phenotypic variance into different sources of variation (Fig. 1B) highlights the significant contribution of the year effect (from 0.1% for glucose to 13.5% of the sum of squares for malic acid) as well as the fruit maturity stage modelled by fruit groups (from 0.01% for fructose to 18.6% of the sum of squares for ethylene). The highest contribution to the phenotypic variation is attributed to the genetic pattern reflected in the genotype effect (from 38.4% for sucrose to 72.9% of the sum of squares for glucose) as well as to the interactions between genotype and year (from 8.2% for ethylene to 25.6% of the sum of squares for sucrose). This trend was endorsed by the moderate to high heritability estimates of apricot quality traits (Table 1), with broad-sense H^2 ranging from 0.56 for sucrose content to 0.92 for glucose content. Moreover, analysis of apricot fruit quality traits revealed high positive pairwise correlations between TA and Citric.A (r = 0.83), RI and Sucrose (0.73), Glucose and Fructose (r = 0.56), RI and Fructose (r = 0.50) and Ethylene production and Citric.A. (r = 0.47) and Ethylene production and Malic.A. (r = -0.44) (Figure 2).

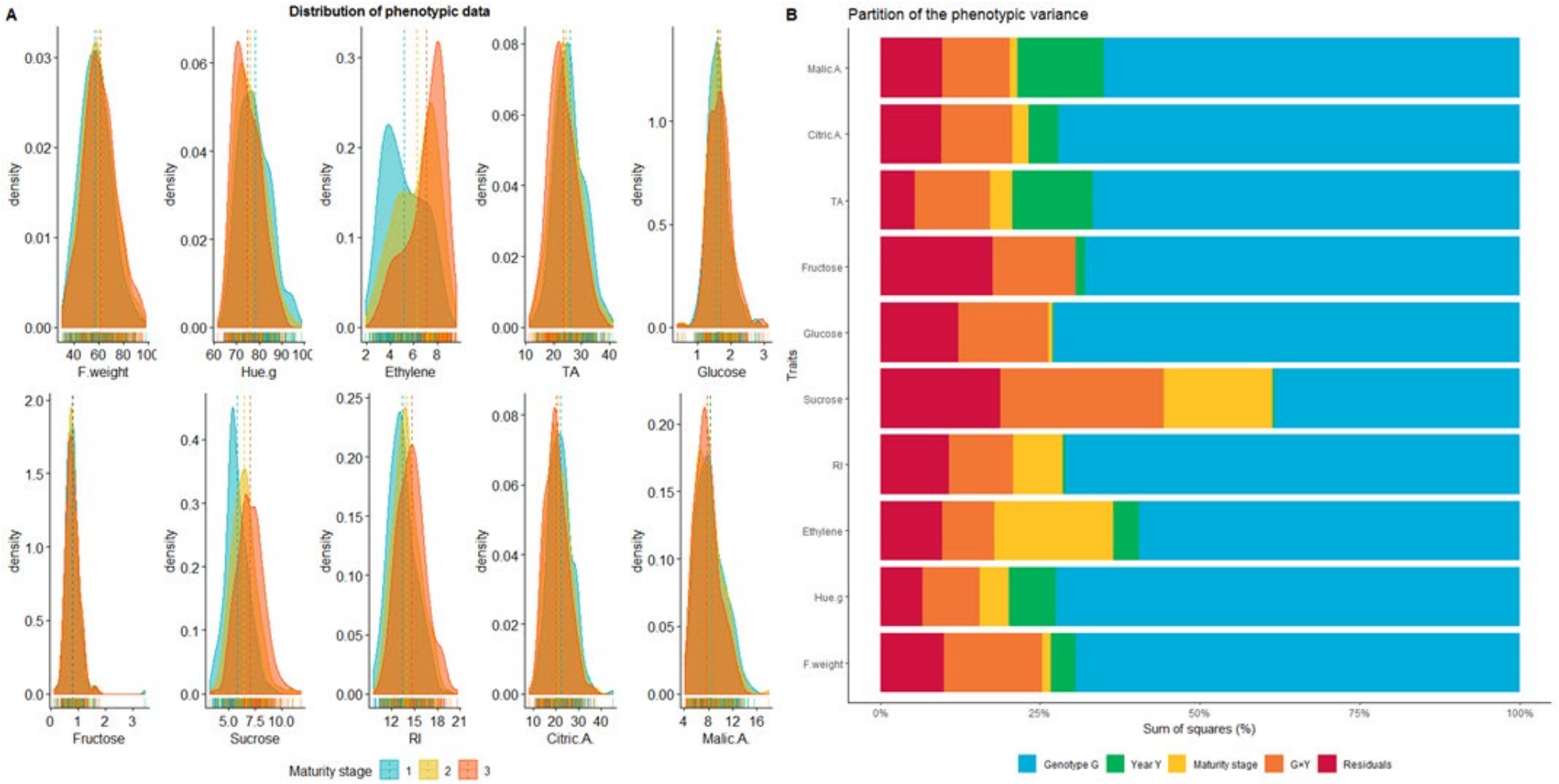


Figure 1 | Exploration of the phenotypic data : distribution of phenotypic values for the 10 apricot quality traits in the Go X Mo progeny (A) and components of the phenotypic variation with reference to genotype (G) effect, year (Y) effect, G×Y interaction and fruit stage of maturity (B).

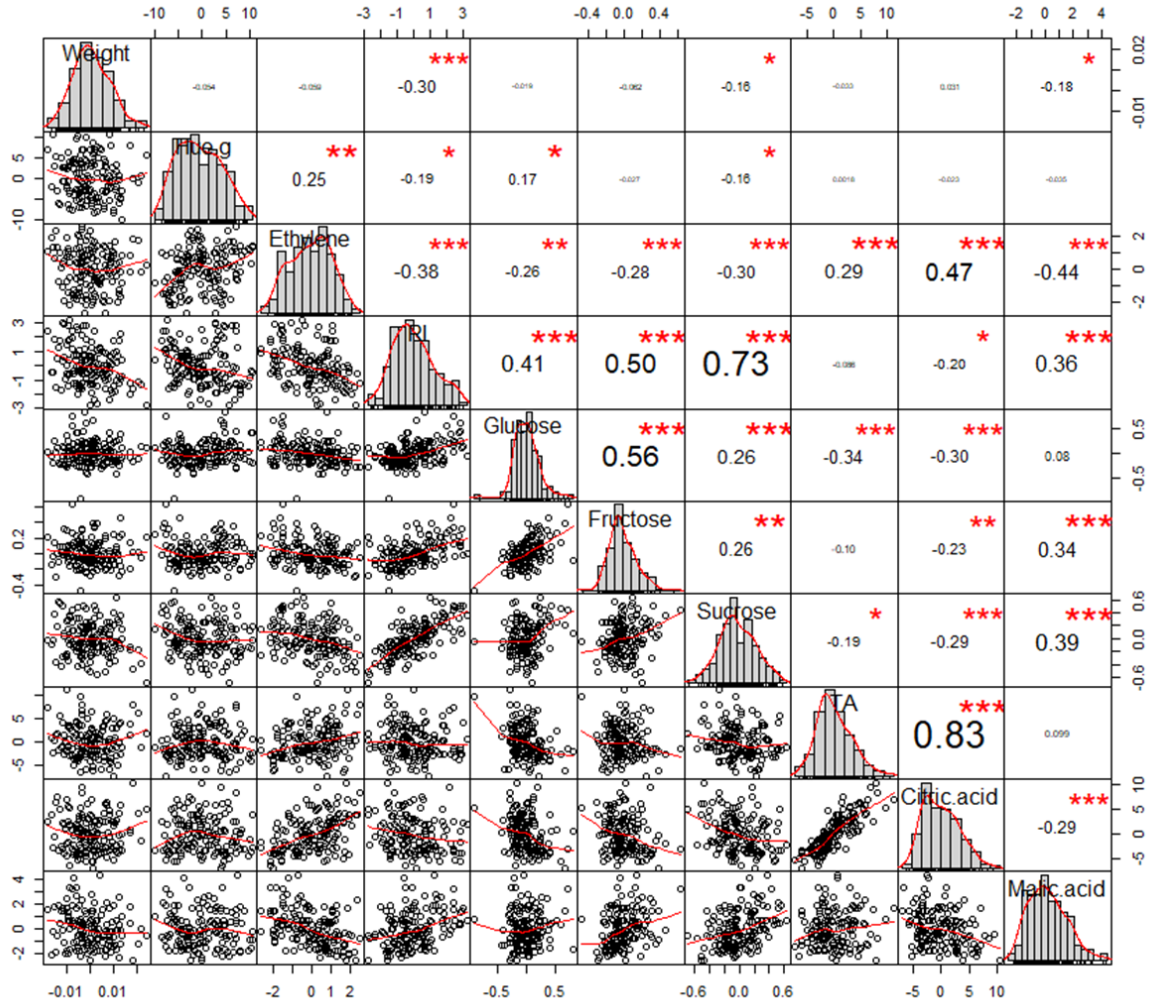


Figure 2 | Matrix of pairwise correlations: Bivariate scatterplots (lower off-diagonal) and correlation values between phenotypic values (upper off-diagonal) for 10 apricot fruit quality traits. The distribution of adjusted phenotypic values are shown in the diagonal.

Linkage map construction

The linkage mapping for apricot fruit quality was performed according to a pseudo-testcross mapping strategy. Beforehand, markers filtering was implemented according to Mendelian segregation leading to a final set of 4,922 SNPs, of which 2,311 were heterozygous for Goldrich and 1,395 were heterozygous for Moniqui. Then, we applied a stringent quality control by removing markers presenting more than 1% of genotyping error. A total of 366 SNPs was retained for Goldrich and 250 SNPs for Moniqui. Hence, two parental linkage maps were generated. SNPs mapped for Goldrich were distributed on eight linkage groups, which present an overall length of 562 cM and an average distance of 1.6 cM between adjacent SNPs. For the male parent Moniqui, the genetic map spanned an overall length of 842.3 cM with an average spacing of 3.5 cM between SNPs. The 250 SNPs mapped for Moniqui were positioned on 10 linkage groups (LG), where chromosomes 1 and 7 were split into 2 LGs each.

QTL detection

The linkage analysis was performed using BLUPs. The across-years analysis undertaken using composite interval mapping revealed 20 significant QTLs spread over all LGs except LGs 5 and 8 (Figure 3), which explained from 7.6% (TA) to 51.2% (Hue.g) of the observed phenotypic variance and whose peak LOD values varied from 3.44 (TA) to 23.8 (Hue.g) for the ten fruit quality traits (Table S4). Two major QTLs, that explain 23.1% and 21.6% of the phenotypic variability, were detected for refractive index (RI). One major QTL explaining 51.2% of the observed variability was found for ground color (Hue.g), one major QTL explaining 43.7% for Ethylene, one major QTL explaining 24.6% for Citric.A and one major QTL explaining 22.9% for Sucrose.

With reference to the annual linkage analyses, 19 QTLs were detected for apricot fruit quality in 2006 and 19 QTLs in 2007. Seven QTLs showed stability across years, being consistently detected in 2006 and in 2007. The amount of explained variance ranged from 7.9% (Ethylene) to 46.1% (Hue.g) for 2006 and from 0.5% (Fruit weight) to 46.7% (Hue.g) for 2007, while LOD values varied from 3.3 (TA) to 20.5 (Hue.g) for 2006 and from 3.5 (TA) to 20.9 (Hue.g) for 2007. Detailed information is provided in Files S4 and S5.

In terms of colocalization, QTLs for Sucrose coincided with QTLs for RI on LG2 for Moniqui and on LG4 for Goldrich, while QTLs for TA coincided with QTLs for Citric.A on LG7 for Moniqui, with QTLs for Glucose on LG6 for Goldrich and with Ethylene on LG2 for Goldrich (Figure 3). QTLs for RI and Ethylene colocalized on LG2 for Moniqui. In addition, QTLs intervals for Malic.A, Sucrose, Fructose, RI, Ethylene and Hue.g presented overlapping on LG 2 of Moniqui parental linkage map. Finally, QTLs associated to Glucose and TA coincided on LG6, while those for Fructose, Ethylene and F.weight overlapped on LG1 of Goldrich parental linkage map.

Genomic prediction accuracy (PA)

The GS prediction accuracy was assessed using different statistical models, different sizes of training population and different subsets of markers randomly distributed along the genome. Results are provided in Figure 4.

Factors controlling genomic prediction accuracy

Impact of statistical prediction models

We investigated prediction performance for six models (RR-BLUP, Bayes A, Bayes B, Bayes C, BL and BRR). Across the traits, the overall average PA varied from 0.31 with BL (for Glucose) to 0.78 with RR-BLUP (for Ethylene) (Fig. 4A). RR-BLUP outperformed Bayes A, Bayes B, Bayes C, BL and BRR for six traits out of 10 (F.weight, Ethylene, RI, Sucrose, Fructose and Malic.A). For three traits (Glucose, TA and Citric.A) similar PA was yielded using RR-BLUP and Bayes A, while Bayes B exhibited the best prediction performance for Hue.g. BL, BRR and Bayes C models provided similar prediction performance across all traits. Among the ten apricot fruit quality traits, Glucose and TA displayed the lowest average prediction accuracy regardless of the investigated model.

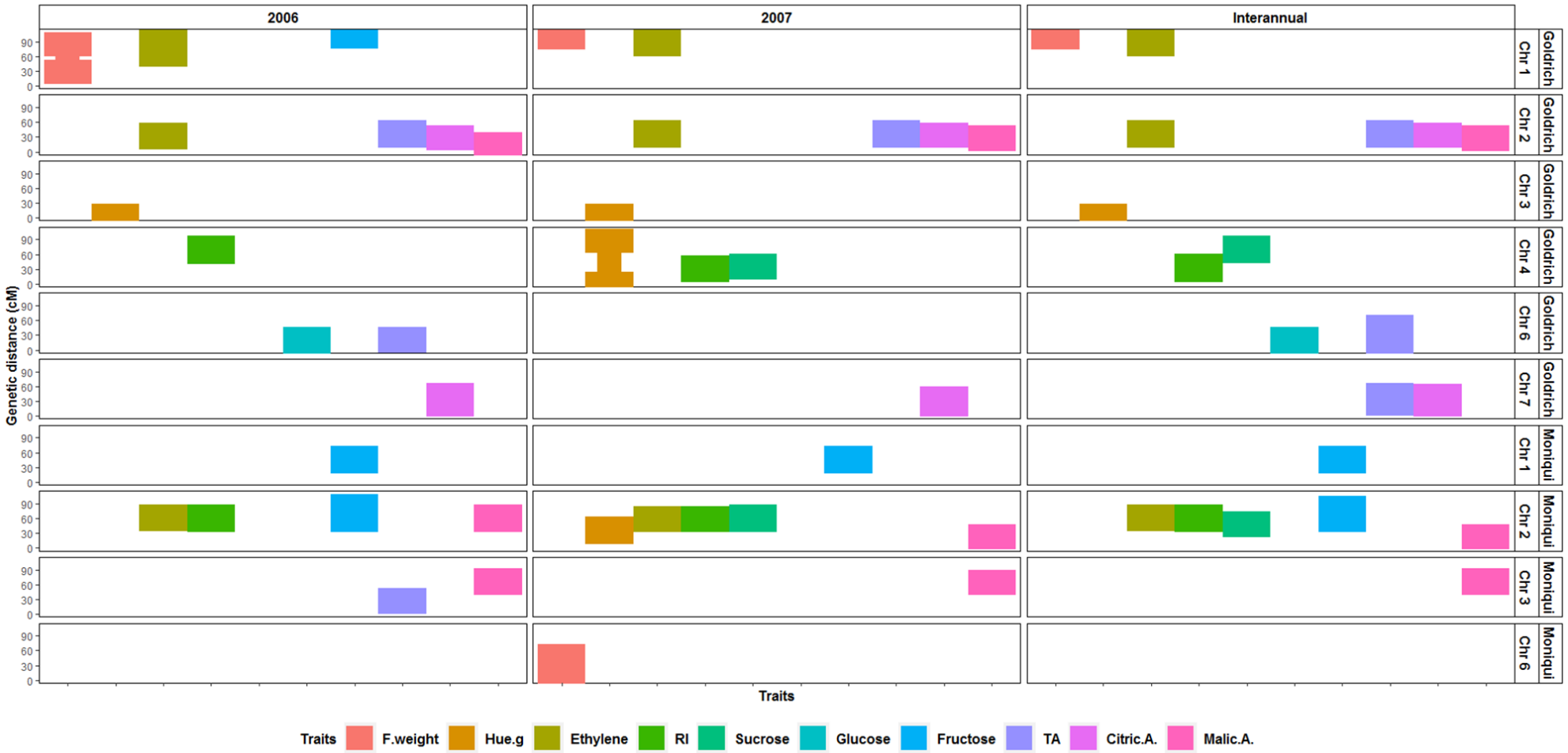


Figure 3 | QTL detection performed on adjusted phenotypic values of 10 apricot fruit quality traits according to a pseudo-test cross mapping strategy where two genetic linkage maps were constructed for Goldrich and Moniqui. A joint linkage analysis was carried out across years on two independent datasets recorded in 2006 and 2007. On the y axis, positions of SNPs significantly linked to targeted traits expressed in cM. On the x axis, traits distributed across years and genetic backgrounds. Only chromosomes enclosing significant SNPs are represented.

To further explore the impact of statistical models on prediction performance, we assessed the magnitude of variation appraised by standard deviation of PA of the tested models. Averaged over 10 traits, RR-BLUP showed the lowest variation in PA (0.09) and BRR the highest (0.11). The lowest variation in PA was noted for Ethylene (0.05) and Citric.A (0.09) and the highest variation for Glucose (0.15) and Fructose (0.14).

Then, assessment of variation in PA according to factors such as marker density and training population size was performed with RR-BLUP model, which represents an optimal compromise between prediction performance and computational time.

Impact of training population (TP) size

To assess the impact of TP size on accuracy, we performed genomic prediction using 43, 76 and 115 individuals. As shown in Figure 4B, the increase in TP size resulted in a substantial increase in PA, which ranged from 11% to 24% as a response to the increase in TP size from 25% to 75%.

Impact of the number of markers

PA increased with the number of markers, regardless of the trait genetic architecture, and became steady reaching a plateau at about 6,103 SNPs corresponding to 10% of the total number of markers (Figure 4C). No significant improvement in accuracy was noted when more than 6,103 SNPs were used and with only a rather small number of markers, the model was able to accurately predict the phenotypes in the validation set. Conversely, the average PA dropped from 0.55 to 0.25 across traits when the number of markers used in the prediction model dropped from 6,103 to 100, and the drop was steeper when the number of SNPs was below 50. In addition, decreasing the number of markers resulted in an increase in the variability of PA for all the traits under investigation. For instance, for ethylene production, the standard deviation ranged from 0.24 for 50 SNPs to 0.06 for 61,030 SNPs. Furthermore, it should be noted that traits with a rather simple genetic architecture due to the contribution of major QTLs to the phenotypic variation such as Ethylene, RI as well as Citric.A were the most sensitive to the variation of the number of markers.

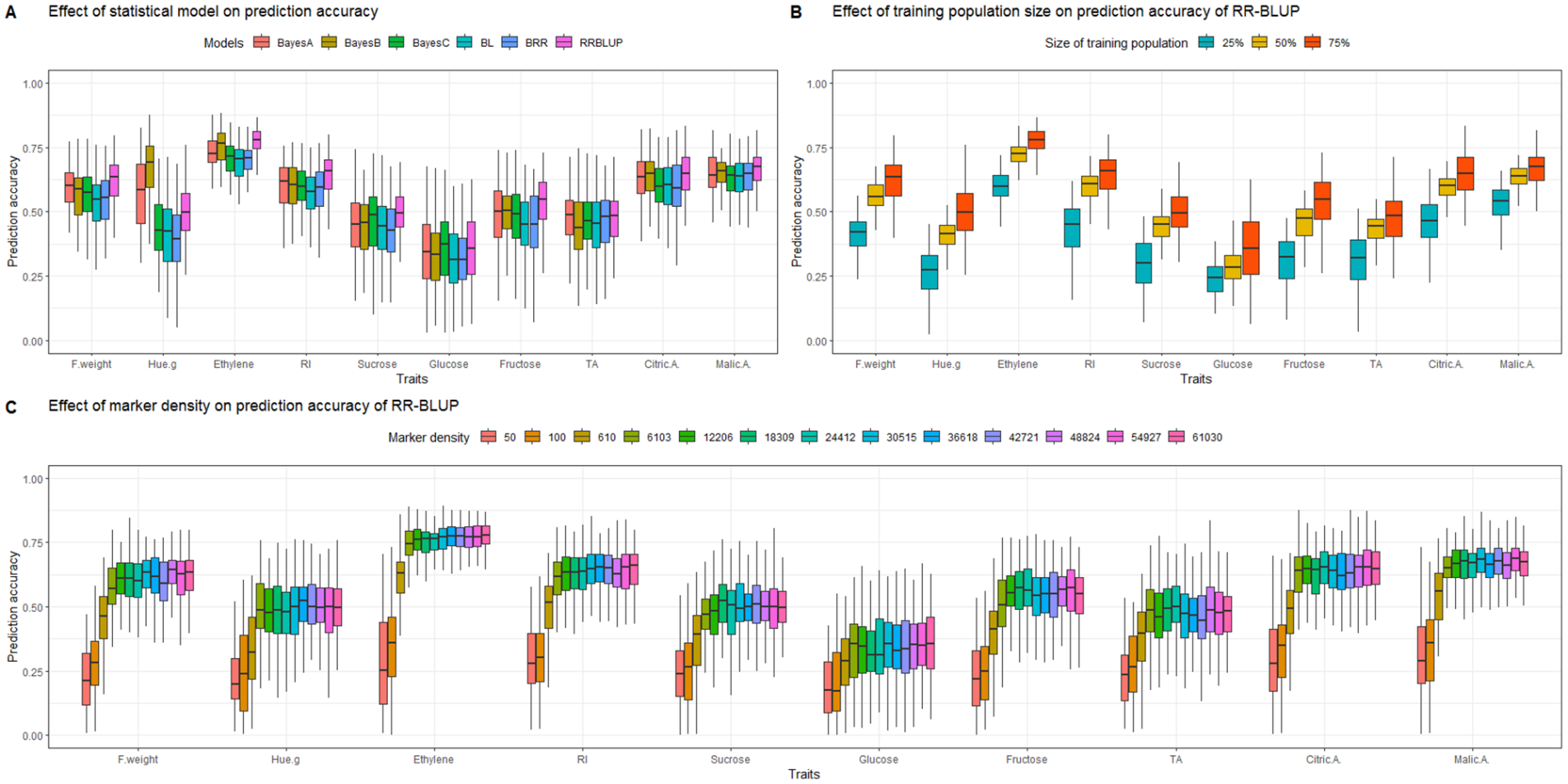


Figure 4 | Variation in accuracy of genomic prediction for ten apricot fruit quality traits using a random cross-validation (100 replicates).

Effect of the GS statistical model: RR-BLUP, Bayes A, Bayes B, Bayes C, Bayesian LASSO and BRR (A); Effect of the training population size using subsets of 43, 76, and 115 individuals randomly drawn corresponding to 25%, 50% and 75% of the study population using RR-BLUP model (B). Effect of marker density using randomly selected SNP subsets: 50, 100 and from 1% to 100% out of 61,030 SNPs using RR-BLUP model.

Optimization of genomic prediction

Optimization of the GS models

We investigated the effect of integrating prior knowledge about the trait genetic architecture on the accuracy of genomic prediction. The prediction performance of models in response to the inclusion of QTLs significantly linked to the traits of interest varied across models and traits (Figure 5). Accuracy gain derived from models with QTLs as fixed effects in comparison with models with markers as random effects was more pronounced for Hue.g for which adding two QTLs explaining 51.2% and 10% of phenotypic variance respectively resulted in an accuracy gain of 23.4% across the six investigated models. The magnitude of gain in prediction accuracy was tightly linked to the proportion of variance explained by QTLs added to prediction models with $R^2 = 0.49$. Furthermore, models built with regard to trait genetic architecture also permitted to enhance PA for Fructose, Sucrose and RI. For these traits, gains ranged from 4.7% to 11.3% subsequently to the inclusion of QTLs explaining 13.7% to 20.5% of variance. Nevertheless, the integration of QTLs as fixed effects in the models slightly decreased the prediction performance for Malic.A, F.weight and TA, with decreases ranging from -0.7% to -2.2%. PA for Ethylene also decreased but to a lesser extent (-0.5%). Furthermore, the prediction models differed in their response to the integration of prior genomic information. The gain in PA ranged from 2.3% to 6.0%. The highest gain was observed for the BRR model, although Bayes B depicted the lowest increase in PA.

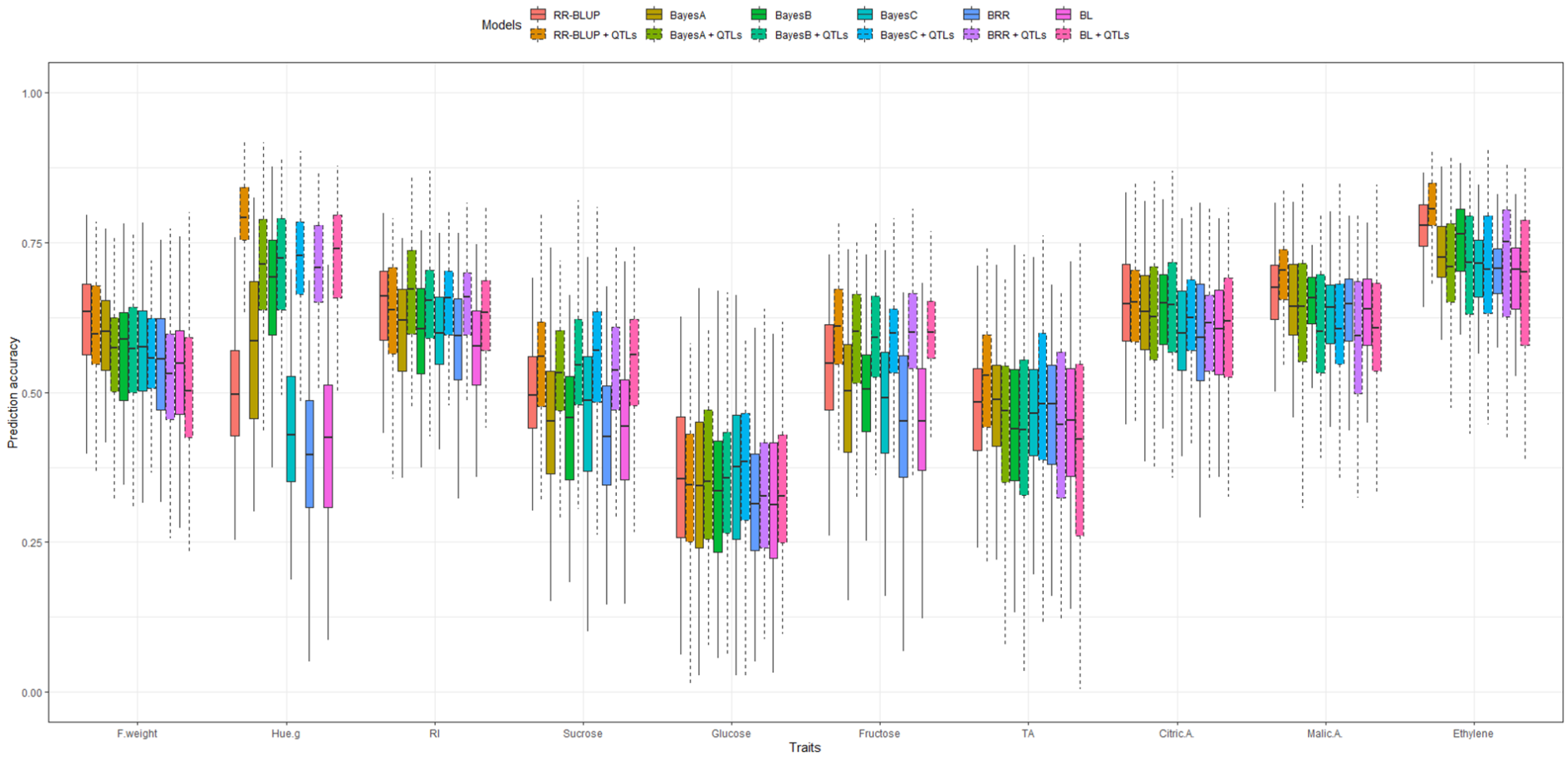


Figure 5 | Accuracy of genomic prediction for 10 apricot fruit quality traits using six statistical models treating all SNPs as random effects (boxplots with solid line type) and models where SNPs significantly linked to apricot quality traits were considered as fixed effects (boxplots with dashed line type). Prediction accuracies were computed using a random cross-validation scheme replicated 100 times.

Multi-trait genomic prediction

In order to improve the accuracy of genomic prediction, we assessed the univariate prediction in comparison with multivariate prediction, by leveraging the information on secondary traits that are easy to measure, such as F.weight, Hue.g, RI and TA, in order to predict the ethylene production rate and the apricot fruit content in organic acids (Malic.A and Citric.A) and in soluble sugars (Sucrose, Glucose and Fructose) (Figure 6). Multi-trait prediction using genetic values yielded improvement in PA notably for traits which showed high positive pairwise genetic correlations such as sugars and RI, Citric.A and TA, ethylene and Hue.g. By contrast, model-based index was approximately equivalent to univariate models for non-correlated traits (Figure 6A). Nevertheless, multivariate models using phenotypic information on secondary traits to predict focal traits displayed the lowest overall PA for all the traits under investigation (Figure 6). Prediction performance for ethylene production, organic acids and sugars was significantly improved subsequent to the integration of estimated genetic values of ground color, refractive index and titratable acidity (Figures 6B, 6C and 6D, respectively). The model-based index showed a gain in accuracy which reached 0.25 for Citric.A. informed by TA (Figure 6D), 0.23 for sucrose content informed by RI, 0.17 for glucose informed by RI, and 0.08 for fructose informed by RI (Figure 6C). Nevertheless, the model-based selection index displayed a slight drop in accuracy when genetic correlation between secondary traits and focal traits was low. The decrease ranged from 0.01 to 0.04 for traits where genetic correlation ranged from -0.86 to 0.24.

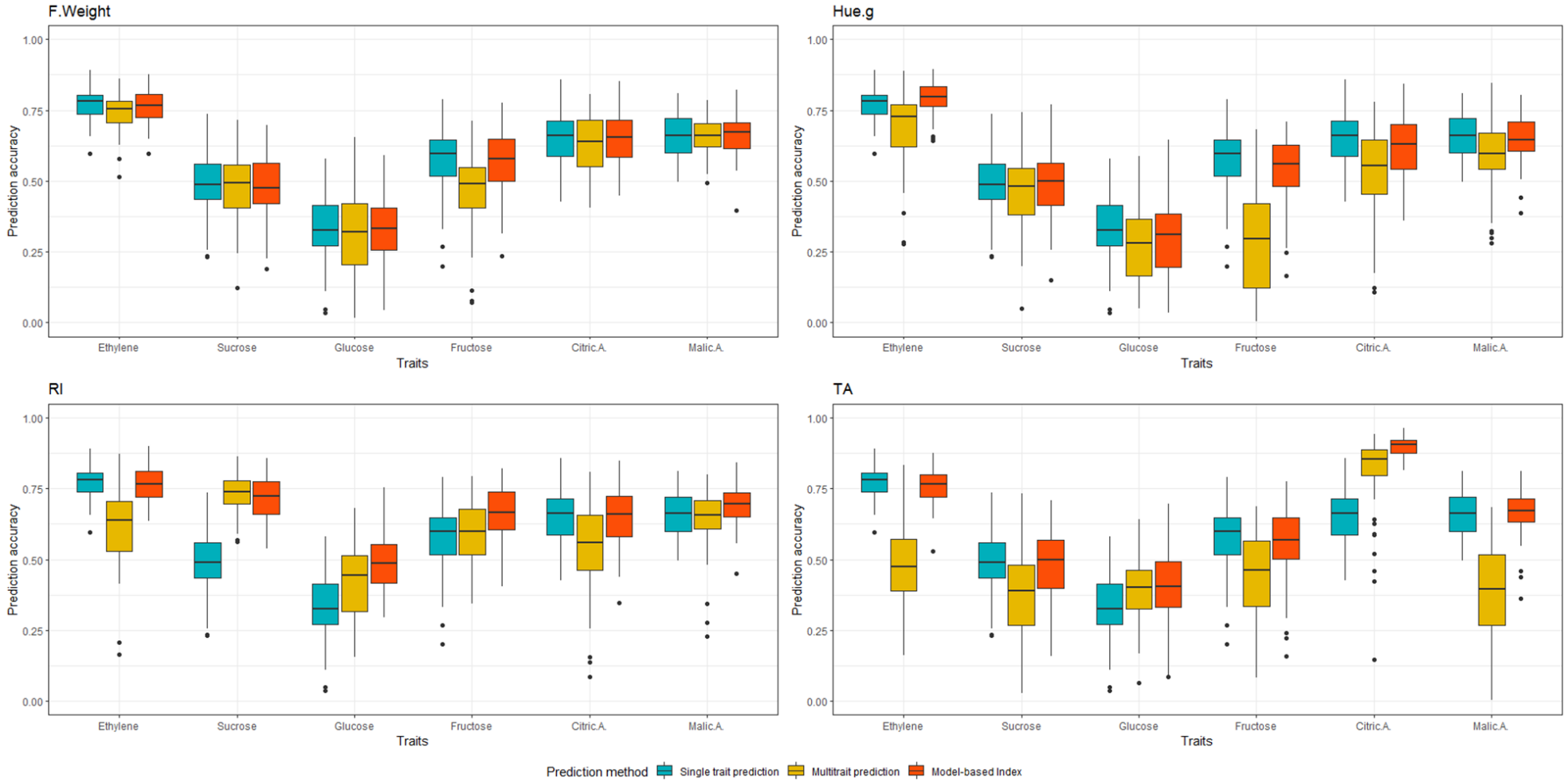


Figure 6 | Assessment of multivariate genomic prediction of ethylene production, content in organic acids (malic acid and citric acid) and soluble sugars (sucrose, glucose and fructose) using F.weight, Hue.g, RI and TA as proxy predictands in comparison to univariate prediction. Multivariate models leveraged the information provided by secondary traits using either phenotypic values or model-based index (estimated genetic values).

DISCUSSION

The objective of our study aimed at investigating GS within a biparental cross between two apricot varieties Goldrich and Moniqui contrasting for fruit traits. The emphasis on GS in apricot breeding stems from the need to deliver genetic progress by facing constraints that hamper or slow down genetic improvement in this species such as the length of juvenility period and the complex genetic architecture of several traits of interest. In this context, our work represents the first evaluation of GS in apricot with a focus on key fruit quality attributes in order to help selection decisions within quality-oriented breeding schemes. Moreover, our study provided new insights into the genetic architecture of fruit quality parameters for fruit weight, ethylene production, and the content in soluble sugars and organic acids. The optimization of genomic PA was also explored by including QTLs significantly linked to targeted traits in prediction models and within a multivariate prediction framework.

Exploration of the phenotypic data

Exploring the phenotypic data revealed that fruit ground color and ethylene production showed skewed distributions, which feeds to the concomitant existence of orange skinned fruits with low ethylene production and white colored fruits with high ethylene outburst reflecting the scope of variation of fruit quality traits within the Go×Mo progeny. Nevertheless, fruit weight and contents in glucose and fructose present symmetric distributions coupled with low standard deviation. A particular variability was found for the sucrose content, the main soluble carbohydrate derived from photosynthesis, which accounts for 60% to 80% of total soluble sugars stored in the apricot fruit. The related variability could result from source-sink relationships. Indeed, carbohydrate translocation from source leaves to 'non-photosynthetic' sink organs, notably fruits, depends largely on agricultural practices such as fruit thinning which tends to increase the availability of photo-assimilates in the remaining fruits (Roussos *et al.* 2011). Therefore, as no thinning was performed in the Go×Mo progeny, this likely resulted in a stronger competition for carbohydrates between sink organs thereby affecting fruit weight and carbohydrate concentration over the two years of observation. In addition, a broad range of variability was mirrored in ground color within the Go×Mo population, which originates from color-contrasted varieties, with a variation magnitude that ranges from deep-orange to white. Indeed, the fruits of Goldrich tend to accumulate colored carotenoids (β-carotene) while those of Moniqui mainly accumulate uncolored compounds (phytoene, phytofluene)(Marty et al. 2005; Adami et al. 2013; Jiang et al. 2019b). Likewise, Goldrich and Moniqui show different climacteric behaviors with regard to ethylene production that ranges from fruits with a high ethylene production rate and fast evolution towards maturity (Moniqui), to fruits characterized by lower ethylene production (Goldrich). With respect to fruit organic acidity, the Go×Mo progeny exhibited a wide range of phenotypic variability for TA (from 11.3 to 41 meq $100g^{-1}$), citric acid (from 8.0 to 45.3 meq $100g^{-1}$) and malic acid (from 4.1 to 17.9 meq $100g^{-1}$). The citric to malic acid ratio ranged from 0.9 to 8.5 with a predominance of citric acid in Go×Mo. These results are in accordance with previous studies expressing large genetic diversity in apricot germplasm (Gurrieri *et al.* 2001; Bassi *et al.* 2006).

It is noteworthy that the genotype effect contributes significantly to the variability of target traits within the Go×Mo population. Therefore, the phenotypic variation linked to apricot fruit quality is largely due to genetic differences and thus the contribution of genetic pattern to the overall phenotypic variance, which is confirmed by estimation of trait heritability. Indeed, except for the sucrose content, all the apricot fruit quality traits were highly heritable. Moreover, the quality attributes varied in response to the fruit physiological stage, the year effect and the genotype × year interaction. Hence, components of phenotypic variance resulted from an interplay of several factors including both genetic and environmental conditions. However, the

phenotypic variation differed between traits. For instance, climatic fluctuations tended to exert minor effect on sucrose, glucose and fructose contents, although genotype by year interaction highly contributes to the expression of sugar-related traits. Contrastingly, acid-related traits (TA, contents in citric and malic acids) were considerably dependent on the year effect. This trend has been confirmed by several studies, which highlighted that a higher acidity pertains to fruits that were produced under cold and wet weather while fruits with low acid flavor are produced under warmer temperatures (Wang and Camp 2000; Gautier *et al.* 2005; Etienne *et al.* 2013).

QTL detection

The genetic architecture underlying the fruit quality attributes has been assessed in order to deepen our knowledge of the genomic regions harboring QTLs for apricot quality. The exploration of genetic determinism of target traits revealed highly stable QTLs across the two years for all traits under consideration, except for sucrose and glucose contents. Several QTLs were identified at the same position on the two parental maps for eight traits out of 10, except for the contents in glucose and citric acid, which suggests the stability of these QTLs over genetic backgrounds. Three QTL clusters were identified: (i) ethylene, fructose and fruit weight on LG1 of Goldrich, (ii) ethylene, TA, citric acid and malic acid on LG2 of Goldrich and (iii) ethylene, RI, fructose, sucrose, on one hand, malic acid and ground color on the other hand on LG2 on Moniqui. QTL clusters might arise from pleiotropic effects of one QTL or the presence of tightly linked QTLs (Eduardo *et al.* 2011). QTLs linked to ethylene production and acid-related traits which clustered on LG2 on Goldrich as well as QTLs clustered for sugar-related traits, organic acids and ethylene on LG2 Moniqui underline overlapping patterns in the expression of apricot quality attributes, notably for ethylene. Indeed, ethylene is a phytohormone interfering in several metabolic pathways and whose biosynthesis is coupled with a respiratory burst. Linked with subsequent ethylene signal and related transcription factors, several physical and biochemical changes occur in fruit maturation such as chlorophyll degradation, carotenoid accumulation and modulation of sugar content, as well as changes in organic acids profiles (Paul *et al.* 2012). Additionally, QTL colocalization between ethylene production and organic acids content in the Go×Mo progeny, confirmed by significant phenotypic correlations, demonstrates that these traits are likely to segregate together so that the fruits whose ethylene production is high, can also produce large amounts of citric acid and low amounts of malic acid. Our results support the causality link between ethylene production and organic acids, which is consistent with several studies postulating that metabolic pools of citric and malic acids are under ethylene regulation (Fan *et al.* 1999; Defilippi *et al.* 2004; Gao *et al.* 2007; Valdés *et al.* 2009; Etienne *et al.* 2013; Batista-Silva *et al.* 2018).

A major and robust QTL was mapped for ground color, located on the LG3 of Goldrich supporting results of (Socquet-Juglard *et al.* 2013). Similarly, different studies on *Prunus* species have also reported that LG3 is associated with skin and flesh pigmentation in conjunction with carotenoid and anthocyanin contents (Frett *et al.* 2014; Salazar *et al.* 2017; García-Gómez *et al.* 2019). Further, a cluster including QTLs for ethylene and ground color was only detected for Moniqui. This is in adequacy with the results of (Marty et al. 2005) according to which the synthesis of colorless carotenoid precursors, phytoene and phytofluene, is upregulated by ethylene, whilst β-carotene, the main colored carotenoid pigment, is ethylene-independent. This trend has been confirmed by a steeper expression of carotenogenic genes in white fruits in comparison to orange ones (Marty *et al.* 2005).

Genomic prediction accuracy

Our investigation of the potential of GS for apricot fruit quality revealed that within-family predictions hold a great promise since accuracies varied across traits within a range from 0.31 to 0.78. Our results are in accordance with (Riedelsheimer et al. 2013) who evidenced the need to train prediction models using full-sibs for the validation set. Hence, cross-validation performed within-family provides richer information than that issued from distant relatives, given the identity-by-descent (IBD) relationships among full-sibs (Legarra *et al.* 2008). Therefore, within-population prediction presents a valuable tool for the implementation of genomic prediction. It allows to achieve higher genomic prediction accuracy than predictions across multiple populations (de Roos *et al.* 2009). Moreover, for a full-sib family, the high level of relatedness and the strong LD between SNPs and causal QTLs underlying the traits under investigation result in high PA. Our results are in accordance with previous studies that pointed out the importance of the inclusion of relatedness in the prediction model (Lenz *et al.* 2017). Interestingly, the highest PA were obtained for ethylene production rate, a trait whose direct measurement is time-consuming, making it unsuitable for high-throughput studies. Conversely, the lowest PA was found for the fruit content in sucrose, glucose and fructose. This might be attributed to the non-genetic part of the observed variation due to environmental factors. Indeed, postharvest performance strongly depends upon various preharvest factors that modulate the source-sink relationships. It is noteworthy that the fruit sweetness has a tendency to increase in response to cultural practices such as fruit thinning. As sucrose is prominent metabolites in the photosynthetic carbon scheme, and as the increase in availability of this carbohydrate is considerably dependent upon fruit load during the secondary fruit growth phase (Roussos *et al.* 2011), the lack of precise source-sink control by thinning strongly impacted the quality of the prediction.

Factors controlling genomic prediction accuracy

Impact of statistical prediction models

Across all prediction models, the average accuracy for apricot quality traits were moderate to high. This trend is in adequacy with the extent of linkage disequilibrium (LD) between SNPs and QTLs. Hence, in a single generation cross, as a limited number of recombination events occurs per meiosis, leading to large linkage blocks and therefore, more phenotypic records per chromosome segment are available in order to derive GEBVs which leads to more accurate predictions (Lorenzana and Bernardo 2009). Within the framework of model comparison, RR-BLUP tended to outperform Bayes A, Bayes B, Bayes C, BL and BRR for 6 traits out of 10. RR-BLUP proved to be the best-performing statistical model notably for traits controlled by several QTLs that explain each a small amount of the phenotypic variance. In addition, RR-BLUP is more efficient with regards its computational speed in comparison to Bayesian models (Tan et al. 2017). Bayes B showed a superior prediction performance compared to RR-BLUP for ground color where a QTL accounts for a large proportion of the phenotypic variation. The aforementioned outcome is in agreement with (Daetwyler et al. 2010a), in which Bayes B gave higher accuracies than GBLUP when the number of QTLs N_{QTL} was low. However, this trend diminished as N_{QTL} increased past the equivalence point where N_{QTL} equates to independently segregating chromosome segments M_e. The deviation from superiority of RR-BLUP for oligogenic traits is likely due to the model over-parameterization as a response to fitting a large number of SNPs to model variation within a trait controlled by few major QTLs (Resende et al. 2012d).

Impact of training population size

Lowering the TP size led to a decay in PA. Accordingly, further improvement on PA could be attained by increasing the total training population size, as a larger reference set provides more acurate predictions due to less biased estimation of marker effects. Furthermore, besides the prominent effect of the size of training set, the design of the reference population in respect of resemblance between training and validation partitions, depicts a potent factor that considerably affects prediction performance. Thereby, closer relationships between training and validation populations has been reported to lead to a higher PA. Conversely, adding genetic diversity within the reference population lead to a reduction in PA compared with smaller training populations including highly related individuals (Lorenz and Smith 2015; Riedelsheimer et al. 2013; Lehermeier et al. 2014). More importantly, highly related individuals share long haplotypes and linkage blocks due to limited recombination events and thereby lead to minor bias while computing GEBVs within the validation set (Lorenzana and Bernardo 2009; Hickey *et al.* 2014; Lozada *et al.* 2019). Hence, higher accuracies linked to richer information issued from closely related individuals rather than distant individuals arise from more precise estimation of marker effects. Therefore, higher degree of IBD sharing between full-sibs is likely to provide accurate estimation of genetic variation for quantitative traits exempt from confounding non-genetic factors in comparison to genetically distant individuals (Visscher *et al.* 2006).

Impact of the number of markers

The number of markers used to train prediction equation represents a prominent factor that affects the prediction performance. Hence, the larger the set of markers the higher the probability to be in LD with QTLs controlling target traits, which provided richer genomic information. This trend has been shown by (de los Campos et al. 2013) who highlighted that the inclusion of all available markers resulted in a considerable increase of the proportion of variance explained. Thereby on a broader scope, the number of SNPs required to obtain accurate predictions depends on the number of independently segregating chromosome segments M_e as well as the span of LD within the study population (Goddard 2008; Daetwyler et al. 2010a). Nevertheless, in the present study, the PA tended to reach a plateau for 6,103 SNPs, indicating that only 10% of the initial markers set were sufficient to capture SNPs-QTLs LD, related to several traits. Our results are consistent with those of (Covarrubias-Pazaran et al. 2018b) who showed that a medium marker density (500 to 750 SNPs) was sufficient to achieve high accuracy due to extensive LD typically present in biparental populations. Similarly, phenotypic records collected in maize biparental populations that were closely related to the selection candidates, only a small number of SNPs (200 – 500) and small number of phenotypes (1,000) were needed to achieve accurate predictions of GEBVs. Otherwise, in more distantly related populations, 10,000 SNPs as well as 5,000 to 20,000 records are needed (Hickey *et al.* 2014). In our study, no accuracy gain was noted beyond 6,103 SNPs. Similarly, (Hickey et al. 2014) reported no benefit in terms of prediction accuracy beyond 10,000 SNPs. Further, in an Eucalyptus breeding population of 949 F1 hybrids, no significant accuracy improvement was obtained using more than 5,000 SNPs to predict growth and wood traits (Tan et al. 2017). In this latter study, only 500 to 1,000 informative, non-redundant and randomly distributed markers were needed to reach sufficient coverage of the genome (Tan et al. 2017). More importantly, our study showed that higher marker density can lead to a reduction in PA whatever the trait. This decrease in accuracy might be attributed to multicollinearity between SNPs due to overfitted prediction models, overestimating the marker effects (Meuwissen et al. 2001a; Muir 2007).

Genomic prediction optimization

Optimization of the GS models

The genomic prediction including QTL mapping outcomes considerably depends on the genetic architecture of the traits under consideration. Hence, this study showed that the magnitude of accuracy gain in prediction was heterogeneous across the traits studied. For instance, for ground color, a significantly higher accuracy was obtained as a result of the inclusion of two QTLs that represent more than 58% of phenotypic variation, compared to models where all markers were fit as random effects. Similar patterns were observed for RI and contents in sucrose and fructose due to the large proportion of genetic variance captured by fixed factors in the prediction models. Hence, accounting for prior genomic information provided a steeper increase in accuracy for traits controlled by major QTLs such as ground color, fructose, sucrose and refractive index than the other fruit quality metrics controlled by several QTLs explaining lower proportion of phenotypic variance. Our findings are in agreement with those of (Morgante et al. 2018), who showed that the integration of a priori information on the genetic architecture underlying quantitative trait variation resulted in a valuable increase in accuracy within samples of unrelated individuals. Similarly, (Zhang et al. 2014) reported that GBLUP informed by the genetic architecture in rice diversity panel an increased the accuracy by 5.4 %. However, when QTLs were modeled as fixed effects in models, a slight decrease in prediction for fruit weight, glucose and TA was observed, since these traits are controlled by QTLs covering only a small proportion of variation. An additional feature of these QTLs was their instability across years.

Multi-trait genomic prediction

Within the multivariate genomic predictions, our findings highlighted that prediction models targeting multiple traits are greatly dependent upon genetic correlations. Thereby, we showed that multivariate models generally provided more accurate predictions compared to univariate models for genetically highly correlated traits. Nevertheless, accuracies showed an equivalent or slight decrease under a low genetic correlation framework. Similar outcomes were reported by several studies. For instance, in bread wheat, (Michel et al. 2017) exploited the availability of easy-to-measure traits such as the protein content, which is genetically highly correlated with costly and labor-intensive traits linked to baking quality in order to breed for superior genotypes. In addition, our results are in accordance with those of (Calus and Veerkamp 2011) where the accuracy gain ranged from 0.03 to 0.14, when genetic correlation of target traits varied from 0.25 to 0.75. However, multivariate models grounded on phenotypic values performed poorly in comparison to model-based index which accounted for estimated genetic values despite phenotypic information. An exception was noted for the prediction of sucrose content informed by RI, where multi-trait model outperformed the two previously cited models, given that genetic correlation between sucrose and RI is very close to 1, so that the residual correlation between these two traits is almost null. Besides their superiority with respect to prediction performance, model-based selection index are computationally less demanding than phenotype-based multivariate models (Michel *et al.* 2017). More importantly, the accuracy gain is more pronounced for slightly heritable traits that are genetically correlated with a highly heritable trait such as sucrose content ($H^2 = 0.31$), which is highly genetically correlated to RI (Jia and Jannink 2012a; Guo et al. 2014; Karaman et al. 2018). Furthermore, the drop-off in PA for traits that are not genetically correlated is attributed to the residual correlation which potentially adds noise to predictions with respect to single trait models and thus leads to biased computation of GEBVs. Our results are in agreement with (Covarrubias-Pazaran et al. 2018b), which showed no benefit over single-trait models under a low genetic correlation framework. Therefore, broad phenotypic information provided at high-throughput

on easy-to-measure traits such as F.weight, Hue.g, RI and TA could offer the opportunity to enlarge the selection candidate population. This potentially enhance the PA for costly and labor-intensive traits. Hence, multivariate prediction grounded on easy-to-phenotype traits might help selection decisions and thus potentially deliver genetic progress notably for perennial species for which the length of breeding cycles is a significant impediment to genetic improvement process. In addition, multivariate predictions might offer a valuable opportunity to predict expensive and difficult to measure traits which is in line with the breeding goals.

CONCLUSION

Our findings highlighted that GS holds a valuable potential with reference to prediction of fruit quality within a biparental design in apricot. Indeed, genomic prediction yielded interesting outcomes in terms of prediction accuracy which encourages further investigations about valuing whole-genome information with the aim of assessing agronomical relevant traits in apricot and potentially orienting breeding strategies towards the implementation of GS within breeding schemes. Furthermore, the outcomes of this study provided insights into the genetic architecture of apricot fruit quality whose integration in prediction models led to a higher PA. As expected, the prediction accuracy gain is higher for the traits that are governed by QTLs explaining a substantial part of phenotypic variation. Besides, genomic predictions might be improved by optimizing factors controlling the predictive performance of GS models such as larger training populations, for example. However, with reference to markers' density, only 6,103 SNPs were enough to reach accurate predictions. In terms of prediction modeling, RR-BLUP outperformed Bayesian models and provided an outstanding compromise between statistical performance and computational time. Moreover, optimal accuracies were obtained under a multivariate prediction framework for fruit quality traits that are strongly and positively correlated to their different proxies, and thus predictions of ethylene content informed by ground color, organic acids by titratable acidity and sugars by refractive index are more accurate than univariate predictions.

In terms of prospects, regarding that the response of phenotypes to genomic prediction is tightly linked to the observed variability within the study population, a greater attention should be paid to orchard management practices through mastering source-sink relationships in order to optimize the potential performance of genotypes and achieve an optimal fruit quality. In addition to that, in the present study, conclusions on the efficiency of GS in apricot were drawn for a biparental design in which genotypes share the same LD pattern and relatedness between training partition and validation partition is high. Therefore, further evidence ought to be assessed in a genetic diversity panel potentially covering a broader range of allelic combinations of traits of agronomical interest.

AUTHORS' CONTRIBUTIONS

JMA conceived and designed the field trials. BG, JMA and SB performed the experiment. TF provided us with the genomic dataset. MN analyzed the data and wrote the manuscript. JLR, JMA and CS guided through the study and reviewed the manuscript. All authors read and approved the manuscript. The authors declare that they have no conflict of interest.

ACKNOWLEDGMENTS

This work was funded by the Ministry of Higher Education and Scientific Research (Tunisia) and was supported by INRAE GAFL. Genotyping costs were supported by FruitSelGen project (2015). We acknowledge the experimental and the scientific teams of UR-GAFL. We would like to express our gratitude to Sebastian Michel for sharing script for multivariate analysis and Brigitte Mangin for her kind assistance and statistical support in multi-trait and optimization

approaches. Also, we would like to thank Eric Duchene for his insightful guidance through the genetic mapping analysis.

LITERATURE CITED

Adami, M., P. de Franceschi, F. Brandi, A. Liverani, D. Giovannini *et al.*, 2013 Identifying a carotenoid cleavage dioxygenase (ccd4) gene controlling yellow/white fruit flesh color of peach. *Plant Molecular Biology Reporter* 31 (5):1166-1175. https://doi.org/10.1007/s11105-013-0628-6

Akdemir, D., W. Beavis, R. Fritsche-Neto, A.K. Singh, and J. Isidro-Sánchez, 2018 Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity* 122:672-683. https://doi.org/10.1038/s41437-018-0147-1

Aranzana, M.J., V. Decroocq, E. Dirlewanger, I. Eduardo, Z.S. Gao *et al.*, 2019 Prunus genetics and applications after de novo genome sequencing: achievements and prospects. *Horticulture Research* 6 (1):58. https://doi.org/10.1038/s41438-019-0140-8

Arumuganathan, K., and E.D. Earle, 1991 Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* 9 (4):208-218. https://doi.org/10.1007/BF026720169

Bartholomé, J., M. Bink, J. Heerwaarden, E. Chancerel, C. Boury *et al.*, 2016 Linkage and association mapping for two major traits used in the maritime pine breeding program: height growth and stem straightness. *PloS one* 11:e0171439. https://doi.org/10.1371/journal.pone.0165323

Bassi, D., F. Bartolozzi, and E. Muzzi, 2006 Patterns and heritability of carboxylic acids and soluble sugars in fruits of apricot (*Prunus armeniaca* L.). *Plant Breeding* 115:67-70. https://doi.org/10.1111/j.1439-0523.1996.tb00873.x

Bates, D., M. Mächler, B. Bolker, and S. Walker, 2014 Fitting linear mixed-effects models using lme4. *ArXiv e-prints* arXiv:1406:1-48. https://doi.org/10.18637/jss.v067.i01

Batista-Silva, W., V.L. Nascimento, D.B. Medeiros, A. Nunes-Nesi, D.M. Ribeiro *et al.*, 2018 Modifications in Organic Acid Profiles During Fruit Development and Ripening: Correlation or Causation? *Frontiers in Plant Science* 9:1689. https://doi.org/10.3389/fpls.2018.01689

Broman, K., H. Wu, S. Sen, and G. Churchill, 2003 R/QTL: QTL mapping in experimental crosses. *Bioinformatics* 19:889-890. https://doi.org/10.1093/bioinformatics/btg112

Bureau, S., D. Ruiz, M. Reich, B. Gouble, D. Bertrand *et al.*, 2009 Rapid and non-destructive analysis of apricot fruit quality using FT-near-infrared spectroscopy. *Food Chemistry* 113:1323-1328. https://doi.org/10.1016/j.foodchem.2008.08.066

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp, 2008 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* 124:362-368. https://doi.org/10.1111/j.1439-0388.2007.00691.x

Calus, M.P.L., and R.F. Veerkamp, 2011 Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution* 43:26. https://doi.org/10.1186/1297-9686-43-26

Chambroy, Y., M. Souty, G. Jacquemin, R.-M. Gomez, and J.-M. Audergon, 1995 Research on the suitability of modified atmosphere packaging for shelf-life and quality improvement of apricot fruit. *Acta Hort.* 384:633-638. https://doi.org/10.17660/ActaHortic.1995.384.99

Covarrubias-Pazaran, G., 2016 Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PloS one* 11:1-15. https://doi.org/10.1371/journal.pone.0156744

Covarrubias-Pazaran, G., B. Schlautman, L. Diaz-Garcia, E. Grygleski, J. Polashock *et al.*, 2018 Multivariate GBLUP Improves Accuracy of Genomic Selection for Yield and Fruit Weight in Biparental Populations of Vaccinium macrocarpon Ait. *Frontiers in Plant Science* 9:1310. https://doi.org/10.3389/fpls.2018.01310

Crossa, J., P. Perez-Rodriguez, J. Cuevas, O. Montesinos-Lopez, D. Jarquin *et al.*, 2017 Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci* 22 (11):961-975. https://doi.org/10.1016/j.tplants.2017.08.011

Daetwyler, H.D., B. Villanueva, and J. Woolliams, 2008 Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PloS One* 3:e3395. https://doi.org/10.1371/journal.pone.0003395

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J. Woolliams, 2010 The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics* 185:1021-1031. https://doi.org/10.1534/genetics.110.116855

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327-345. https://doi.org/10.1534/genetics.112.143313

de Roos, A.P.W., B.J. Hayes, and M.E. Goddard, 2009 Reliability of Genomic Predictions Across Multiple Populations. *Genetics* 183:1545-1553. https://doi.org/10.1534/genetics.109.104935

Defilippi, B.G., A.M. Dandekar, and A.A. Kader, 2004 Impact of suppression of ethylene action or biosynthesis on flavor metabolites in apple (*Malus domestica* Borkh) fruits. *Journal of Agricultural and Food Chemistry* 52:5694-5701. https://doi.org/10.1021/jf049504x

Desta, Z.A., and R. Ortiz, 2014 Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci* 19 (9):592-601. https://doi.org/10.1016/j.tplants.2014.05.006

Eduardo, I., I. Pacheco, G. Chietera, D. Bassi, C. Pozzi *et al.*, 2011 QTL analysis of fruit quality traits in two peach intraspecific populations and importance of maturity date pleiotropic effect. *Tree Genetics & Genomes* 7:323-335. https://doi.org/10.1007/s11295-010-0334-6

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6 (5):e19379. https://doi.org/10.1371/journal.pone.0019379

Endelman, J.B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* 4:250-255. https://doi.org/10.3835/plantgenome2011.08.0024

Etienne, A., M. Genard, P. Lobit, A.M.D. Mbeguie, and C. Bugaud, 2013 What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells. *J Exp Bot* 64 (6):1451-1469. https://doi.org/10.1093/jxb/ert035

Fan, X., M.B. Sylvia, and P.M. James, 1999 1-Methylcyclopropene Inhibits Apple Ripening. *Journal of the American Society for Horticultural Science* 124 (6):690-695. https://doi.org/10.21273/JASHS.124.6.690

Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu *et al.*, 2017 Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genetics Selection Evolution* 49:44. https://doi.org/10.1186/s12711-017-0319-0

Fodor, A., V. Segura, M. Denis, S. Neuenschwander, A. Fournier-Level *et al.*, 2014 Genome-Wide Prediction Methods in Highly Diverse and Heterozygous Species: Proof-of-Concept through Simulation in Grapevine. *PloS One* 9:e110436. https://doi.org/10.1371/journal.pone.0110436

Frett, T.J., G.L. Reighard, W.R. Okie, and K. Gasic, 2014 Mapping quantitative trait loci associated with blush in peach [*Prunus persica* (L.) Batsch]. *Tree Genetics & Genomes* 10:367-381. https://doi.org/10.1007/s11295-013-0692-y

Gao, H.Y., B.Z. Zhu, H.L. Zhu, Y.L. Zhang, Y.H. Xie *et al.*, 2007 Effect of suppression of ethylene biosynthesis on flavor products in tomato fruits. *Russ. J. Plant Physiol.* 54:80-88. https://doi.org/10.1134/S1021443707010128

García-Gómez, B.E., J.A. Salazar, L. Dondini, P. Martínez-Gomez, and D. Ruiz, 2019 Identification of QTLs linked to fruit quality traits in apricot (*Prunus armeniaca* L.) and biological validation through gene expression analysis using qPCR. *Molecular Breeding* 39. https://doi.org/10.1007/s11032-018-0926-7

Gatti, E., B.G. Defilippi, S. Predieri, and R. Infante, 2009 Apricot (*Prunus armeniaca* L.) quality and breeding perspectives. *Journal of Food, Agriculture and Environment* 7:573-580.

Gautier, H., A. Rocci, M. Buret, D. Grasselly, and M. Causse, 2005 Fruit load or fruit position alters response to temperature and subsequent cherry tomato quality. *Journal of the Science of Food and Agriculture* 85:1009-1016. https://doi.org/10.1002/jsfa.2060

Gianola, D., M. Pérez-Enciso, and M.A. Toro, 2003 On marker-assisted prediction of genetic value: Beyond the ridge. *Genetics* 163:347-365.

Goddard, M., 2008 Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136:245-257. https://doi.org/10.1007/s10709-008-9308-0

Goddard, M.E., and B.J. Hayes, 2007 Genomic selection. *J Anim Breed Genet* 124 (6):323-330. https://doi.org/10.1111/j.1439-0388.2007.00702.x

Grattapaglia, D., and M.D.V. Resende, 2011 Genomic selection in forest tree breeding. *Tree Genetics & Genomes* 7:241-255. https://doi.org/10.1007/s11295-010-0328-4

Grattapaglia, D., and R. Sederoff, 1994 Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* 137 (4):1121-1137.

Guo, G., F. Zhao, Y. Wang, Y. Zhang, L. Du *et al.*, 2014 Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics* 15:30. https://doi.org/10.1186/1471-2156-15-30

Gurrieri, F., J.-M. Audergon, G. Albagnac, and M. Reich, 2001 Soluble sugars and carboxylic acids in ripe apricot fruit as parameters for distinguishing different cultivars. *Euphytica* 117:183-189. https://doi.org/10.1023/A:1026595528044

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink, 2012 Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science* 56:146-160. https://doi.org/10.2135/cropsci2011.06.0297

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu *et al.*, 2014 Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science* 54 (4):1476-1488. https://doi.org/10.2135/cropsci2013.03.0195

Infante, R., P. Martínez-Gomez, and S. Predieri, 2008 Quality oriented fruit breeding: Peach [Prunus persica (L.) Batsch]. *Journal of Food, Agriculture and Environment* 6:342-356.

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2014 Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics* 128:145–158. https://doi.org/10.1007/s00122-014-2418-4

Jia, Y., and J.-L. Jannink, 2012 Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. *Genetics* 192:1513–1522. https://doi.org/10.1534/genetics.112.144246

Jiang, F., J. Zhang, S. Wang, L. Yang, L. Yingfeng *et al.*, 2019 The apricot (Prunus armeniaca L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis. *Horticulture Research* 6:128. https://doi.org/10.1038/s41438-019-0215-6

Karaman, E.M.S., M.T. Lund, M. Anche, L. Janss, and G. Su, 2018 Genomic prediction using multi-trait weighted GBLUP accounting for heterogeneous variances and covariances across the genome. *Genes Genomes Genetics* 8:3549-3558. https://doi.org/10.1534/g3.118.200673

Kumar, S., D. Chagné, M.C.A.M. Bink, R.K. Volz, C. Whitworth *et al.*, 2012 Genomic selection for fruit quality traits in apple (*Malus* × *domestica* Borkh.). *PloS one* 7:e36674. https://doi.org/10.1371/journal.pone.0036674

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen, 2008 Performance of Genomic Selection in Mice. *Genetics* 180:611-618. https://doi.org/10.1534/genetics.108.088575

Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan *et al.*, 2014 Usefulness of Multiparental Populations of Maize (Zea mays L.) for Genome-Based Prediction. *Genetics* 198:3-16. https://doi.org/10.1534/genetics.114.161943

Lenz, P.R.N., J. Beaulieu, S.D. Mansfield, S. Clément, M. Desponts *et al.*, 2017 Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (*Picea mariana*). *BMC Genomics* 18:335. https://doi.org/10.1186/s12864-017-3715-5

Liu, H., H. Zhou, Y. Wu, X. Li, J. Zhao *et al.*, 2015 The Impact of Genetic Relationship and Linkage Disequilibrium on Genomic Selection. *PloS One* 10:e0132379. https://doi.org/10.1371/journal.pone.0132379

Liu, X., H. Wang, H. Xiaojiao, K. Li, Z. Liu *et al.*, 2019 Improving Genomic Selection With Quantitative Trait Loci and Nonadditive Effects Revealed by Empirical Evidence in Maize. *Frontiers in Plant Science* 10:1129. https://doi.org/10.3389/fpls.2019.01129

Lopes, M.S., H. Bovenhuis, M. van Son, Ø. Nordbø, E. Grindflek *et al.*, 2017 Using markers with large effect in genetic and genomic predictions. *Journal of Animal Science* 95:59-71. https://doi.org/10.2527/jas.2016.0754

Lorenz, A.J., and K.P. Smith, 2015 Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. *Crop Science* 55:2657. https://doi.org/10.2135/cropsci2014.12.0827

Lorenzana, R.E., and R. Bernardo, 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics* 120:151-161. https://doi.org/10.1007/s00122-009-1166-3

Lozada, D.N., R.E. Mason, J.M. Sarinelli, and G. Brown-Guedira, 2019 Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC Genetics* 20:82. https://doi.org/10.1186/s12863-019-0785-1

Makowsky, R., N.M. Pajewski, Y.C. Klimentidis, A.I. Vazquez, C.W. Duarte *et al.*, 2011 Beyond Missing Heritability: Prediction of Complex Traits. *PLoS Genetics* 7:e1002051. https://doi.org/10.1371/journal.pgen.1002051

Marty, I., S. Bureau, G. Sarkissian, B. Gouble, J.-M. Audergon *et al.*, 2005 Ethylene regulation of carotenoid accumulation and carotenogenic gene expression in colour-contrasted apricot varieties (*Prunus armeniaca*). *Journal of Experimental Botany* 56:1877-1886. https://doi.org/10.1093/jxb/eri177

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20 (9):1297-1303. https://doi.org/10.1101/gr.107524.110

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (4):1819-1829.

Michel, S., C. Kummer, M. Gallee, J. Hellinger, C. Ametz *et al.*, 2017 Improving the baking quality of bread wheat by genomic selection in early generations. *Theoretical and Applied Genetics* 131:1-17. https://doi.org/10.1007/s00122-017-2998-x

Minamikawa, M., K. Nonaka, E. Kaminuma, H. Kajiya-Kanegae, A. Onogi *et al.*, 2017 Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits. *Scientific Reports* 7:1–13. https://doi.org/10.1038/s41598-017-05100-x

Morgante, F., W. Huang, C. Maltecca, and T.F.C. Mackay, 2018 Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. *Heredity* 120:500–514. https://doi.org/10.1038/s41437-017-0043-0

Muir, W., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics* 124:342-355. https://doi.org/10.1111/j.1439-0388.2007.00700.x

Munoz-Sanz, J.V., E. Zuriaga, I. Lopez, M.L. Badenes, and C. Romero, 2017 Self-(in)compatibility in apricot germplasm is controlled by two major loci, S and M. *BMC Plant Biol* 17 (1):82. https://doi.org/10.1186/s12870-017-1027-1

Muranty, H., M. Troggio, B.S. Ines, M. Rifaï, A. Auwerkerken *et al.*, 2015 Accuracy and responses of genomic selection on key traits in apple breeding. *Horticulture Research* 2:15060. https://doi.org/10.1038/hortres.2015.60

Obenchain, V., M. Lawrence, V. Carey, S. Gogarten, P. Shannon *et al.*, 2014 VariantAnnotation: A Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* 30:2076-2078. https://doi.org/10.1093/bioinformatics/btu168

Onogi, A., 2020 Connecting mathematical models to genomes: Joint estimation of model parameters and genome-wide marker effects on these parameters. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa129

Paul, V., R. Pandey, and G.C. Srivastava, 2012 The fading distinctions between classical patterns of ripening in climacteric and non-climacteric fruit and the ubiquity of ethylene-An overview. *Journal of Food Science and Technology* 49:1-21. https://doi.org/10.1007/s13197-011-0293-4

Pérez, P., and G. de los Campos, 2014 Genome-Wide Regression & Prediction with the BGLR Statistical Package. *Genetics* 198:483–495. https://doi.org/10.1534/genetics.114.164442

Ramstein, G.P., S.E. Jensen, and E.S. Buckler, 2019 Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theoretical and Applied Genetics* 132:559–567. https://doi.org/10.1007/s00122-018-3267-3

Rana, M., A. Sood, W. Hussain, R. Kaldate, T. Sharma *et al.*, 2019 Gene Pyramiding and Multiple Character Breeding, pp. 83-124 in *Lentils: Potential Resources for Enhancing Genetic Gains*. edited by M. Singh. Academic Press, London.

Resende, M.D.V., M.F.R. Resende, C.P. Sansaloni, C.D. Petroli, A.A. Missiaggia *et al.*, 2012a Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist* 194:116-128. https://doi.org/10.1111/j.1469-8137.2011.04038.x

Resende, M.F.R., P. Munoz, M.D.V. Resende, D.J. Garrick, R.L. Fernando *et al.*, 2012b Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (*Pinus taeda* L.). *Genetics* 190:1503-1510. https://doi.org/10.1534/genetics.111.137026

Riedelsheimer, C., J.B. Endelman, M. Stange, M.E. Sorrells, J.-L. Jannink *et al.*, 2013 Genomic Predictability of Interconnected Biparental Maize Populations. *Genetics* 194:493-503. https://doi.org/10.1534/genetics.113.150227

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel *et al.*, 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.Rincent, R., A. Charcosset, and L. Moreau, 2017 Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics* 130:2231–2247. https://doi.org/10.1007/s00122-017-2956-7
https://doi.org/10.1534/genetics.112.141473

Roussos, P.A., V. Sefferou, N.-K. Denaxa, E. Tsantili, and V. Stathis, 2011 Apricot (*Prunus armeniaca* L.) fruit quality attributes and phytochemicals under different crop load. *Scientia Horticulturae* 129:472-478. https://doi.org/10.1016/j.scienta.2011.04.021

Ruiz, D., P. Lambert, J.-M. Audergon, L. Dondini, S. Tartarini *et al.*, 2010 Identification of QTLs for fruit quality traits in apricot. *Acta Horticulturae* 862:587-592. https://doi.org/10.17660/ActaHortic.2010.862.93

Salazar, J.A., I. Pacheco, P. Shinya, P. Zapata, C. Silva *et al.*, 2017 Genotyping by Sequencing for SNP-Based Linkage Analysis and Identification of QTLs Linked to Fruit Quality Traits in Japanese Plum (*Prunus salicina* Lindl.). *Frontiers in Plant Science* 8:476. https://doi.org/10.3389/fpls.2017.00476

Socquet-Juglard, D., D. Christen, G. Devènes, C. Gessler, B. Duffy *et al.*, 2013 Mapping Architectural, Phenological, and Fruit Quality QTLs in Apricot. *Plant Molecular Biology Reporter* 31:387-397. https://doi.org/10.1007/s11105-012-0511-x

Solberg, T.R., A.K. Sonesson, J.A. Woolliams, and T.H. Meuwissen, 2009 Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution* 41:29. https://doi.org/10.1186/1297-9686-41-29

Spindel, J.E., H. Begum, D. Akdemir, B. Collard, E. Redoña *et al.*, 2016 Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116:395–408. https://doi.org/10.1038/hdy.2015.113

Tan, B., D. Grattapaglia, G.S. Martins, K.Z. Ferreira, B. Sundberg *et al.*, 2017 Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biology* 17:110. https://doi.org/10.1186/s12870-017-1059-6

Taylor, J., and D. Butler, 2017 R Package ASMap: Efficient Genetic Linkage Map Construction and Diagnosis. *Journal of Statistical Software* 79:1-29. https://doi.org/10.18637/jss.v079.i06

Testolin, R., 2011 Kiwifruit breeding: From the phenotypic analysis of parents to the genomic estimation of their breeding value (GEBV). *Acta Horticulturae* 913:123-130. https://doi.org/10.17660/ActaHortic.2011.913.14

Valdés, H., M. Pizarro, R. Campos-Vargas, R. Infante, and B.G. Defilippi, 2009 Effect of Ethylene Inhibitors on Quality Attributes of Apricot cv. Modesto and Patterson during Storage. *Chilean Journal of Agricultural Research* 69:134-144. https://doi.org/10.4067/S0718-58392009000200002

Verde, I., A.G. Abbott, S. Scalabrin, S. Jung, S. Shu *et al.*, 2013 The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics* 45 (5):487-494. https://doi.org/10.1038/ng.2586

Visscher, P.M., S.E. Medland, M.A.R. Ferreira, K.I. Morley, G. Zhu *et al.*, 2006 Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLoS Genetics* 2:e41. https://doi.org/10.1371/journal.pgen.0020041

Voorrips, R.E., 2002 MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *The Journal of Heredity* 93:77-78. https://doi.org/10.1093/jhered/93.1.77

Wang, S.Y., and M.J. Camp, 2000 Temperatures after bloom affect plant growth and fruit quality of strawberry. *Scientia Horticulturae* 85:183-199. https://doi.org/10.1016/S0304-4238(99)00143-0

Whittaker, J.C., R. Thompson, and M.C. Denham, 2000 Marker-assisted selection using ridge regression. *Genetical Research* 75:249-252. https://doi.org/10.1111/j.1469-1809.1999.ahg634_0351_17.x

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus, 2013 The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics* 193:621-631. https://doi.org/10.1534/genetics.112.146290

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao *et al.*, 2014 Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. *PloS One* 9:e93017. https://doi.org/10.1371/journal.pone.0093017

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J.-L. Jannink, 2009 Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. *Genetics* 182:355-364. https://doi.org/10.1534/genetics.108.098277

Chapitre 4 : Évaluation de la sélection phénomique dans une descendance biparentale

4.1. Présentation du chapitre

Dans ce chapitre, nous avons évalué comparativement les performances de la SP et de la SG dans la population biparentale Go×Mo, de 153 individus phénotypés pour la qualité des fruits et la phénologie, génotypés avec 61 030 SNPs et caractérisés deux années consécutives par 2 307 spectres couvrant la gamme de nombres d'onde allant de 12 493 à 3 498 cm^{-1} obtenus sur les deux faces de 4 fruits intacts dans le PIR et 1 736 spectres MIR compris entre 3 996 et 700 cm^{-1} et obtenus sur des homogénats de pulpe de 4 fruits.

Dans un premier temps, nous avons montré l'incidence du prétraitement des spectres sur la précision de prédiction phénomique. Différents algorithmes de prétraitement incluant la normalisation, la dérivation et le detrend ont été appliqués sur les données spectrales afin d'éliminer le biais apporté par la déviation de la ligne de base. Les modèles de SP ont présenté une large gamme de réponses en fonction du caractère ciblé et de la bande spectrale évaluée. En revanche, les spectres prétraités ont présenté de meilleures performances par rapport aux spectres bruts, avec un prétraitement optimal qui dépend du caractère cible.

4.2. Précision des modèles de sélection phénomique

Le partitionnement de la variance spectrale a montré une forte contribution du facteur génotypique à la variation notamment dans le MIR. La capacité des spectres à capturer la variance génétique a été confirmée par les valeurs de l'héritabilité au sens large H^2 calculées pour les différents nombres d'onde le long des spectres PIR et MIR allant de 0 à 0.52 pour le PIR et de 0.26 à 0.88 pour le MIR. Ceci s'est traduit par une précision de prédiction moyenne allant de 0.13 à 0.97.

Il est à noter que l'interaction Génotype – Année ($G \times A$) contribue fortement à la variance spectrale. De même, la contribution cumulée de G et de G×A varie de 42.6 à 95.5 % pour le PIR et de 52 à 92.7 % pour le MIR, respectivement. En d'autres termes, la variance spectrale due à G + G×A a atteint plus de 50 % pour tous les nombres d'onde dans la région MIR, tandis que dans la région PIR, 68 % des nombres d'onde présentaient un pourcentage cumulé supérieur à 50 %. Ceci souligne le potentiel de prédiction inhérent aux modèles de sélection phénomique en termes de précision de prédiction de la composante liée à l'interaction.

La comparaison entre les modèles de SP et de SG a révélé que les modèles de prédiction fondés sur les spectres étaient plus performants que les modèles fondés sur les SNPs pour huit

caractères sur 12. Les performances de prédiction ont varié fortement en fonction du caractère et du domaine spectral. La précision moyenne a varié de 0,13 ± 0,12 pour la couleur à 0,94 ± 0,02 pour l'IR pour les modèles de basés sur le PIR et de 0,17 ± 0,14 pour la couleur à 0,97 ± 0,01 pour l'IR et l'AT pour les modèles basés sur MIR. De même, pour le modèle de SG, la précision moyenne allait de 0,31 ± 0,10 pour la date de floraison à 0,77 ± 0,05 pour la production d'éthylène.

La figure 25 montre un exemple de deux caractères liés à la qualité des fruits et deux caractères phénologiques indépendants de la composition biochimique des échantillons analysés. Comme indiqué dans la figure, la performance prédictive des modèles de SP est supérieure à celle du modèle de SG pour les caractères liés à la constitution biochimique notamment ceux basés sur les spectres MIR sur pulpe de fruit. De même, les modèles de SP se sont révélés supérieurs pour les caractères liés à la phénologie.

Se basant sur l'inférence de la proximité génétique entre les individus de la descendance Go×Mo, la SP s'est montrée aussi performante que la SG laissant suggérer qu'elle est capable de capturer le DL ainsi que l'échantillonnage mendélien, piliers de la précision des modèles de SG.

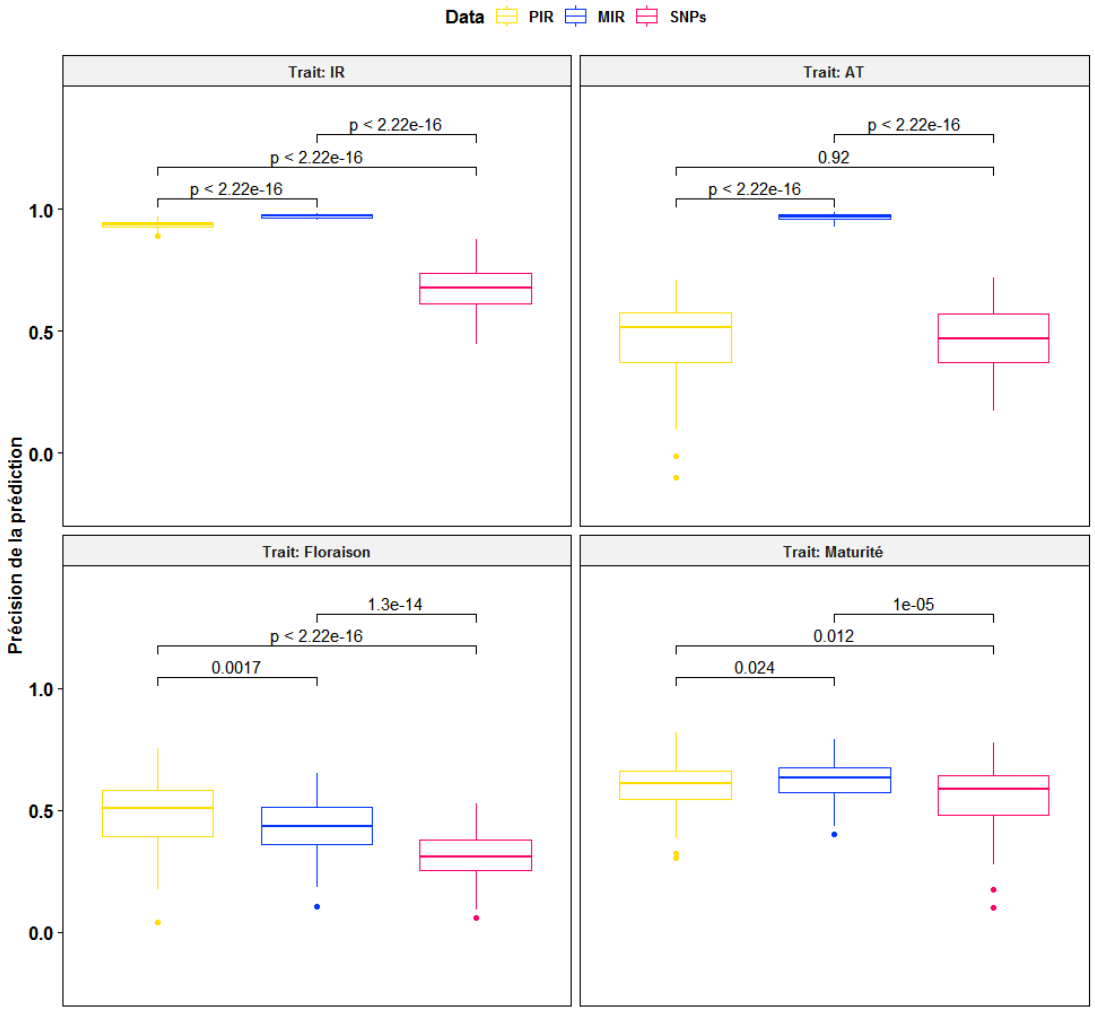


Figure 25: Précision de prédiction des modèles de sélection phénomique comparée à la précision du modèle de sélection génomique

4.3. Optimisation de la précision de la sélection phénomique

Afin d'optimiser la performance prédictive des modèles de SP, nous avons évalué le couplage entre des informations a priori sur l'architecture génétique des caractères et l'information spectrale. Ce cadre d'optimisation a permis d'améliorer la précision pour tous les caractères modélisés par le Bayes B. Quant à RR-BLUP, les modèles PIR- et MIR-BLUP ont permis d'améliorer la précision pour sept et six caractères sur 12, respectivement. En outre, le gain en précision, dérivé des modèles de pondération, a augmenté avec la variance génétique représentée par les QTLs. Par exemple, le gain moyen en précision est passé de 0.26 à 0.28 pour la production d'éthylène et de 0.61 à 0.70 pour la couleur de fond, traits pour lesquels l'architecture génétique est déterminée par des QTLs majeurs représentant plus de 40 % de la variance phénotypique (Figure 26).

La réponse de la précision de la SP à la modélisation des QTL en covariables est donc fortement influencée par la part de variance expliquée par ces QTLs et par le type modèle de prédiction.

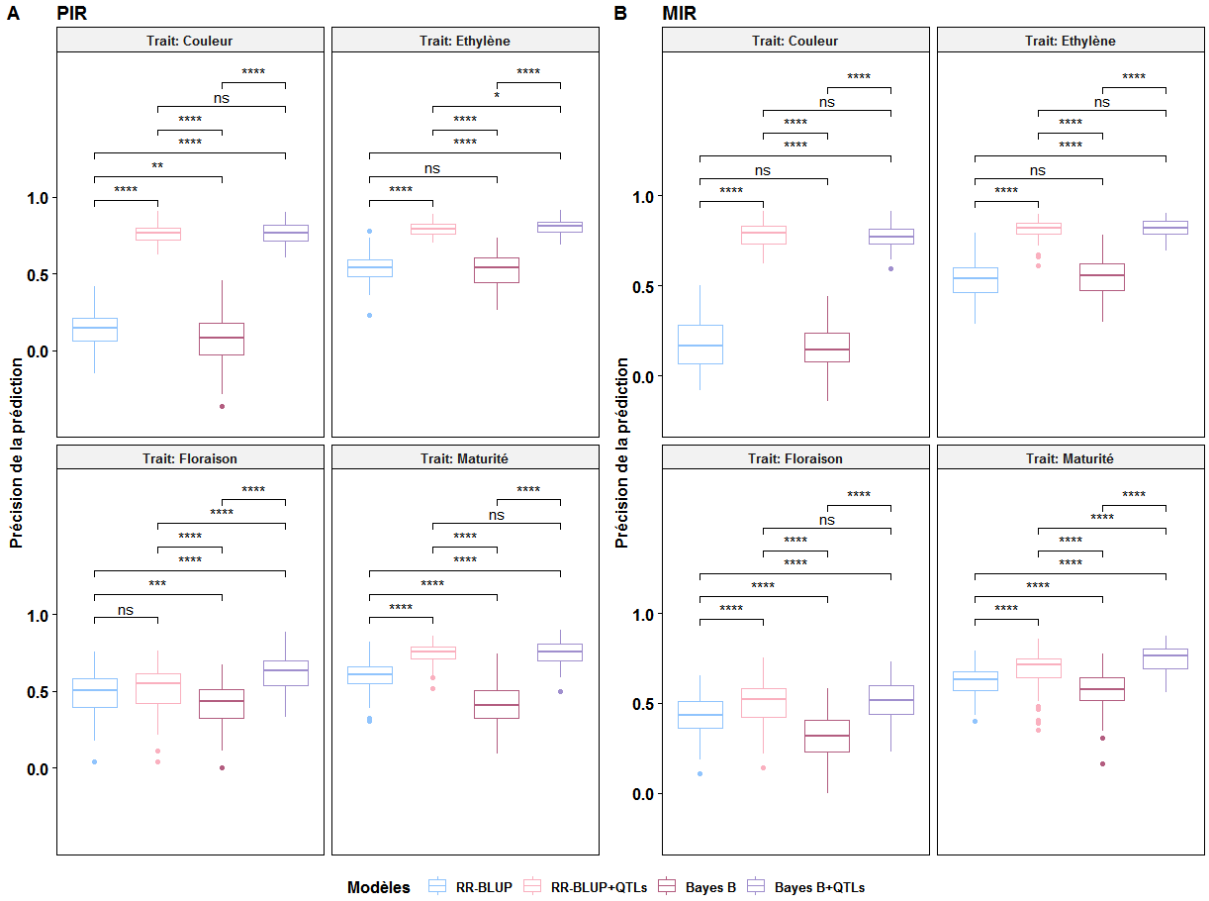


Figure 26: Précision de la prédiction des modèles phénomiques optimisés comparée à la précision des modèles de référence

4.4. Conclusion

En conclusion, les données phénomiques s'avèrent être plus informatives que les données génomiques pour les caractères étroitement liés à la composition biochimique du fruit. Les mesures de réflectance dans l'infrarouge sont considérées comme des phénotypes intermédiaires permettant la prédiction de phénotypes d'intérêt agronomique. Par conséquent, les phénotypes spectraux sont capables non seulement de capturer l'expression des séquences génomiques mais aussi des interactions entre ces régions chromosomiques et les conditions environnementales. Il s'agit de caractères complexes intégrant différents types d'information correspondants aux différents niveaux d'expression du génome à l'échelle moléculaire. Ceci est en cohérence avec maintes études ayant confirmé la supériorité de la valeur prédictive des endophénotypes par rapport à celle des marqueurs moléculaires. Il est également à noter que

les modèles mobilisant les spectres MIR présentent une meilleure performance prédictive par rapport à ceux utilisant les spectres PIR, notamment pour les caractères biochimiques.

Assessment of phenomic selection accuracy for fruit quality and phenology in Apricot

Mariem Nsibi*, Barbara Gouble †, Sylvie Bureau†, Christopher Sauvage*,#, Jean-Marc Audergon*, [1], Jean-Luc Regnard§

* INRAE, Génétique et Amélioration des Fruits et Légumes, 84143 Montfavet Cedex, France,

† INRAE, Avignon University, UMR SQPOV, 84914 Avignon, France,

Syngenta SAS France, 1228 Chemin de l'Hobit, 31790 Saint Sauveur, France,

§ AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France

ORCID IDs: 0000-0002-4175-2431 (M.N.); 0000-0003-1269-7733 (B.G.); 0000-0002-9335-5091 (S.B.); 0000-0001-5466-9955 (C.S.); 0000-0002-3132-5815 (J.-M.A.); 0000-0001-8614-0618 (J.-L.R.)

Abstract

Despite the technological breakthroughs that have fostered the success of genomic selection within crop breeding programs, the acquisition of molecular marker and phenotypic information at high-throughput scale remains problematic given the costs of genotyping strategies in addition to the laborious implementation of phenotyping and selection procedures. These challenges gave rise to a potential alternative to genomic selection: a cost-effective and high-throughput approach called phenomic selection, based on endophenotypes potentially revealed by spectral data. Phenomic selection in apricot was assessed on a pseudo-F1 biparental population of 153 individuals and revealed that spectral information derived from near- and mid-infrared spectroscopy (NIRS and MIRS) captured a wide spectrum of heritable variation for traits of interest. The preprocessing of spectral data helped towards the optimization of phenomic selection accuracy. Yet, pretreatment algorithms yielded variable outcomes with reference to prediction accuracy according to the 12 traits under investigation (fruit quality and tree phenology). Moreover, our findings shed light on the optimization of phenomic selection models by means of valuing the information on trait genetic architecture. Indeed, upweighting QTLs that underpin variation of target traits within prediction models allowed to improve their predictive performance, notably for traits driven by major QTLs, such as fruit ethylene production and ground color. Correspondingly, these traits yielded higher accuracy gain

compared to traits displaying more complex genetic architecture, ranging from 0.26 to 0.28 and from 0.61 to 0.7, respectively, depending on the prediction model. Globally, we provide additional proof that phenomic selection is of outstanding interest to capture additional heritable variance compared to markers in order to sustain genetic gain.

INTRODUCTION

Prevailing challenges linked to the pace of population growth and the steadily degrading biodiversity and natural resources have emphasized the urgency of implementing efforts aligned with the growing concerns of achieving food security and promoting sustainable development (Misra 2014; Banik 2019). Such challenging context highlights the urgency of reducing costs and enhancing throughput of information sources leveraged in plant breeding programs. Although high-throughput genomic technologies rendered possible acquisition of increasingly available genetic make-up of selection candidates, genotyping costs are in decline but could remain unaffordable, notably for crop species that lack genomic resources such as SNP arrays and high-quality reference genomes. Hence, the emphasis has shifted to alternative tools that permit acquisition of high-throughput and high-dimensional datasets to unravel molecular variation. A potential alternative is the information brought by transcripts and metabolites derived from omics technologies. In this regard, omics resources provide a powerful tool to shape breeding designs. However, despite the ability of omics data to elucidate genetic variants that contribute to the agronomic performance at a molecular scale, leveraging such data at a large scale could be problematic. Alternatively, phenomic information driven by spectroscopic technologies provides an indirect prediction of key traits, stemming from phenomena beyond the genetic control (Bureau et al. 2009c, 2009b; Ruiz et al. 2008). The adoption of spectroscopy was achieved decades ago in breeding stations and is attributable to its ease of implementation, its cost-effectiveness and computational affordability. It generates high-dimensional source of information, across different geographical sites and seasons that allows to capture polymorphisms associated with endophenotypes (transcriptome, proteome and metabolome that modulate the expression of phenotypes) across hierarchical levels of biological knowledge.

Within this context, Rincent *et al*. (2018) put forward a novel breeding strategy, called Phenomic Selection (PS), in order to partially circumvent the shortcomings of genomic selection with reference to the expenses of genotyping technologies and the inability to capture genotype by environment (G×E) interactions. This proof-of-concept study provided a

framework for predicting unknown phenotypes using large-scale, non-destructive and cost-efficient near-infrared spectroscopy (NIRS). Besides, phenome-oriented selection strategies have attracted considerable interest given their ability to uncover the link between phenotypic variation and the genetic make-up of selection candidates, thus bridging the gap between phenome and genome. Spectral knowledge permits to forecast crop characteristics and pinpoint high-performing individuals among selection candidates. Indeed, a breadth of studies has been focusing in phenomics to screen crop attributes with regards to their physico-chemical features, mainly linked to endophenotypes and genetics. Conversely, the drive behind phenomic selection stems from its ease of use, cost-effectiveness and importantly its ability to predict phenotypes independently from the biochemical composition of the biological samples. Particularly in apricot, this provides a valuable breeding framework for a range of agronomically relevant traits with emphasis on resistance to biotic and abiotic stresses, adaptation to climate change, phenology and more importantly fruit quality in order to meet the expectations of apricot sector stakeholders with reference to consumer preferences. Although PS is still in its infancy, it has shown its potential in predicting accurately phenotypic performance that are independent from analyzed tissues within contrasted environment in an association population of poplar and a panel of elite winter wheat (Rincent et al. 2018b). Similarly, (Hayes et al. 2017a) reported enhancement of prediction models performance that incorporate NIR and nuclear magnetic resonance‐derived (NMR) phenotypes for end-use quality traits in wheat. Moreover, (Lane et al. 2020) evidenced that PS was instrumental in predicting maize grain yield within a diversity panel and highlighted the promise of NIR reflectance spectroscopy in assessing genetically independent breeding panels.

It is noteworthy that deployment of phenomics within multi-environment field trials might significantly sustain genetic gain through the reduction of costs and labor intensity associated with key phenotypes, assessed at an early developmental stage (Krause *et al.* 2019). Moreover, phenomic records are likely to bring insights into similarities between selection candidates without the need for genotyping, which represents a significant hindrance within breeding programs. Similarities between candidates are inferred from their reflectance profiles that capture a wide spectrum of phenotypic variation resulting from genetic mutation and environmental impact (Gegas *et al.* 2014). Therefore, phenomic-based predictions are grounded on genetic relatedness captured by spectral data. Besides capturing covariance between genotypes, phenomic selection attempts to capture genotype-by-environment interactions. Therefore, novel insights can be gained from an experiment performed in one environment to

predict traits in any other environment through computing kinship matrix or using NIR reflectances relative to environment 1 as regressors in the prediction model (Rincent et al. 2018b). Within this context, (Krause et al. 2019) showed that relationship matrix, derived from high-dimensional spectral data and modeled in order to predict economically relevant phenotypes, yielded equivalent or superior prediction accuracies in comparison to models that incorporate only molecular marker information and pedigree-derived relationship matrices. The afore-mentioned studies revealed the superiority of NIRS-based models as well as multi-trait models based on joint predictions of target traits and NIR-derived phenotypes, as compared to SNPs-based models. This highlights the valuable potential of NIRS to achieve breeding goals and supports the resources shifting from genomics to phenomics. In this sense, we aimed at assessing the potential of spectral data, as input for PS, in predicting apricot fruit quality and tree phenology.

Our study revolves around the following objectives: (1) to evaluate the potential value of phenomic selection in comparison with genomic selection to predict traits linked to the biochemical constitution of the fruit samples and tree phenology traits, (2) to assess the effect of the spectral data pretreatment on the phenomic selection prediction accuracy and (3) to optimize the phenomic selection models by upweighting QTLs related to the traits under investigation within randomly drawn training partitions (Nsibi *et al.* 2020).

MATERIALS AND METHODS

Plant material

The plant material comprised 153 pseudo-F1 progenies derived from a cross between two apricot cultivars ‘Goldrich’ and ‘Moniqui’. This progeny, hereafter referred as Go×Mo, is characterized by contrasting patterns of genetic variation notably for traits linked to fruit quality and phenology (Salazar *et al.* 2016; Nsibi *et al.* 2020). The experimental orchard was planted in 2004 at Bellegarde in southern France (INRAE, L’Amarine). The trial design consisted of randomly distributed hybrids, observed on their own roots, with blocs of parent lines localized within each row and arranged across the whole experimental design by diagonalization. Borders were protecting the core-experimental field. Conventional integrated management practices were applied all over the experimentation in the hybrid orchard.

Phenotyping for apricot fruit quality

The evaluation of apricot fruit quality was performed over two consecutive years (2006 and 2007). A set of 40 fruits per genotype were harvested and assessed according to fruit color and firmness. Representative physical and biochemical quality metrics were evaluated using reference methods used in routine at SQPOV research unit (INRAE, Avignon). Fruit firmness was assessed through the determination of pressure (kPa) required to achieve 3% deformation of fruit height with a multipurpose texture analyzer (Pénélaup, Serisud, Montpellier, France). Firmness was used to sort fruits into three homogeneous maturity lots (of 4 fruit each) representative of the variability within harvest lots: commercial maturity stage with pressure from 130 to 80 kPa, half-ripe stage from 80 to 50 kPa and mature fruits with firmness less than 50 kPa. Fruit color was measured using a CR-400 chromameter (Minolta, Osaka, Japan) and expressed, according to color space coordinates a* and b*, through computation of Hue angle.

Ethylene production rate (nmol $kg^{-1}h^{-1}$) was determined by analyzing a headspace aliquot by gas chromatography after confinement in hermetically closed jars (see Chambroy *et al.*, 1995 for detailed protocol).

The biochemical characterization of fruit samples was performed on purees. Soluble solid content was determined by the index of refraction (RI) using a digital refractometer (PR-101 ATAGO, Norfolk, VA) and expressed in % Brix at 20°C. Titratable acidity was measured by neutralization up to pH 8.1 with 0.1 N NaOH and expressed in meq 100 g^{-1} of fresh weight using an autotitrator (Methrom, Herisau, Switzerland). Sugars (glucose, fructose and sucrose) and organic acids (citric acid and malic acid) were measured using enzymatic methods using kits for chemical food analysis (Boehringer Mannheim Co., Mannheim, Germany) and expressed in g 100 g^{-1} of fresh weight and meq 100 g^{-1} of fresh weight, respectively.

Phenotyping for phenological traits

The phenology was evaluated for three years (2006, 2007 and 2009) with reference to two key phenological traits: flowering date and maturity date expressed in Julian days. Flowering date was determined when 50% of flower buds had reached the flowering stages (BBCH60, 65, 69). As for maturity date (BBCH87), it was recorded according to the softening of fruit flesh at physiological maturity (Dirlewanger *et al.* 2012).

Genotyping data

The Go×Mo progeny was genotyped using the high throughput genotyping by sequencing (GBS) technique according to Elshire *et al.*'s protocol (2011) within the framework of

FruitSelGen project (2015). Genotyping using the NGS technology Illumina HiSeq2000 was performed at the genotyping platform hosted at the AGAP research unit (INRAE, Montpellier). Digestion of DNA and ligation with adapters was performed using the ApeKI restriction endonuclease. Then, single-end sequencing of 96-plex libraries was carried out on a Genome Analyzer II (Illumina, Inc., San Diego, CA) (100-bp reads). The quality of reads was checked using GATK software. High stringency filters were applied to the reads and low-quality sequences were discarded. Given the non-availability of a complete genome sequence for *Prunus armeniaca*, the DNA sequence alignment (individual paired end reads of 100-bp) was performed using the peach reference genome (Peach v1.0) (Verde *et al.* 2013) via the Burrows-Wheeler alignment tool (BWA) (Li and Durbin 2009) using standard parameters and covering 34.27% of the genome length. Thereafter, variant calling identified a total of 2,565,573 SNPs with an average coverage depth of 19.33×. The missing genotyping data threshold was set at 5% and SNPs with a genotype quality lower than 20 were declared as missing. The outcome of filtering process consisted of 61,030 SNPs. Missing SNP genotypes were imputed with the mean value for each marker.

Acquisition of spectral data

Phenomic records were acquired in NIRS on intact fruits and in MIRS on fruit puree homogenates. For NIRS, diffuse reflectance corresponding to each spectrum was recorded by an average of 32 scans and carried out using a multi-purpose analyzer spectrometer (Bruker Optics®, Wissembourg, France) on both the blushed and the unblushed sides of each fruit (four fruits per genotype for each of the three maturity lots) (Bureau *et al.* 2009a). For MIRS, attenuated total reflectance (ATR) spectra were collected at room temperature using a Tensor 27 FTIR spectrometer (Bruker Optics, Wissembourg, France) equipped with ATR zinc selenide (ATR-ZnSe) crystal (dimensions of 6 cm × 1 cm and six internal reflections) and a deuterated triglycine sulphate (DTGS) detector (Bureau *et al.*, 2009b). The samples of apricot slurries (one sample per genotype for each maturity lot) were placed in the crystal for spectrum acquisition and background spectra corresponding to the 32 recorded scans were collected once every 20 samples. NIR spectra were registered at wavenumbers ranging from 12,500 to 4,000 with 4 cm^{-1} wavelength increment, while MIR spectra were recorded across a range of 4,000 to 700 cm^{-1} at a 2 cm^{-1} spectral resolution.

Spectra investigation and pretreatment

The exploration of spectral data was performed via principal component analysis (PCA) by plotting phenomic records corresponding to MIR and NIR spectra in order to assess the spectral variation and detect potential outliers. Preprocessing of spectral data was performed on raw spectra prior to the prediction analysis in order to preclude spectral noise derived from measurement errors and identify the appropriate preprocessing technique for the downstream analysis. In this scheme, data implemented in PS were subjected to several chemometric pretreatment protocols, performed using 'signal' (signal developers, 2014) and 'prospectr' (Stevens and Ramirez-Lopez 2020) R packages to compute the derivation and detrend, respectively. Detrend pretreatment consisted on performing a standard normal variate (SNV) transformation followed by fitting a second order linear model to correct for wavelength-dependent scattering effects (Barnes *et al.* 1989)). Derivatives were computed using Savitzky–Golay filter (Savitzky and Golay 1964) with a smoothing window length of 37 data points (74 cm^{-1}) and polynomial order n=2. Thereafter, the mean reflectance across lots and years generated 2,307 averaged wavenumbers in NIRS region of the spectrum and 1,736 wavenumbers in MIRS region. The preprocessed data as well as raw spectra were subjected to the inspection of the accuracy of phenomic selection in order to explore the effect of data pretreatment on prediction performance.

Partition of spectral variance and heritability estimation

In order to explore the landscape of spectral variation with regards to the contribution of the genetic factor, as well as interaction between genotypes and years, components of spectral variance were partitioned using the following mixed model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + \alpha\beta_{ij} + e_{ijk} \quad (1)$$

where y_{ijk} denotes the reflectance value corresponding to the genotype i for the year j and the maturity group k at a given wavelength, α_i is the random effect of the genotype i, β_j refers to the fixed effect of the year j, δ_k is the fixed effect of the maturity group k, $\alpha\beta_{ij}$ is the interaction effect of the genotype i and the year j and e_{ijk} is the residual effect.

Within the mixed modeling framework, broad-sense heritability of spectral data was computed as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_y^2}{n_y} + \frac{\sigma_l^2}{n_l} + \frac{\sigma_{gy}^2}{n_y} + \frac{\sigma_e^2}{n_y * n_l}} \quad (2)$$

where σ_g^2 is the genetic variance, σ_l^2 denotes the variance attributed to the maturity stage l, σ_y^2 denotes the variance attributed to the year of measurement y, σ_{gy}^2 is the variance attributed to the interaction between genotype and year and σ_e^2 is the residual variance, n_y refers to the number of years and n_l is the number of fruit lots.

Phenomic and genomic selection modeling

Predictive modeling of target phenotypes using infrared spectral data was performed through random cross-validation with 100 replicates, adopted to investigate the accuracy of prediction. Within this cross-validation scheme, phenotypes of the individuals belonging to validation partition were set to zero then predicted using the remaining individuals in the training partition. For each partition, 75% of records were assigned to training prediction model and the remaining 25% were assigned to validation. This prediction scheme was implemented in order to predict performance of supposedly unphenotyped individuals. To determine the performance of GS and PS models, R Pearson's correlation coefficient between observed phenotypes and the corresponding predicted ones was estimated for each iteration and each fold.

Prediction accuracy (PA) of the ten apricot fruit quality traits and the two tree phenological traits was performed using NIRS data and MIRS data separately. In addition, wavelengths from the NIR and MIR ranges were used concomitantly in order to evaluate the accuracy of phenomic selection using a combination of wavebands in NIR and MIR regions. Assessment of phenomic selection was performed in comparison with genomic selection using RR-BLUP model provided in 'rrBLUP' R package (Endelman 2011b).

Optimization of phenomic selection accuracy

The optimization strategy grounded on prior knowledge on the genetic architecture of target traits, consisted on incorporating SNPs statistically linked to the phenotypic variation in phenomic selection models. QTL detection was performed via composite interval mapping strategy within randomly drawn training partitions with 100 runs. The linkage analysis was carried out using the R/qtl package (Broman et al. 2003b) according to a pseudo-testcross mapping approach (Grattapaglia *et al.* 1995) using the two parental genetic maps (Nsibi *et al.* 2020). The logarithm of odds (LOD) threshold was determined using a permutation test of 1000 replicates. SNPs with a significant LOD score ($p < 0.01$) were declared as QTLs.

In order to investigate prediction performance of phenomic-based models, including SNPs tightly linked to QTLs as covariates, PA was computed in comparison to that of phenomic selection models where only spectral data were used. Herein, two modelling approaches were performed, in contrast to the upstream analysis: RR-BLUP and Bayes B, as their assumptions regarding the distribution of marker effects and consequently their response to weighting QTLs tend to diverge. Indeed, RR-BLUP overlooks trait genetic architecture by attributing equal variance to all SNPs, whilst Bayes B is more prone to marker variance and tends to outperform RR-BLUP for traits with lower QTL number.

Data availability

The phenotypic, spectral and genotypic data are available in Files S1-6. File S1 contains the raw phenotypic data. File S2 contains broad-sense heritability estimates. The genotypic data are available in File S3. Files S4 and S5 provide the output of the decomposition of spectral variance in NIR and in MIR, respectively. Files S6-8 enclose the outcome of the spectral pre-treatment.

RESULTS

Partition of spectral variance and heritability estimation

Partition of spectral variance, outlined in Figure 1, showed that the genotype G and genotype by year interaction G×Y represented the main range of variation of NIR and MIR spectra. In comparison to NIR, spectral variance due to G and G×Y was consistently higher along the MIR spectrum. The genetic component G accounted for a rather low variation in NIR spectrum compared to MIRS. In NIRS it varied from zero to 37.2%, and for wavenumber ranges below 7,240 cm-1, genetic variance was lower than 10%. Contrastingly, a substantial fraction of variation within MIR signal was attributed to the genetic component G, ranging from 16% to 58%, with 156 wavenumbers (9.4%) displaying a percentage of G above 50%. As for the G×Y component, 989 wavenumbers (42.9%) showed a contribution of more than 50% for the NIR region and only 98 wavenumbers (6%) for the MIR region. Besides, the genotypic effect G was higher than the G × Y interaction across 55% of NIRS wavenumbers and 36.8% of MIRS wavenumbers. Additionally, the cumulative proportion of G and G×Y varied from 42.6 to 95.5% for NIRS and from 52 to 92.7% for MIRS, respectively. The capacity of spectra to capture genetic variance was further confirmed by computing the broad-sense heritability (H^2) along the NIR and MIR spectra. The H^2 values, reported in File S2, ranged from zero to 0.52

for NIRS and from 0.26 to 0.88 for MIRS. Namely, spectral variance due to G + G×Y reached more than 50% for all wavenumbers in the MIR region, whereas in the NIR region, 68% of the wavenumbers displayed a cumulative percentage higher than 50%.

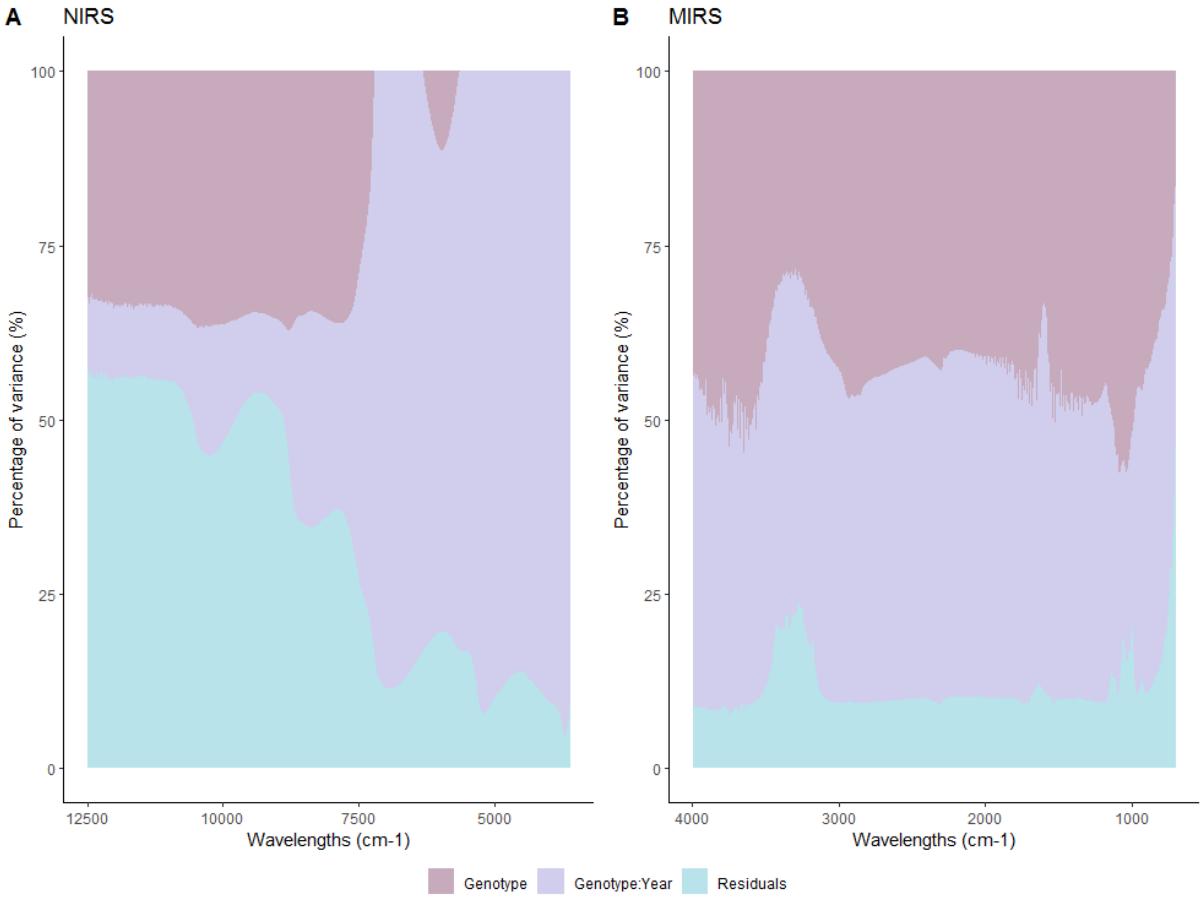


Figure 1 | Partition of spectral variance into genotype component (G) and genotype × year interaction (G×Y) attributed to near (A) and mid-infrared (B) spectra spanning the ranges 12.500 cm-1 to 4.000 cm-1 and 4.000 cm-1 to 700 cm-1, respectively.

Effect of spectral pretreatment on phenomic selection accuracy

The spectral pretreatment algorithms, performed to remove spurious signals due to uncontrolled spectral distortions, displayed a wide range of responses according to the targeted trait and the spectral band. The response of PA to the preprocessing techniques is presented in Files S7-8, provided in the supplemental materials.

The pretreatment of spectra led to a consistent baseline compared to raw spectra. Yet, spectral data transformation greatly influenced PA. The raw reflectance spectra depicted the largest

variation in accuracy. For instance, the accuracy range of MIR raw spectra reached 1.1 for TA and 0.8 for citric acid. Likewise, some pretreatment algorithms resulted in a broad range of variation within PA such as the de-trending pretreatment of TA in NIRS, for which accuracy range reached 1.26. Normalized spectrum provided the highest PA for six traits out of 12 in MIRS and for two traits out of 12 for NIRS. The first derivative displayed the best-performing pretreatment for F.weight (NIRS) and Hue.g (MIRS). For ethylene production, normalization coupled to detrending in NIRS and to derivative in MIRS depicted the highest accuracy. Correspondingly, de-trending pretreatment, either coupled to normalization or derivatives, outperformed all pretreatments for eight traits out of 12 in NIRS. By contrast, detrending the MIR signal was the optimal preprocessing algorithm for only three traits out of 12. Considering these findings and given that PA is trait-specific, the most efficient pretreatment technique in terms of PA was for each trait for the downstream analysis.

Assessment of phenomic selection accuracy

The comparison between phenomic selection and genomic selection (Figure 2) revealed that prediction models grounded on NIRS or MIRS spectra outperformed SNPs-based models for eight traits out of 12. Prediction performance varied greatly depending on the trait and the infrared signal. The average PA for 100 iterations of random cross-validation ranged from 0.13 ± 0.12 for Hue.g to 0.94 ± 0.02 for RI for phenomic selection models based on NIRS and from 0.17 ± 0.14 for Hue.g to 0.97 ± 0.01 for both RI and TA for models based on MIRS. Correspondingly, for genomic selection model, the average accuracy ranged from 0.31 ± 0.10 for flowering date to 0.77 ± 0.05 for ethylene production.

MIR-BLUP models exhibited a higher predictive performance than SNPs-based models for seven traits out of 12, notably for traits linked to fruit biochemical composition, with the exception of malic acid content and ethylene production. Moreover, results indicated a significant difference in average accuracy between the phenomic selection and genomic selection, except for fructose, TA, malic acid and maturity date, for which p-values were 0.78 (NIRS), 0.92 (NIRS), 0.12 (MIRS) and 0.012 (NIRS), respectively. Similarly, NIRS-based models and MIRS-based models performed equally well for Hue.g, Ethylene and flowering date with 0.015, 0.84 and 0.0017 as p-values. However, spectral data acquired in MIR region achieved a higher PA in comparison to NIRS data for all traits except for these three traits.

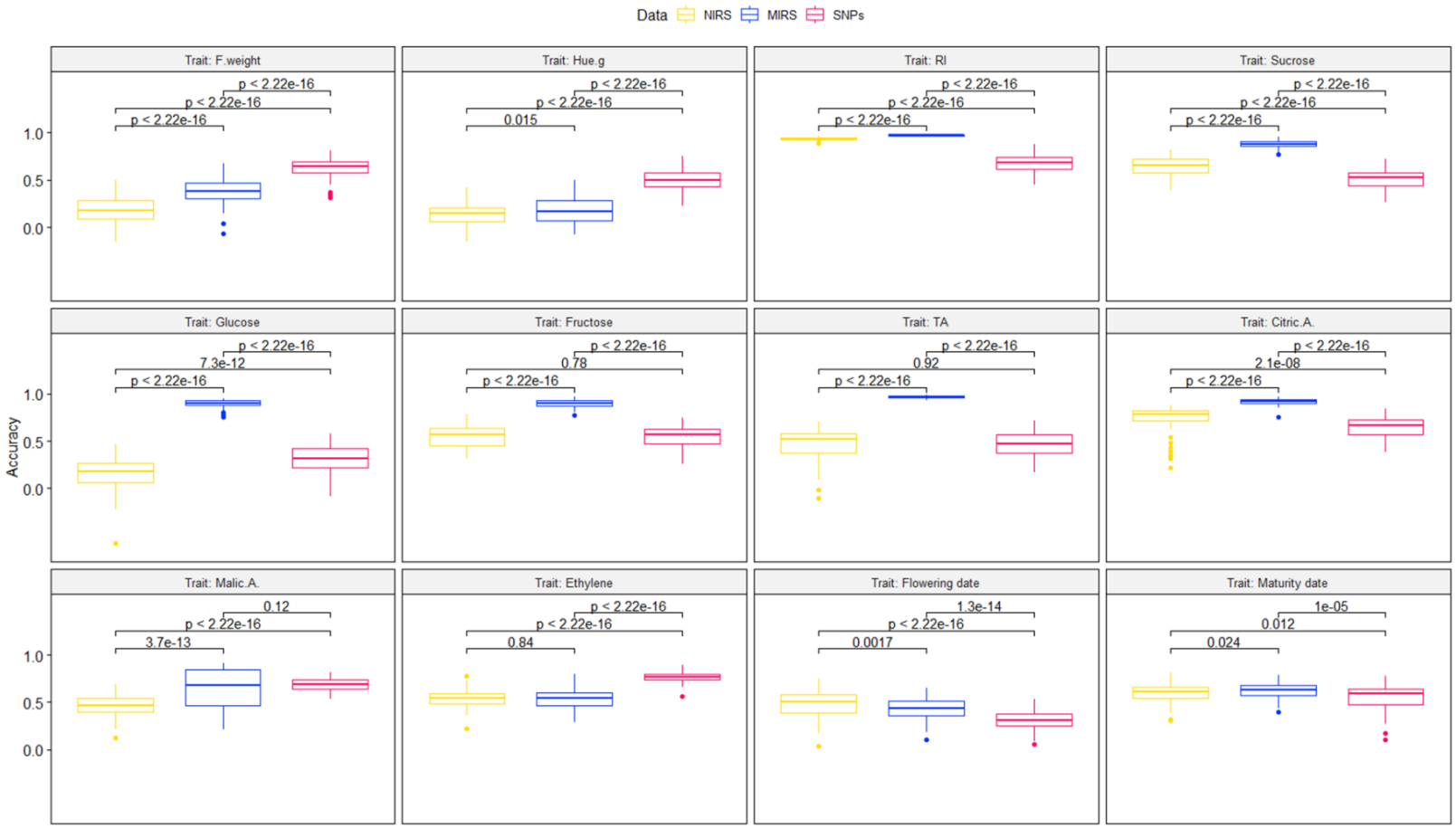


Figure 2 | Prediction performance of phenomic selection in comparison to genomic selection assessed through computation of Pearson's correlation between observed phenotypes and predicted ones using phenomics-based models (NIRS and MIRS) and marker-based models (SNPs).

Optimization of phenomic selection accuracy

Optimization of predictive performance of phenomic selection models was grounded on the exploitation of prior knowledge on trait genetic architecture, provided in Nsibi *et al.* (2020). As outlined in Figure 3, response of PA for phenomic selection models to the inclusion of QTLs as covariates was highly dependent on the investigated trait. The adoption of NIRS wavelengths along with the genomic information improved RR-BLUP model performance for seven traits out of 12. Likewise, MIR-BLUP model with QTLs defined as covariates, provided a higher PA than models without QTLs for six traits out of 12. For the Bayes B model, a remarkable gain in accuracy was consistently found for all traits using both NIRS- and MIRS-based models, including genetic information of QTLs.

Hence, this optimization strategy resulted in a predictive accuracy gain ranging from 0.03 for tree flowering date to 0.63 for Hue.g and from 0.06 for maturity date to 0.61 for Hue.g for NIRS- and MIRS-BLUP models, respectively. For Bayes the B model, the improvement in PA varied from 0.11 for F.weight to 0.70 for Hue.g and from 0.17 for maturity date to 0.63 for Hue.g for NIRS- and MIRS-based models.

The magnitude of accuracy gain was steeper for traits driven by major QTLs such as Hue.g and ethylene production for which, average accuracy was up 0.63 and 0.26 for NIRS and up 0.61 and 0.28 for MIRS, respectively. Regarding the predictive accuracy gain derived from Bayesian prediction of supposedly unknown phenotypes, accuracy improvement was up 0.70 and 0.28 for ethylene and Hue.g, respectively within NIR range and up 0.63 and 0.27 within MIR range.

For the two phenological traits, PA improvements were consistently marginal on average compared to the reference RR-BLUP model based on reflectance of NIRS and MIRS wavenumbers. In addition, incorporating QTLs in spectra-BLUP models resulted in a decrease in predictive accuracy for citric acid, TA, sucrose, fructose and RI. Concerning Bayes B model, PA was consistently higher than that of RR-BLUP model for most traits with the exception of F.weight, Hue.g and Malic.A, for which only marginal differences were recorded between BLUP and Bayesian models, ranging from 0.01 to 0.02.

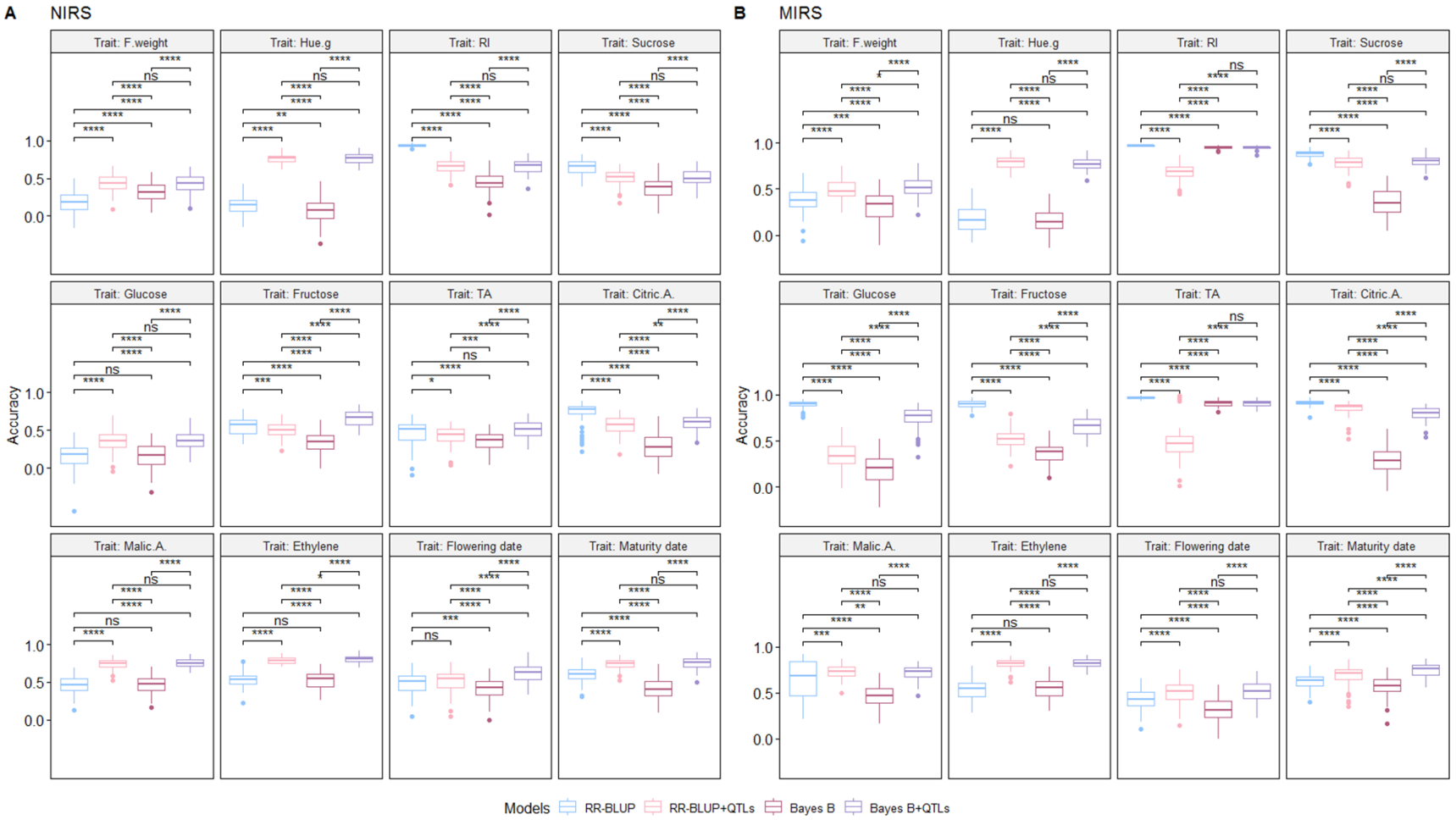


Figure 3 | Assessment of prediction performance of models including QTLs as covariates in comparison to spectra-based models.

(A) Near-infrared-based models
(B) Mid-infrared-based models

DISCUSSION

Our study aimed at investigating the potential of PS to predict 10 fruit quality and two phenological traits and assessing the relevance of further implementation of PS in the apricot. The PS concept lies on prediction of traits of interest through computation of covariance between individuals derived from spectral signature of analyzed samples. Crucially, phenomic selection is potentially relevant for traits that are expressed at a late developmental stage, an advantage of a paramount interest notably for crop species with long breeding cycles.

Partition of spectral variance and heritability estimation

Broad-sense heritability estimates revealed that spectral variation is moderately to highly heritable, except for some wavebands ranging from 7,201 cm-1 to 9,195 cm-1, which exhibited rather low heritability. Indeed, more than 64% of wavenumbers spanning the NIR region depicted heritability estimates lower than 0.4. Conversely, a high level of heritability was found notably in the MIR range coupled with a high cumulative proportion of G + G×Y. This implies that variation within spectra is due to genetics and thus IR spectrum especially MIR spectrum includes a paramount proportion of information potentially valuable for breeding and likely to provide guidance on G×Y interactions to aid selection for stable genotypes across environments.

In accordance with broad-sense heritability estimates, dissection of spectral variability revealed that apricot fruit quality, flowering date and fruit maturity date are under the control of moderate to high number of loci. In addition to the genotypic factor, G×Y interactions were proved to contribute to the observed spectral variability. Therefore, a substantial partition of spectral variation was attributed to the genetic contribution, which highlights the potential of NIRS and, to a greater extent, that of MIRS to capture Mendelian sampling and instruct the inheritance patterns. Our results are in accordance with previous results that reported that phenomics enable to assess genetic variation within relevant traits (Montes *et al.* 2007; White *et al.* 2012; Araus and Cairns 2014; Gebreselassie *et al.* 2017). Therefore, this offers a framework to infer the genetic individual relatedness and the underlying linkage disequilibrium, two major concepts genomic selection is based on.

Effect of spectral pretreatment on phenomic selection accuracy

Mathematical transformation of spectra is performed to cope with interferences originating from light scattering and extract effective information. Hence, removing artifacts prior to

spectra modeling is mandatory for downstream analysis. Besides, spectral preprocessing allows removing irrelevant variance due to background signals and redundancies due to overlapping signatures and thus enhancing prediction performance (Rinnan *et al.* 2009; Barbin *et al.* 2014).

Regardless of the trait under consideration, preprocessed spectra outperformed raw reflectance spectra. Conversely, the effect of spectral pretreatment on PA of phenomics-based models tend to vary according to the trait and the pretreatment algorithm. For instance, Savitsky-Golay algorithm is destined to smooth the spectrum in order to reduce signal-to-noise ratio (Savitzky and Golay 1964). Baseline correction is performed using derivative methods: First order derivative eliminates only the baseline, whereas second derivative discards both baseline and linear trend (Rinnan *et al.* 2009). Although derivation removes irrelevant baseline signals and increases spectral resolution, derivative algorithms are susceptible to spectral noise (Wang and Zhou 2011; Pizarro et al. 2004). Accordingly, first and second derivatives are only recommended when the extent of spectral interferences is low as derivation emphasizes noise and eliminates a proportion of the valuable information (Buddenbaum and Steffens 2012). Regarding spectral normalization, this scatter corrective technique aims to achieve normal distribution of spectra with unit variance. Normalization overlooks the least squares fitting, so that this technique is prone to noisy entries within the spectrum (Rinnan *et al.* 2009). Regarding detrending correction, it is worth noting that the detrended signal accounts for variation derived from curvilinearity and baseline shift (Barnes *et al.* 1989). Moreover, several studies that investigated the scope of influence of preprocessing on PA, confirmed that coupling multiple ranges of spectral pretreatment techniques is of a great interest as compared to untransformed spectra. Therefore, in an attempt to enhance the prediction performance of phenomics-based models, we investigated the optimal spectral pretreatment as well as combinations of several mathematical transformations. In this regard, coupling multiple pretreatments greatly improved PA for all traits under investigation except for Hue.g. Our findings are in accordance with a myriad of studies that reported a reduction of bias and an enhancement of PA (Pizarro *et al.* 2004). However, preprocessing might introduce noise information so that coupling several pretreatment techniques amplifies noisy signals and thus influences performance of the predictive model (Barnes *et al.* 1989; Rinnan *et al.* 2009). Additionally, as evidenced by (Huang *et al.* 2020), spectral pretreatment might be associated with overcorrection, which leads to inaccurate predictions.

Overall, pretreated NIR and MIR spectra both exhibited higher average and lower variation of PA compared to non-preprocessed spectra. Conversely, response of PA to the preprocessing of

spectral data is dependent upon the considered trait. On one hand, considering a given pretreatment specific to each trait can provide a valuable optimization framework. On the other hand, the discrepancies derived from pretreatment, mirrored in a high accuracy range and a high standard deviation in comparison to average accuracy, the amplification of spectral noise or loss of spectral information (Wang and Zhou 2011). Further, additional parameters are worth exploring such as window length and order for Savitzky-Golay filter. Their incidence on pretreatment performance was previously investigated in literature (Geladi and Dåbakk 1995; Nicolaï *et al.* 2007; Gautam *et al.* 2015). However, their effect on phenomic selection accuracy prompts further investigation.

Assessment of phenomic selection accuracy

In the present study, the Prediction performance of phenomic selection and genomic selection varied across traits. Genomic prediction was slightly more accurate than phenomic prediction for physical properties (F.weight and Hue.g) and ethylene production. Moreover, despite marginal differences between NIRS and MIRS for four traits out of 12, MIRS led to a higher PA for all fruit chemical composition. Although comparison between NIRS and MIRS is beyond the scope of our article, it has been reviewed in several studies that NIRS has been standing behind MIR spectroscopy due to the scarcity of relevant chemical information. Indeed, the superiority of MIRS is attributed to fundamental absorption bands. Nonetheless, NIRS is grounded on overtones and combinations of fundamental absorption bands derived from vibrations and rotational transitions of chemical bonds in MIR range (Blanco and Villarroya 2002; Porep *et al.* 2015; Bureau *et al.* 2019). Additionally, overtones and combinations are broad and overlapped giving rise to a wide range of multicollinearity (Ozaki *et al.* 2006; Porep *et al.* 2015). Therefore, MIRS is more reliable than NIRS for identification of organic compounds and thus to predict traits related to the biochemical composition of the samples. Hence, in contrast to NIR, MIR peaks are assigned to major chemical compounds of apricot fruit such as sugars and organic acids, as demonstrated in (Bureau et al. 2009b), which is in alignment with structural investigation of biological matrices. In addition, diffuse reflectance of NIRS region originates from superficial layers of the intact fruit, whereas fruit quality is rather appraised from mesocarp (Walsh *et al.* 2020). The assessment performed in MIR region offers the opportunity to inspect biological matrices representative of the whole sample and thereby quantify individual compounds within homogenized materials (Bureau *et al.* 2019).

Furthermore, NIR spectroscopy is prone to interferences due to water absorption bands. Indeed, given that water presents the prevalent constituent of fruits and vegetables, spectral wavebands recorded in NIR are dominated by the water peaks, which penalizes the potential of prediction performance of chemical attributes that are present in relatively low concentration (Nicolaï *et al.* 2007; Cozzolino *et al.* 2011; Porep *et al.* 2015). Besides that MIR is recommended to predict desired traits linked to chemical constitution of biological samples, MIRS performed better than NIRS to predict their physical properties notably fruit weight and fruit ground color.

On the other hand, regardless of the biochemical composition of investigated tissues, phenomics-based models outperformed SNP-based model and non-significant differences were noted between NIRS and MIRS for phenological traits. Correspondingly, phenomic prediction is valuable for traits that are independent from fruit chemical constitution.

Although several studies have shown the significant contribution of spectra to unveil chemical properties of analyzed samples, spectral reflectance might inform about traits regardless of biochemical background.

Optimization of phenomic selection accuracy

Optimization of PS models by means of leveraging prior information on trait genetic architecture allowed for accuracy improvement for all traits modeled by Bayes B model. As for RR-BLUP model, NIRS- and MIRS-BLUP models provided improved accuracies for seven and six traits out of 12, respectively. In addition, accuracy gain, derived from weighting models, increased with genetic variance accounted for by QTLs. For instance, the average gain in PA ranged from 0.26 to 0.28 and from 0.61 to 0.70 for ethylene production and ground color, respectively, for which genetic architecture is shaped by major QTLs accounting for higher than 40% of phenotypic variance (Nsibi *et al.* 2020). Therefore, response of phenomic selection accuracy to modelling QTLs as fixed-effect covariates is influenced by the shape of the genetic contribution to target traits and the prediction model.

It is noteworthy that prediction performance was consistently higher for Bayes B compared to RR-BLUP. Therefore, model response to weighting QTLs is tightly dependent upon the gap between model assumption and the trait genetic architecture. This was evidenced by multiple studies that of Bayes B outperforms BLUP based model for traits controlled by large-effect QTLs (Daetwyler et al. 2013; Hayes et al. 2010; Zhang et al. 2014). In addition, (Ren et al. 2020) investigated several traits with a broad range of genetic architecture and proved that

Bayes B, assuming variable selection and shrinkage priors, as well as GWAS methods were more valuable, in terms of PA, than GBLUP in inferring genomic relationship matrix and thus predicting target phenotypes. Furthermore, it is worth noting that in the MIR range, some traits exhibited marginal significance in terms of accuracy gain such as RI and TA, traits for which PA was close to unity, so that there is no further scope for improvement.

Our findings are aligned with several studies that highlighted the potency of upweighting QTLs underpinning variation of target traits in genomic selection. This optimization strategy either relies on QTLs encoding for known functions associated with key traits or on GWAS output (Bernardo 2014; Spindel et al. 2016). Herein, QTLs revealed by linkage analysis were proven to contribute to the enhancement of the predictive performance of phenomic selection models for phenological and fruit quality trait. Hence, information on QTLs is likely to complement spectral signature of the evaluated genotype and thus increase the frequency of favorable allelic combinations.

Beyond genetic architecture, including heterogeneous sources of information in prediction models is likely to ensure a valuable accuracy gain. For instance, (Krause et al. 2019) reported that multi-kernel genomic prediction models that integrate genomic markers, pedigree and hyperspectral reflectance, outperformed single-kernel models in forecasting grain yield in wheat within a multi-environment framework. Likewise, (Hayes et al. 2017a) reported that predictions inferred from the concomitant use of SNP information and NIR or nuclear magnetic resonance yielded higher predictive performance within a multivariate modelling context. In addition, (Guo et al. 2016) showed that integration of gene expression and metabolic information along with SNPs in GBLUP model led to a higher PA compared to the benchmark model using only markers. They attributed the accuracy gain to the additional genetic information captured by transcripts and metabolites.

The availability of affordable and cost-efficient phenomic information at an unprecedented scale might translate into genetic gain not only within apricot hybrid breeding but also in collections of genetic resources, through phenomic selection. This breeding strategy potentially permits to screen extensively selection candidates and identify high-performing individuals, either to identify potential elite progenitors for further crosses in the context of breeding or to introgress desirable alleles for germplasm enhancement. In addition, the added value of phenomic selection lies in the capability of spectra to capture G×E interactions and thus the ability to breed for stable genotypes and to sustain fruit production.

CONCLUSION

Our findings underpin the value of phenomic selection, as a novel breeding approach, in predicting apricot fruit quality features as well as traits that are independent from the biochemical constitution of the analyzed samples. The spectral signature is likely to mimic genetic fingerprints and thus renders valuable information on genetic relatedness and underlying linkage disequilibrium, which opens an avenue for integrating phenomics to guide through breeding programs. Indeed, infrared spectroscopy spanning near and mid regions proved efficient to emulate the genomic information and predict target phenotypes. Within this regard, MIRS has proven its capability to portray the biochemical composition of the samples and thus predict the corresponding fruit quality attributes. Also, MIRS-based models outperformed GS models in forecasting tree phenological traits.

In the framework of the optimization of phenomic selection accuracy, spectral preprocessing provides a prominent tool, with regards to its outcome in terms of PA gain. However, response of prediction models to mathematical pretreatment is highly dependent upon the trait. Moreover, weighting QTLs as covariates in NIR- and MIR-models along with reflectance permits to further improve PA, notably for Bayes B model. Therefore, adoption of phenomic selection is valuable not only to ensure enhancement of data acquisition throughput at low cost, but also to enhance PA, which contributes to optimization of genetic gain per unit time.

ACKNOWLEDGMENTS

The authors acknowledge Renaud Rincent and Vincent Segura for their insightful guidance through the elaboration of the analysis. We thank the experimental and the scientific teams of GAFL (Genetics and Breeding of Fruits and Vegetables) and SQPOV (Safety and Quality of Processed fruits and Vegetables) research units as well as the technical staff that helped with the experimentations including phenotyping, spectral acquisition and genotyping. Special thanks to Patrick Lambert, Carole Confolent, Anne-Marie Ferréol, Marielle Boge, Maggy Grotte …

FUNDINGS

This work was funded by the Ministry of Higher Education and Scientific Research (Tunisia) and was supported by INRAE GAFL (France).

LITERATURE CITED

Araus, J.L., and J.E. Cairns, 2014 Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science* 19 (1):52-61.

Banik, D., 2019 Achieving Food Security in a Sustainable Development Era. *Food Ethics* 4 (2):117-121.

Barbin, D., A.L. Felicio, D.-W. Sun, S. Nixdorf, and E. Hirooka, 2014 Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview. *Food Research International* 61.

Barnes, R.J., M.S. Dhanoa, and S.J. Lister, 1989 Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy* 43 (5):772-777.

Bernardo, R., 2014 Genomewide Selection when Major Genes Are Known. *Crop Science* 54 (1):68-75.

Blanco, M., and I. Villarroya, 2002 NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry* 21 (4):240-250.

Broman, K.W., H. Wu, S. Sen, and G.A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19 (7):889-890.

Buddenbaum, H., and M. Steffens, 2012 The Effects of Spectral Pretreatments on Chemometric Analyses of Soil Profiles Using Laboratory Imaging Spectroscopy. *Applied and Environmental Soil Science* 2012:274903.

Bureau, S., D. Cozzolino, and C.J. Clark, 2019 Contributions of Fourier-transform mid infrared (FT-MIR) spectroscopy to the study of fruit and vegetables: A review. *Postharvest Biology and Technology* 148:1-14.

Bureau, S., D. Ruiz, M. Reich, B. Gouble, D. Bertrand *et al.*, 2009a Application of ATR-FTIR for a rapid and simultaneous determination of sugars and organic acids in apricot fruit. *Food Chemistry* 115 (3):1133-1140.

Bureau, S., D. Ruiz, M. Reich, B. Gouble, D. Bertrand *et al.*, 2009b Rapid and non-destructive analysis of apricot fruit quality using FT-near-infrared spectroscopy. *Food Chemistry* 113 (4):1323-1328.

Cozzolino, D., W.U. Cynkar, N. Shah, and P. Smith, 2011 Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality. *Food Research International* 44 (7):1888-1896.

Daetwyler, H.D., M.P. Calus, R. Pong-Wong, G. de Los Campos, and J.M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193 (2):347-365.

Dirlewanger, E., J. Quero-García, L. Le Dantec, P. Lambert, D. Ruiz *et al.*, 2012 Comparison of the genetic determinism of two key phenological traits, flowering and maturity dates, in three Prunus species: peach, apricot and sweet cherry. *Heredity* 109 (5):280-292.

Endelman, J.B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4:250-255.

Fisher, R.A., 1918 The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* 52 (2):399-433.

Gautam, R., S. Vanga, F. Ariese, and S. Umapathy, 2015 Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation* 2 (1):8.

Gebreselassie, M.N., K. Ader, N. Boizot, F. Millier, J.-P. Charpentier *et al.*, 2017 Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from Populus nigra (L.) natural populations. *Industrial Crops and Products* 107:159-171.

Gegas, V., A. Gay, A. Camargo, and J. Doonan, 2014 Challenges of Crop Phenomics in the Post-genomic Era, pp. 142-171.

Geladi, P., and E. Dåbakk, 1995 An Overview of Chemometrics Applications in near Infrared Spectrometry. *Journal of Near Infrared Spectroscopy* 3 (3):119-132.

Grattapaglia, D., F.L. Bertolucci, and R.R. Sederoff, 1995 Genetic mapping of QTLs controlling vegetative propagation in Eucalyptus grandis and E. urophylla using a pseudo-testcross strategy and RAPD markers. *Theoretical and Applied Genetics* 90 (7):933-947.

Guo, Z., M.M. Magwire, C.J. Basten, Z. Xu, and D. Wang, 2016 Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor Appl Genet* 129 (12):2413-2427.

Hayes, B., J. Panozzo, C. Walker, A. Choy, S. Kant *et al.*, 2017 Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theoretical and Applied Genetics* 130:1-15.

Hayes, B.J., J. Pryce, A.J. Chamberlain, P.J. Bowman, and M.E. Goddard, 2010 Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. *PLOS Genetics* 6 (9):e1001139.

Huang, H., J.U. Qureshi, S. Liu, Z. Sun, C. Zhang *et al.*, 2020 Hyperspectral Imaging as a Potential Online Detection Method of Microplastics. *Bulletin of Environmental Contamination and Toxicology*.

Krause, M.R., L. González-Pérez, J. Crossa, P. Pérez-Rodríguez, O. Montesinos-López *et al.*, 2019 Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat. *G3: Genes|Genomes|Genetics* 9 (4):1231.

Lane, H., S. Murray, O. Montesinos-López, A. Montesinos‐López, J. Crossa *et al.*, 2020 Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. 3.

Misra, A., 2014 Climate change and challenges of water and food security. *International Journal of Sustainable Built Environment* 3.

Montes, J.M., A.E. Melchinger, and J.C. Reif, 2007 Novel throughput phenotyping platforms in plant genetic studies. *Trends in Plant Science* 12 (10):433-436.

Nicolaï, B.M., K. Beullens, E. Bobelyn, A. Peirs, W. Saeys *et al.*, 2007 Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology* 46 (2):99-118.

Nsibi, M., B. Gouble, S. Bureau, T. Flutre, C. Sauvage *et al.*, 2020 Adoption and Optimization of Genomic Selection To Sustain Breeding for Apricot Fruit Quality. *G3: Genes|Genomes|Genetics* 10 (12):4513.

Ozaki, Y., S. Morita, and Y. Du, 2006 Spectral Analysis, pp. 47-72.

Pizarro, C., I. Esteban-Díez, A.-J. Nistal, and J.-M.a. González-Sáiz, 2004 Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Analytica Chimica Acta* 509 (2):217-227.

Porep, J., D. Kammerer, and R. Carle, 2015 On-line application of near infrared (NIR) spectroscopy in food production. *Trends in Food Science & Technology* 46:211–230.

Ren, D., L. An, B. Li, L. Qiao, and W. Liu, 2020 Efficient weighting methods for genomic best linear-unbiased prediction (BLUP) adapted to the genetic architectures of quantitative traits. *Heredity*.

Rincent, R., J.-P. Charpentier, P. Faivre-Rampant, E. Paux, J. Le Gouis *et al.*, 2018 Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar. *G3: Genes|Genomes|Genetics* 8 (12):3961.

Rinnan, Å., F.v.d. Berg, and S.B. Engelsen, 2009 Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* 28 (10):1201-1222.

Ruiz, D., M. Reich, S. Bureau, C.M. Renard, and J.M. Audergon, 2008 Application of reflectance colorimeter measurements and infrared spectroscopy methods to rapid and nondestructive evaluation of carotenoids content in apricot (Prunus armeniaca L.). *J Agric Food Chem* 56 (13):4916-4922.

Salazar, J., D. Ruiz, J. Campoy, S. Tartarini, L. Dondini *et al.*, 2016 Inheritance of reproductive phenology traits and related QTL identification in apricot. *Tree Genetics & Genomes* 12.

Savitzky, A., and M.J.E. Golay, 1964 Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36 (8):1627-1639.

Spindel, J.E., H. Begum, D. Akdemir, B. Collard, E. Redoña *et al.*, 2016 Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116 (4):395-408.

Stevens, A., and L. Ramirez-Lopez, 2020 An introduction to the prospectr package.

Verde, I., A.G. Abbott, S. Scalabrin, S. Jung, S. Shu *et al.*, 2013 The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics* 45 (5):487-494.

Walsh, K.B., J. Blasco, M. Zude-Sasse, and X. Sun, 2020 Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use. *Postharvest Biology and Technology* 168:111246.

Wang, X., and G. Zhou, 2011 Study on Pretreatment Algorithm of Near Infrared Spectroscopy, pp. 623-632 in *Computer and Computing Technologies in Agriculture IV*, edited by D. Li, Y. Liu and Y. Chen. Springer Berlin Heidelberg, Berlin, Heidelberg.

White, J., P. Andrade-Sanchez, M. Gore, K. Bronson, T. Coffelt *et al.*, 2012 Field-based phenomics for plant genetics research. *Field Crops Research* 133:101–112.

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao *et al.*, 2014 Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. *PLoS One* 9 (3):e93017.

Chapitre 5 : Evaluation des modèles génomiques et phénomiques dans un panel de diversité

5.1. Présentation du chapitre

Dans les deux précédents chapitres, nous nous sommes concentrés sur l'évaluation des deux stratégies de sélection génomique et phénomique appliquées à la descendance biparentale Go×Mo. Nous avons démontré la pertinence de l'adoption de ces deux approches dans ce dispositif présentant des liens génétiques étroits entre les partitions d'entraînement et de validation. Ce chapitre offre quant à lui un cadre de validation pour évaluer plus largement la capacité prédictive des modèles de SG et de SP dans un contexte de diversité génétique maximisée (core-collection). L'objectif de ce projet de thèse étant de guider les décisions de sélection pour des caractères d'intérêt chez l'abricotier en se basant sur des modèles de prédiction fondés sur des données haut-débit telles que les données génotypiques et spectrales.

A l'encontre du dispositif expérimental sur lequel ont été axés les deux chapitres 3 et 4, le panel de diversité mobilisé dans le cadre de la validation se caractérise par un faible niveau d'apparentement se traduisant sur le plan génétique par de faibles proportions d'allèles identiques par descendance.

L'analyse de la structure du panel par le biais de la méthode de factorisation de matrice non négative parcimonieuse (Sparse Non-negative Matrix Factorization sNMF) a révélé l'existence de huit populations ancestrales minimisant le critère d'entropie croisée (Figure 27A). Par ailleurs, les proportions alléliques ancestrales témoignent d'une forte admixture entre les différents groupes ancestraux (Figure 27B). Il est à noter que cette approche d'inférence de structure ne tient pas compte de l'hypothèse postulant l'équilibre de Hardy-Weinberg et se base sur l'entropie croisée au lieu du maximum de vraisemblance pour l'estimation du nombre de populations ancestrales.

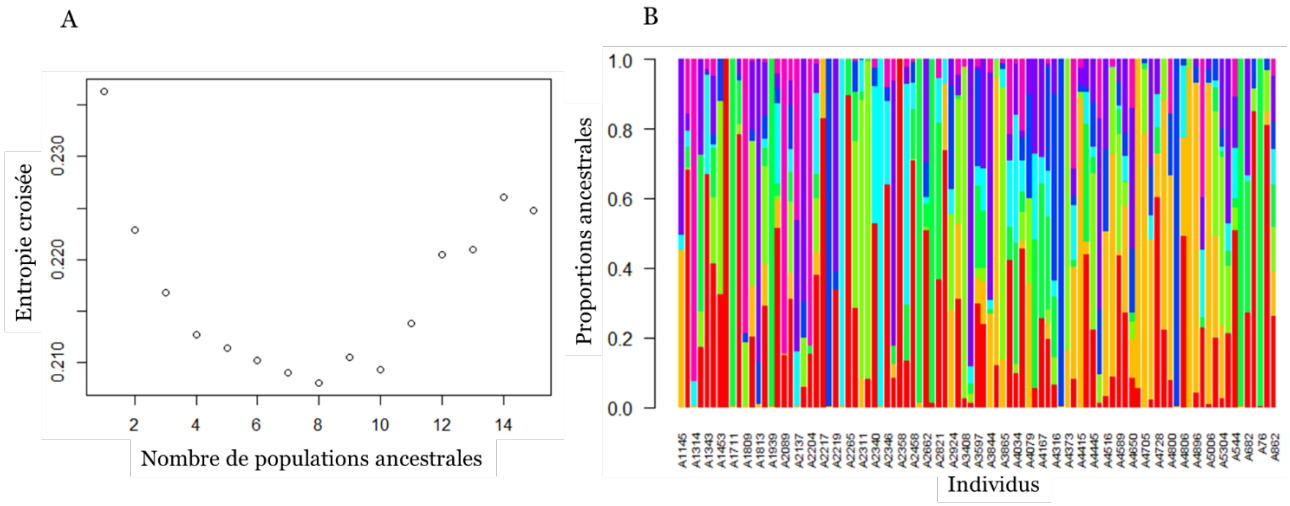


Figure 27: Structure de la diversité génétique de la collection

5.2. Comparaison de la précision des modèles génomiques versus phénomiques

Nous avons évalué la précision des modèles de SP fondés sur des valeurs de réflectances dans le PIR sur feuilles et le MIR sur pulpes des fruits après broyage pour des caractères liés à la qualité des fruits (IR, AT et poids du fruit), la phénologie (dates de floraison et de maturité) ainsi que pour la sensibilité à l'oïdium, au monilia sur fleurs, à la rouille et au chancre bactérien.

La performance de la SP s'est révélée supérieure comparée à celle de la SG pour les deux caractères en lien avec la constitution biochimique des échantillons analysés dans le MIR (IR et AT).

De même, pour la SP basée sur les spectres PIR, des niveaux de précision élevés ont été atteints pour ces deux caractères bien que les spectres aient été acquis sur feuilles. Pour les caractères de phénologie et de sensibilité aux maladies, le modèle de SG s'est montré au contraire le plus performant.

5.3. Optimisation des modèles de sélection phénomique

Au premier abord, nous avons analysé l'architecture génétique des caractères par le biais d'une étude d'association entre variation génétique et variation phénotypique observée. En comparaison avec l'analyse de liaison génétique dans la population biparentale, cette étude offre une meilleure résolution grâce au DL. Pour ce faire, nous avons utilisé le modèle multi-locus prenant en compte l'apparentement moléculaire entre individus afin d'éviter la détection de signaux faux positifs induits par un DL artéfactuel pouvant dériver de la structuration génétique de la population. Les coefficients d'apparentement ont été estimés à partir des

marqueurs moléculaires par le biais de l'inférence de la probabilité d'identité par état (identity by state) des allèles partagés par les individus, pris deux à deux. Les QTLs associés aux différents caractères étudiés sont présentés dans la figure 28.

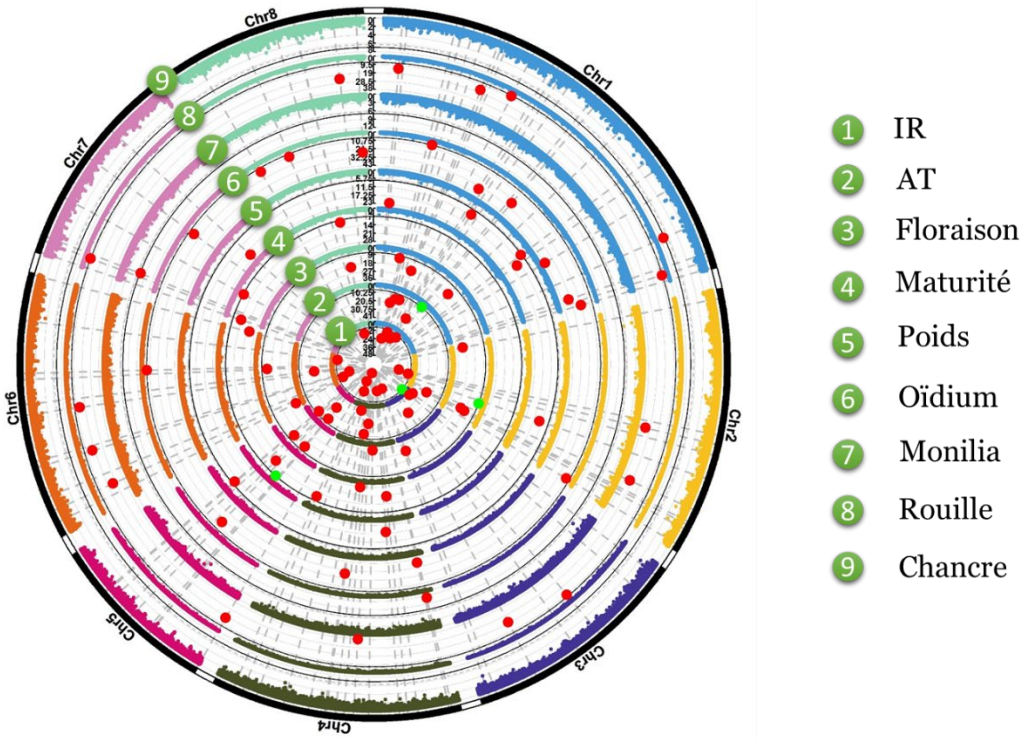


Figure 28: Manhattan plots circulaires indiquant les QTLs identifiés pour les caractères évalués

Etant donné que l'architecture génétique a une incidence fondamentale sur la précision de prédiction et que la modélisation de cette information et son intégration dans les modèles prédictifs s'avère cruciale pour améliorer leurs performances, nous avons intégré les marqueurs les plus significatifs (top SNPs) dans les modèles basés sur les spectres PIR et MIR. Les capacités prédictives des modèles de SP optimisés par la prise en compte des QTLs en effet fixe ont été comparées à celles des modèles de SP de référence incluant les valeurs de réflectance correspondant à chaque génotype pour un nombre d'onde donné, en effet aléatoire.

5.4. Conclusion

Dans ce chapitre, nous avons pu valider la pertinence des modèles génomiques et phénomiques afin de prédire des caractères d'intérêt dans un panel de diversité. Nos résultats soutiennent l'hypothèse portant sur l'efficacité des données pangénomiques et spectroscopiques pour des caractères présentant des architectures génétiques contrastées. Les modèles fondés sur ces données haut-débit se sont montrés plus efficaces que les modèles de la SAM, en termes de

capacité prédictive. Ceci souligne l'intérêt des méthodes de sélection qui sont axées sur la prédiction par rapport à une stratégie de sélection ciblée que représente la SAM. En corollaire, des perspectives intéressantes peuvent s'offrir à la sélection chez l'abricotier non seulement dans dispositifs biparentaux mais également dans des core-collections représentatives de la diversité phénotypique et génétique de l'espèce. Ces collections peuvent constituer une population d'entraînement permettant de calibrer le modèle de prédiction à l'aide de spectres infrarouges ou des marqueurs moléculaires afin d'améliorer des caractères mesurés dans un matériel végétal diversifié.

Par ailleurs, coupler l'information spectrale avec des connaissances a priori sur l'architecture génétique des caractères s'est révélée pertinente afin d'optimiser la performance prédictive des modèles de SP. Néanmoins, au-delà de la précision, la SP offre un cadre d'optimisation du modèle économique par rapport aux méthodes de sélection conventionnelles et à la SG vu les coûts d'acquisition des spectres par rapport à ceux liés au génotypage. Cette hypothèse postulant que la SP pourrait représenter un investissement rentable, mérite d'être validée.

Phenomic selection: What prospects within an Apricot diversity panel?

Abstract

Recent technological breakthroughs have delivered high-throughput data that have shaped the landscape of several breeding schemes pertaining to a variety of crops. For instance, genomic information has shown its worth as a powerful tool in helping selection decisions and accelerating genetic gain through genomic selection. Notwithstanding, a sizeable amount of variation for quantitative traits is not accounted for by genomic selection models including gene – gene interactions. To this is added, phenotyping, regarded as a major bottleneck for genetic progress despite very recent major breakthroughs, such as the use of drones supporting high-throughput evaluation of larger plant populations. Among these technologies, Infrared spectroscopy (IR), which has a wide range of applications in predictive biology since the spectral signature, reflects the structural composition of biological matrices, provides the foundation for phenomic selection, a large-scale and affordable approach enabling to indirectly capture endophenotypic variation and thus complex gene networks underlying the phenotype.

In this study, we assessed comparatively the performance of genomic and phenomic selection models in terms of prediction accuracy within a panel comprising 93 genetically diverse apricot accessions. We used a random sampling cross-validation method in order to compute Pearson's correlation between the observed and the predicted phenotypes linked to nine agronomically important traits encompassing disease susceptibility, fruit quality and tree phenology. The study population was screened using 2.333 and 1.710 IR wavelengths covering the near- and mid-infrared regions, respectively and genotyped with 98.938 SNPs.

Our findings shed light upon the efficiency of phenomic selection in terms of prediction accuracy for refractive index (RI) and titratable acidity (TA) using near- spectra acquired on leaves and mid- infrared spectra acquired on fruits purees, respectively with an accuracy level of 0.98 for RI and 0.81 for TA compared to 0.49 and 0.52, respectively, achieved by means of genomic selection model. Furthermore, phenomic selection models informed by trait genetic architecture outperformed NIRS- and MIRS- based models as well as genomic selection ones for seven out of nine traits. As regards disease-related traits, including top SNPs derived from genome-wide association analysis improved prediction accuracy compared to reference models.

Hence, we clearly demonstrate that phenomic selection provides a valuable predictive framework for agronomically relevant traits in apricot breeding programs.

Keywords: Genomic selection, molecular markers, Phenomic selection, near- and mid- infrared spectroscopy, genome-wide association study, *Prunus armeniaca*

Introduction

In recent years, different technological advancements have paved the way for the release of unprecedented amounts of data that have pioneered the plant breeding landscape and allowed for the discovery of selection strategies that are grounded on large-scale data issued from genomics, transcriptomics, proteomics and metabolomics. These strategies are geared towards prediction-oriented modeling with a view to enhance response to selection and faster the genetic progress.

The development of next-generation sequencing platforms has opened up the path for genome-scale oriented selection strategies such as marker assisted selection (MAS) and subsequently genomic selection (GS) thanks to continuously decreasing genotyping costs and increasing genome coverage.

Supported by genome-wide markers, linkage disequilibrium mapping has permitted to decipher the genetics of important traits and particularly causative loci that have not been identified in QTL mapping population (Alqudah *et al*. 2020). Indeed, in order to overcome the lack of resolution of linkage mapping analysis, due to a limited number of recombination events and low marker density, the focus has shifted towards genome-wide association (GWA) studies, which present a valuable tool to disentangle the genetic architecture of agronomically relevant traits. Thus, with the advent of next-generation sequencing (NGS) techniques, identification of SNPs tightly linked to key traits has been henceforth based on GWA performed in highly diverse genetic panels that have accumulated abundant recombination crossovers (Crossa et al. 2017b). Nevertheless, only a partial proportion of the overall trait heritability is accounted for by the identified loci harboring the association signals, leaving a room for missing heritability which stems from several causes such as unaccounted gene – gene interactions, insufficient sample sizes, rare variants and biased heritability estimates (Makowsky et al. 2011b; Gjuvsland et al. 2013). Despite addressing the missing heritability challenge, the availability of genomic information at unprecedented level has permitted to shape several animal and plant species through GS, which represents an efficient strategy that rhymes with acceleration of genetic gain across generations and time. Its value derives from the genome-wide information provided by

molecular markers that has proved a powerful tool to predict agronomic performance of selection candidates. Indeed, GS values markers tagging QTLs with minor effects and facilitates their use in breeding programs (Meuwissen et al. 2001b; Xu et al. 2020). Even though genome-wide data are able to decipher the hidden genetic control of key traits, GS overlooks complex gene interactions and thus downstream regulations that hold a crucial role in linking genetics to phenotypes (Zhu et al. 2012; Ritchie et al. 2015; Knoch et al. 2021). Hence, the phenotype is determined by the joint contribution of several causal factors and represents the outcome of complex dynamics that include genetic and environmental dimensions (Pigliucci 2005; Gjuvsland et al. 2013). Therefore, contrary to GS where DNA information has a privileged place in interpreting the phenotypic variation, the concept of genotype – phenotype map is based on the contribution of different high-dimensional biological data (omics) including genomics, intralocus dominance and interlocus epistasis to improve our understanding of the mechanisms underlying phenotypes (Alberch 1991; Houle et al. 2010; Pigliucci 2005; Te Pas et al. 2017).

Within this context, several studies outlined the usefulness of endophenotypes such as transcriptomes, proteomes and metabolomes to bridge the gap between genotype and phenotype and estimate the kinship matrix of selection candidates (Akdemir and Isidro-Sánchez 2019). Endophenotypes proved more efficient than genomic data per se in capturing variation across individuals (Mackay et al. 2009; Knoch et al. 2021).

However, endophenotypic information is unaffordable especially at a large scale in breeding programs due to its costs. Therefore, infrared spectroscopy has been proposed as a cost-effective and a large-scale alternative to large-scale biological data including endophenotypic markers (Rincent *et al.*, 2018). It provides unprecedented information on biological samples and thus indirectly captures the endophenotypic variation involved in the control of important traits. (Rincent et al. 2018a) have demonstrated that infrared spectra are able to accurately predict economically relevant traits in winter wheat and poplar using predictive models that are calibrated on non-destructive and cost-effective tools. The deployment of infrared systems encompasses several applications such as chemometric elucidation which has been the pillar of most applications of infrared spectroscopy either to screen candidates in breeding programs or to interpret the biochemical composition of fruits and vegetables to determine their quality. Indeed, different studies highlighted the importance of deploying NIR systems in postharvest technology in order to accurately assess quality attributes of fruits and vegetables (Nicolaï et al. 2007; Bureau et al. 2019). Although PS makes use of infrared spectroscopy that is

implemented routinely within the context of predictive biology, the traits targeted by PS might be or not linked to the biochemical composition of the analyzed samples. Drawing on the proof of concept study on phenomic selection, several studies confirmed the capability of infrared spectra to capture the hidden control contributors related to the quantitative variation. In this regard, deployment of NIR spectra to predict important phenotypes in wheat and poplar has been extrapolated to some crop species. For instance, (Gonçalves et al. 2021) have demonstrated that using NIR wavenumber variables as predictors exhibited higher prediction accuracy for feedstock quality compared to SNP-based models in sugarcane. Moreover, (Krause et al. 2019) showed that models based on relationship matrices derived from hyperspectral reflectance performed similarly to or superior to SNP- and pedigree-based genomic selection models in predicting grain yield in wheat within and across environments. Contrary to vegetation indices that leverage spectral information from a limited number of wavelengths, hyperspectral imaging relies on a large number of narrowband wavelengths spanning the visible and the NIR regions and thus offering a more robust framework for selecting candidates at multiple locations and at an early generation stage (Krause et al. 2019). In (Hayes et al. 2017b), models built using NIR or nuclear magnetic resonance (NMR) along with the genomic information were superior to models including only molecular markers in predicting grain end-use quality traits.

Conceptually, PS accuracy is grounded on linkage disequilibrium and allele sharing between the candidates, as indeed GS (Rincent et al. 2018a). Conversely, beyond genetic variance, this strategy considers the interactions unaccounted for by GS notably linked to environmental factors and thus enables to make inferences by means of infrared spectroscopy about genetic values of candidates and gain insights on the best-performing ones without the need for revealing the causal genetic variation.

Along with the expectations towards GS, the drive behind the recourse to PS in Apricot stems from an urgent need to improve genetic gain in order to tackle the numerous challenges faced by the different stakeholders of apricot sector. Owing to biological constraints of this species linked to gametophytic incompatibility as well as environmental adaptability and susceptibility to a wide range of diseases, the response to selection and thus genetic gain in apricot are limited (Bassi et al. 2016). Therefore, enhancing the genetic gain per unit time is of utmost importance in apricot breeding, whose efforts have been oriented towards quantitatively inherited traits, supporting the need for innovative approaches such as GS and PS to enhance the genetic progress.

For instance, dissection of the genetics of traits linked to apricot fruit quality, tree phenology and disease susceptibility has been performed through linkage mapping using biparental populations (Dirlewanger et al. 2012; Nsibi et al. 2020) as well as association mapping within a diversity panel (Omrani et al. 2019), which has enabled the discovery of valuable Quantitative trait loci (QTL) information underpinning fruit quality and canker sensitivity. Conversely, these studies also demonstrated that a significant amount of genetic variance remains unreachable. Thus, the focus has partly shifted towards GS and PS in order to cover the genetic variance unaccounted for by MAS.

In this regard, the main objectives of our study were to assess the prediction performance of a proof of concept for infrared-based models compared to SNP-based models and to evaluate the extent to which PS models are able to predict key traits in apricot breeding programs. In addition, we aimed at optimizing PS models by means of trait genetic architecture, which was assessed via GWA analysis.

Materials and methods

Plant material

The study panel is composed of 93 apricot cultivars grown under a low phytosanitary input cropping system. The accessions were selected among INRAE germplasm collections with a focus on maximizing the phenotypic and genetic diversity based on results reported in (Bourguiba et al. 2020; Bourguiba et al. 2012) and grafted on Montclar ® Chanturgue peach rootstock. The apricot panel was planted in 2018 at the INRAE UERI experimental site (Gotheron, France) within a randomized design with 5 blocks and 1 replication per block for each accession (single-tree randomized block design). This study panel is part of a multi-trial experimental design located in different apricot production areas in France (Gotheron, L'Amarine and Torreilles) and Switzerland (Conthey).

Phenotyping for fruit quality

The determination of the refractive index (RI), a measure correlated to the fruit soluble solid content, was carried out using a digital refractometer and expressed in $\%\ Brix$ at a temperature equal to $20^{\circ}C$. The fruit titratable acidity (TA) was measured by titration of an aliquot of mesocarp using a $0.1\ N$ sodium hydroxide solution until reaching a pH equal to 8.1 and expressed in milliequivalents per gram of fresh matter.

Phenotyping for phenological traits

Phenotyping for tree phenology focused on the date of flowering when 50% of the flower buds had reached the flowering stages corresponding to BBCH60, 65, 69 and the date of maturity (BBCH87) at the approach of physiological maturity assessed by the softening of the fruits due to the loss of fruit firmness associated with the loss of adhesion of the cell wall components.

Phenotyping for disease susceptibility

The characterization of the study population for susceptibility to bacterial and cryptogamic diseases aims at limiting the use of phytosanitary products to respond to sanitary and environmental issues. In this perspective, the collection was phenotyped for susceptibility to diseases that can be detrimental to production and the sustainability of the orchard. These include powdery mildew on leaves (*Sphaerotheca pannosa*), brown rot on flowers (*Monilia laxa*), rust (*Tranzschelia discolor*) and bacterial canker (*Pseudomonas syringae*). The evaluation of the varietal susceptibility to the above-mentioned diseases was carried out over one year of measurement (2020), as follows:

Susceptibility to powdery mildew

Phenotyping for susceptibility to powdery mildew was performed on 50 fruits per tree that were sampled as the fruits approached maturity. It was based on the scoring of the percentage of powdery mildew fruits.

Susceptibility to brown rot

The characterization of the susceptibility to brown rot was carried out within natural climatic conditions that were favorable to the infection. Scoring of the disease took place 30 days after flowering so that the symptoms on branches are expressed and the vegetation does not hinder the observation and the notation. Disease scoring was based on the percentage of branches with dried flowers following the attack of the fungus compared to the total number of flowering branches on the tree.

Susceptibility to rust

The assessment of rust susceptibility was based on the visual estimation of symptoms' severity in terms of leaves drop percentage. The leaf decay rating scale ranged from 0 for 0% dropped leaves to 5 for more than 80% of leaves dropped due to rust.

Susceptibility to canker

Phenotyping for susceptibility to bacterial canker caused by *Pseudomonas syringae* was performed through observation of gum spots per carpenter under natural inoculation of field conditions. Symptoms were quantified towards the end of flowering before branch desiccations related to moniliosis attacks at flowering appeared. The percentage of carpenters showing symptoms (necrotic buds or spurs with or without the presence of gum and with reddish-brown bark on the surface and brown-necrotic tissue underneath and young non-emerging branches with brown-necrotic tissue underneath the bark and/or the presence of gum) was estimated according to a scale varying from 0 to 100% with an increment of 25%, according to the severity of symptoms.

Genotyping

The study panel was genotyped using Illumina HiSeq2000 NGS technology at the GeT-PlaGe genomics platform (INRAE Toulouse) following the protocol of (Elshire et al. 2011). Library preparation was carried out at AGAP research unit (INRAE Montpellier). Raw reads were filtered according to their quality score (QS>20). DNA sequences were aligned using the genome of a Moroccan apricot cultivar (Marouch) via Burrows-Wheeler software (Burrows-Wheeler Aligner BWA) (Li and Durbin 2009). Duplicated sequences were removed with Samtools (samtools v1.9, sambamba v0.7.1) and alignment quality was checked with multiQC. SNPs and indels were called with GATK (gatk4 v4.1.4.1) and filtered. A total of 98.938 biallelic SNPs with a MAF threshold of 0.05 and a maximum of 20 % missing data were then selected to span all the genome at homogenous intervals providing a coverage density of approximately one marker every 1.958 bp.

Spectral acquisition

Spectral characterization was performed in 2020 for NIRS and over two years of measurements (2019 and 2020) for MIRS. The spectral ranges varied from 12.493 to 4.000 cm^{-1} (NIR) and from 4.000 to 700 cm^{-1} (MIR). The NIR spectra were acquired on leaves collected at a rate of 5 replicates per genotype. MIR spectra were obtained from fruit mesocarp homogenates.

Phenotypic modeling

Decomposition of the phenotypic records was performed via analysis of variance. The amount of variance accounted for by the significant factors was estimated with reference to the sum of

squares of each factor divided by the total sum of squares added to the random error (residual sum of squares).

Phenotypic values corresponding to the nine targeted traits were adjusted according to different factors of the experimental design that contributed to phenotypic variation using lme4 package following the mixed model:

$$y_{ijklm} = \mu + g_i + \alpha\beta_{jk} + \beta\gamma_{kl} + \gamma\delta_{lm} + e_{ijklm} \quad (1)$$

where y_{ijklm} denotes the phenotypic value corresponding to the genotype *i* for the replication *j*, the block *k*, the year *l* and the maturity stage *m*, μ is the overall phenotypic mean, g_i is the random effect of the genotype *i* with $g \sim N(0, A\sigma_g^2)$, α_j refers to the fixed effect of the replication *j*, β_k is the fixed effect of the block *k*, γ_l is the fixed effect of the year *l*, δ_m is the fixed effect of the maturity stage *m* and e_{ijklm} is the residual effect with et $e \sim N(0, I\sigma_e^2)$.

The block effect corresponds to the five experimental blocks, the replication effect corresponds to two repetitions and the year effect corresponds to two years of phenotypic characterization of RI and TA (2019 and 2020). Disease-related and phenological traits were recorded in 2020. All factors were modeled as fixed effects with the exception of genotype and residual factors.

In order to meet the assumption of normality, we used bestNormalize package that offers several transformation functions and selects the best normalizing transformation (Peterson and Cavanaugh 2020). Hence, random effects were assumed to be independently and identically distributed following a normal distribution with mean zero and variance σ_g^2 and σ_e^2 for the genotypic and the residual effects, respectively so that $g_i \sim N(0,\ \sigma_g^2)$ and $e_{ijk} \sim N(0,\ \sigma_e^2)$. Thereafter, best linear unbiased predictions (BLUPs) corresponding to the genotypic effect were obtained using 'lmer' function from 'lme4' R package (Bates et al. 2014b).

Broad-sense heritability was computed as shown in equation (2):

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \quad (2)$$

where σ_g^2 and σ_e^2 denote the genetic variance and residual variance components, respectively.

Spectral pretreatment

Prior to predictive modeling, spectral data were subjected to several preprocessing techniques in order to preclude spectral noise derived from measurement errors. The chemometric

pretreatment protocols were performed using 'prospectr' (Stevens and Ramirez-Lopez 2020) R package to compute the derivation and detrend, respectively. Detrend pretreatment consisted of performing a standard normal variate (SNV) transformation followed by fitting a second order linear model to correct for wavelength-dependent scattering effects (Barnes et al. 1989)(Barnes et al. 1989). Derivatives were computed using Savitzky–Golay filter (Savitzky and Golay 1964) with a smoothing window length of 37 data points (74 cm^{-1}) and polynomial order n=2. Thereafter, the mean reflectance across replicates generated 2.333 averaged wavelengths in NIR region and 1.710 averaged wavelengths in MIR region. The variation range of PS accuracy was inspected and the effect of spectral data pretreatment on prediction performance was investigated.

Genome-wide association

In order to characterize the genetics underlying the targeted traits, a genome-wide association (GWA) analysis was performed using multi-locus mixed model (MLMM) with forward selection, which consists on including the SNP with the smallest p-value as a fixed predictor within the regression and backward elimination. Contrary to single-locus approach, MLMM SNPs are identified by models including cofactors with marginal p-values below the significance threshold. The stopping criterion for the forward inclusion of cofactors within the regression model was defined by the ratio $\hat{\sigma}_g^2/var(y)$. Hence, forward regression stops when the proportion of phenotypic variance explained by polygenic effect is close to zero and is followed by a backward stepwise regression dropping cofactors with the least significant p-values (Segura et al. 2012). With view to minimizing bias derived from false genotype – phenotype association signals, stringent thresholds using multiple-Bonferroni criterion were used. In addition, in order to avoid spurious associations and separate confounding factors linked to population structure from actual association signals, an identity-by-state kinship matrix was included as a covariate.

At each step of the forward selection, SNPs that are significantly associated with the phenotypic variation were included as covariates within the regression model. Afterwards, model selection was performed via eBIC (extended Bayesian Information) (Bonnafous et al. 2018). Only SNPs exhibiting the lowest association signals were considered as representative of a QTL, thus only top SNPs were considered. Only additive genetic variance was considered.

Genomic and phenomic selection modeling

In order to evaluate the performance of SNP- and infrared spectra-based models, we used a random cross-validation scheme with 100 replicates where 75% of records (n=70) were assigned to the training partition and the remaining 25% (n=23) assigned to the validation partition. As for prediction, we used RR-BLUP model provided in 'mixed.solve' function of 'rrBLUP' R package (Endelman 2011a) to estimate SNP and wavenumber effects and predict the performance of supposedly unphenotyped individuals within the partition validation.

Prediction accuracy was assessed using a four-fold cross-validation scheme with 100 iterations, using the statistical model provided in equation (3):

$$y = X\beta + Zu + e \quad (3)$$

where y is the vector of phenotypic records, X is an incidence matrix for fixed effects, β is a vector of fixed effects, Z is an incidence matrix for random effects, u is the vector of random effects and e denotes the vector of random errors.

Optimization of phenomic selection accuracy

The optimization scenario consisted of including top SNPs with the lowest p-values and thus the highest association signals in genomic and phenomic selection models as fixed-effect covariates. QTL scanning via GWA analysis was performed iteratively within training partitions using MLMM approach including kinship matrix to prevent biased estimation of marker effects and thus overestimation of prediction accuracy. Predictive abilities of NIRS- and MIRS-based models were compared to GS models as well as MAS ones that include only top SNPs tagging the putative QTLs associated with the nine investigated traits.

Data availability

Supplemental data are available in Files S1-6. File S1 contains the raw phenotypic data. Broad-sense heritability estimates are presented in File S2. File S3 contains the genotypic data. NIR and MIR spectra are available in Files S4 and S5, respectively. Information on GWA results are available in File S6. Files S7-8 enclose the impact of spectral pretreatment on PS accuracy.

Results

Phenotypic variation

Regarding the partitioning of the investigated traits (Figure 1A), the genotypic variance was shown to contribute significantly to phenotypic variation. It ranged from 21.6% for rust susceptibility to 97% for maturity date. For traits linked to disease susceptibility, the proportion

of variance attributed to the block effect varied from 1.4 for P.mildew to 66% for F.weight. The factor corresponding to replications explained 74% of the variance in rust susceptibility. As for the residual term, it accounted for 0.6% for maturity date to 23.7% for canker susceptibility.

Along with the significant genetic variation associated with the investigated traits, broad-sense heritability estimates varied from 0.29 for canker susceptibility to 0.97 for maturity date (Figure 1B). Disease-related traits except for powdery mildew and rust susceptibility exhibited the lowest heritability estimates. Fruit quality and tree phenology traits were shown to be moderately to highly heritable with heritability estimates ranging from 0.57 for flowering date to 0.97 for maturity date.

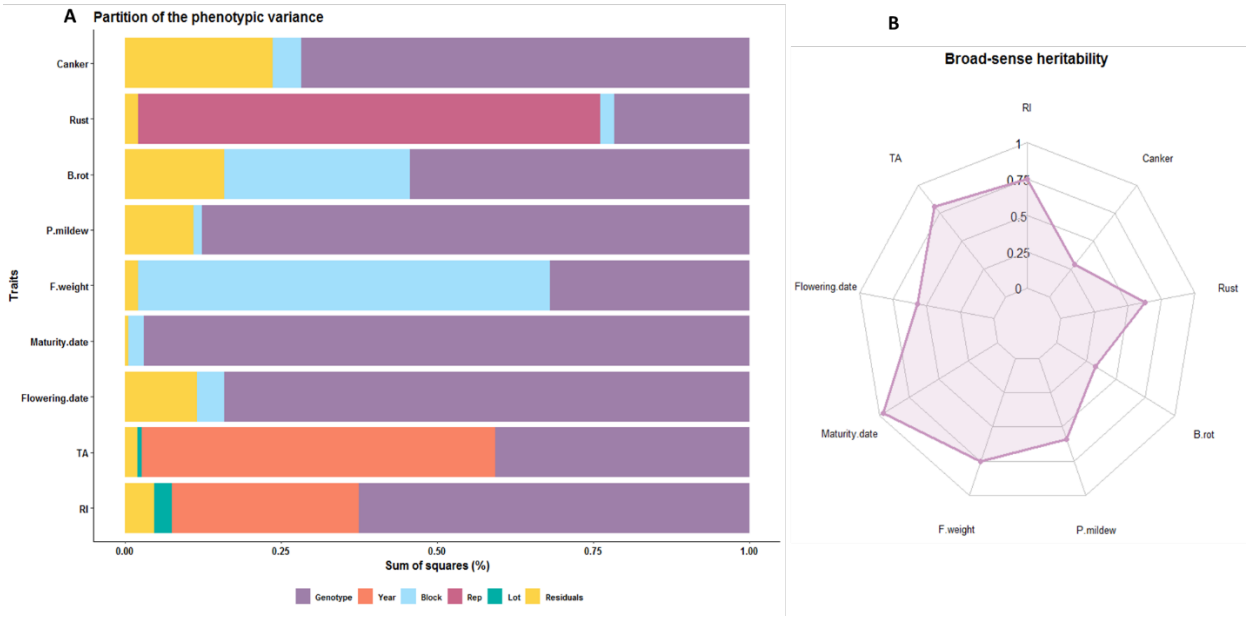


Figure 1: Exploration of the phenotypic records: Partition of the phenotypic variance into genotypic and environmental components (A) and broad-sense heritability estimates.

Genome-wide association

Regression of the phenotypic values on marker frequencies revealed several association signals for all the traits spanning the eight chromosomes. GWA analysis resulted in 114 identified top SNPs that exceeded the 5% threshold of significance for the analyzed traits. For fruit quality, 18 SNPs were detected for RI, 18 for TA and 9 for F.weight. Concerning the disease susceptibility, 18 top SNPs were detected for P.mildew, 5 for B.rot, 16 for Rust and only one SNP for Canker, which was located on chromosome 1. Concerning phenological traits, 18 top SNPs were identified for Flowering.date and 11 for Maturity.date. No overlap between QTLs was noted across the investigated traits.

Manhattan plots displaying the association signals for the nine investigated traits are provided in Figure 2, exhibiting top SNPs with the lowest p-values as well as the genome-wide thresholds. Detailed information on the detected SNPs is provided in File S6.

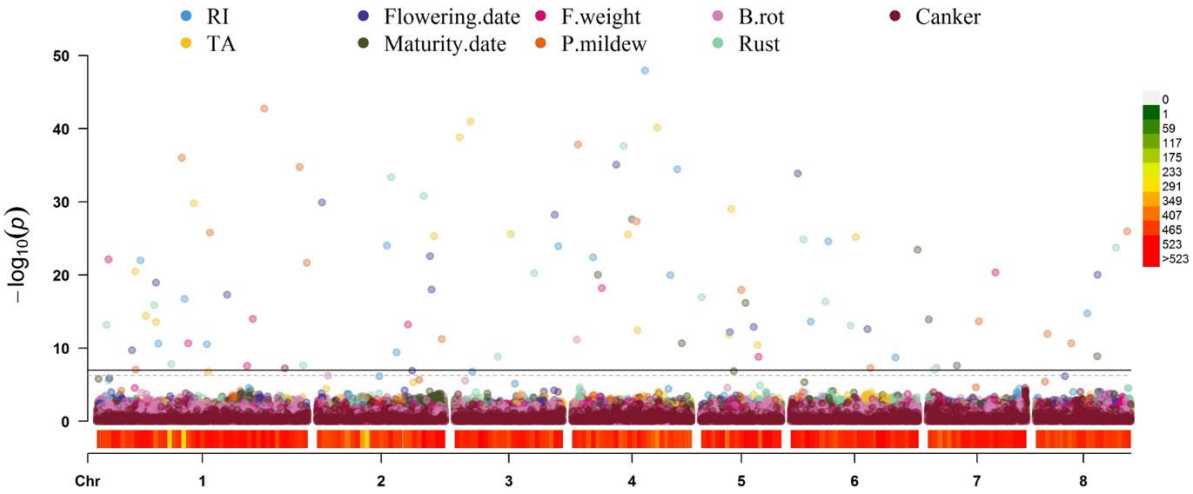


Figure 2: Manhattan plots for nine traits displaying $-log_{10}\,(p)$ plotted against the physical position across the eight chromosomes.

The horizontal solid line indicates a significance threshold of 0.05 divided by markers number. Dashed line represents a significance threshold of 0.01 divided by markers number. SNP above the threshold were declared as significantly associated to the investigated traits.

Genomic versus phenomic selection accuracy

Prior to phenomic prediction, we assessed the effect of spectral pretreatment on phenomic selection accuracy. Preprocessing of the NIR and MIR spectra displayed a wide range of responses in terms of PA according to the targeted traits and the infrared range. As outlined in Figure S2, the raw spectra depicted either the largest variation in accuracy or the lowest accuracy mean for the majority of the traits. Some pretreatment algorithms exhibited a wider range of accuracy variation.

For instance, NIRS- and MIRS- based models depicted a higher PA for fruit quality traits, compared to SNP-based model. As shown in Figure 3, the mean PA of PS models reached 0.98 for RI and 0.81 for TA. Yet, GS model provided 0.49 and 0.52 of PA for these two traits. As for phenological and disease-related traits, GS model outperformed PS models.

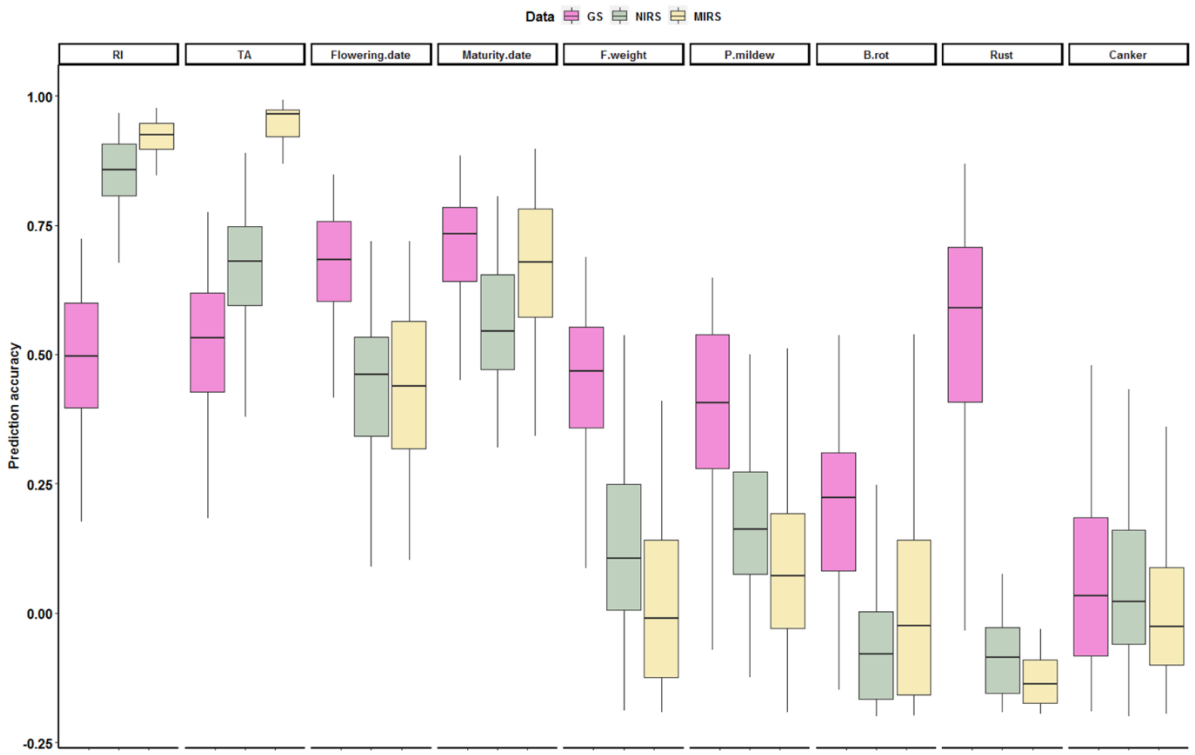


Figure 3: Comparison of prediction models using SNPs, near-infrared (NIR) and mid-infrared (MIR) spectra.

Boxplots illustrate the variation in prediction accuracy computed as Pearson's correlation between observed phenotypes and predicted ones.

Optimization of PS accuracy

As outlined in Figure 4, PS Models informed by trait genetic architecture derived from GWA analysis resulted in a systematic accuracy gain for the majority of the traits with the exception of RI and TA, whose mean PA was 0.85 and 0.66 for NIRS and 0.92 and 0.95 for MIRS, respectively using reference models without top SNPs as covariates. However, the mean of PS accuracy was lower than GS accuracy for five traits out of nine. Yet, PS models outperformed MAS ones for all traits except for canker susceptibility, displaying a slight decrease in mean PA. MIRS+QTL models yielded higher PA than NIRS+QTL models. As for the reference PS models, NIRS-based models outperformed MIRS-based models for five traits out of nine.

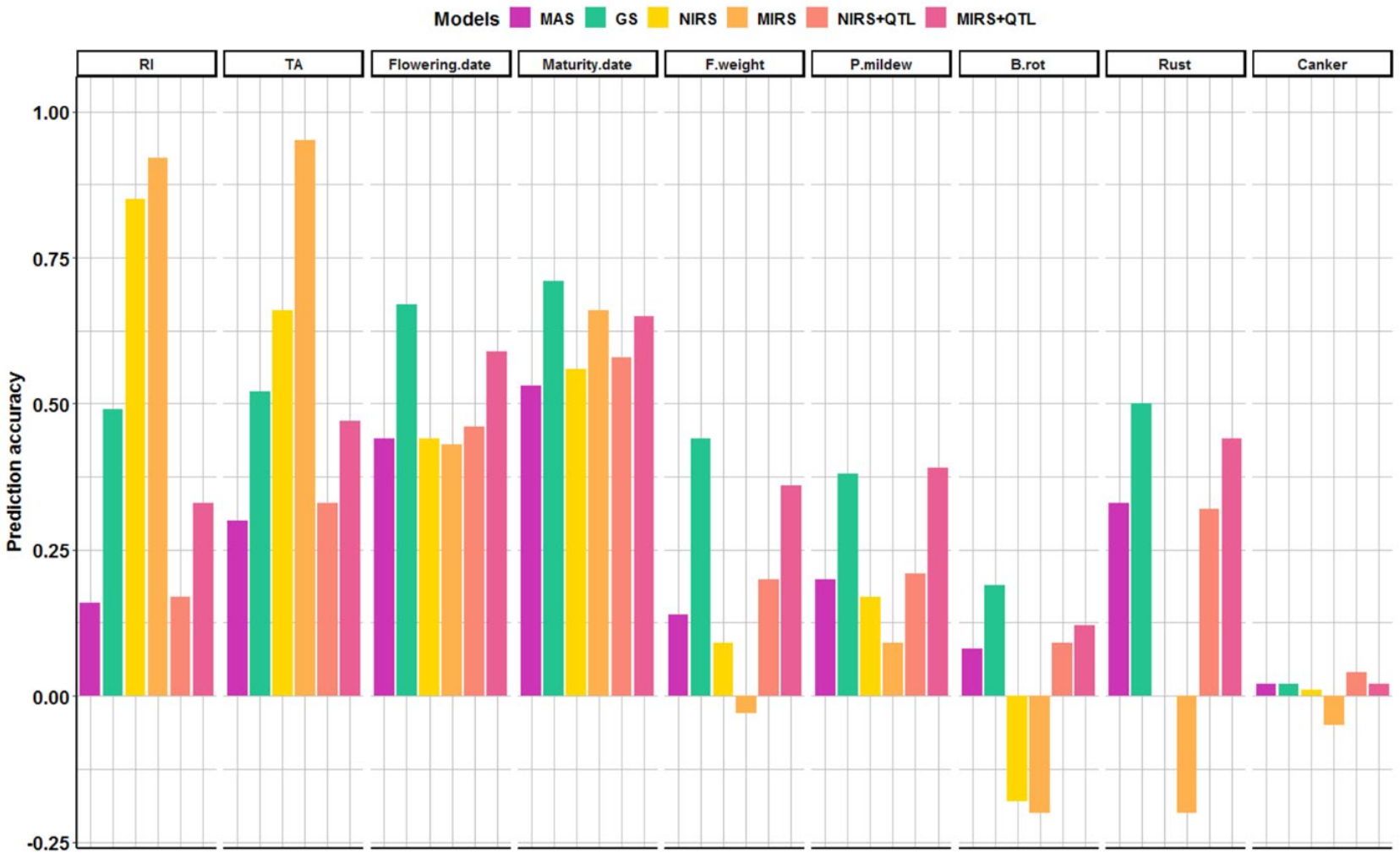


Figure 4: Comparison of prediction accuracy pertaining to MAS, GS and PS models

Barplots displaying the mean prediction accuracy of models integrating only the top SNPs identified by genome-wide association analysis (MAS), SNP-based models (GS) and phenomic selection models (NIR and MIR) informed by the trait genetic architecture where top SNP are included as fixed-effect covariates (NIRS+QTL and MIRS+QTL, respectively).

Discussion

The drive behind marker-based selection strategies is motivated by the ambition to faster the genetic gain per unit of time notably for crop species whose genetic improvement is hampered by several biological and environmental constraints. This is the case for perennial crops including *Prunus armeniaca*. For instance, MAS in apricot reduced breeding cycles as it is implemented at an early developmental stage given that DNA sequence information is acquired on juvenile tissues. However, apricot breeding programs that are endorsed by MAS are restricted to auto-fertility and resistance to sharka, which are controlled by small effect genes. In addition, the lack of knowledge on QTL stability across genetic backgrounds and environments contributed to a limited implementation of MAS within a wide range of genetic materials (Kumar et al. 2012a). This is in concert with several studies that agreed that MAS is not suited for complex traits as it is unable to cover quantitative variation brought by small-effect loci (Bernardo and Yu 2007; Heffner et al. 2011; Crossa et al. 2017b). As a corollary, the focus has been geared towards GS to uncover the genetic variance unaccounted for by MAS notably for traits controlled by polygenes. Over the past decade, GS have led the way for genetic improvement thanks to the advances in molecular genetics. Its implementation in breeding programs aims at circumventing the shortcomings of selection strategies based on QTL identification supported either by population-wide linkage disequilibrium (association mapping) as well as population-wide linkage equilibrium (linkage mapping) (Dekkers 2004). Genomic information is crucial for prediction-oriented selection strategies as it influences three parameters out of four of the breeder's equation including reducing breeding cycle length, enhancing prediction accuracy and increasing selection intensity at a given budget in comparison to phenotypic selection. Thereby it optimizes the genetic gain per unit time which is crucial notably for perennial fruit crops whose breeding cycles are impeded by a slow rate of genetic gain (Heffner et al. 2010).

Along with this, (Rincent et al. 2018a) proved that NIRS-based models outperformed GS-based ones in predicting agronomic performances of selection candidates. Several scenarios were

assayed for winter wheat and poplar taking into consideration a combination of parameters linked to traits, tissues, genotyping and NIRS acquisition costs. Even though wavelengths in NIRS are partly environment specific, predictions based on joint analysis of several wavelengths across a wide range of independent environments outperformed marker-based predictions even when the correlation between the evaluated environments is low. Hence, PS models that are trained in a given environment are able to predict accurately traits that are characterized in independent environmental conditions. More importantly, PS holds a promise for dramatically accelerating the genetic gain which ranged from 11% to 127%.

Beyond prediction accuracy, a successful implementation of PS across multiple generations is determined by the frequency of re-estimation of genetic values inferred by infrared spectra through equation update in order to maintain an acceptable level of accuracy (Rincent et al. 2018a). This is of a paramount importance for both selection strategies, genomic and phenomic, with view to capturing LD information across generations of selection (Calus et al. 2008b). Given the low cost and the facility of acquisition of spectral data, determining the frequency of re-estimating genetic values of selection candidates might be performed at an early developmental stage, which is particularly advantageous for apricot. Indeed, early selection is likely to enable reduced breeding cycle length. Hence, PS could be integrated in apricot breeding programs as a routine activity to screen selection candidates without large investments. Yet, despite the dramatic drop in genotyping costs, these latter remain unaffordable and outpace spectral acquisition costs notably within the context of multitrial breeding programs. Added to this, the need for increasing the density of molecular markers in order to increase LD between SNPs and QTLs resulting in a slower reduction of PA across generations (Calus 2010), which subsequently increases the genotyping costs. However, the lack of access to genome-wide information is not only attributed to limited high-throughput and cost-effective genotyping tools (Furbank and Tester 2011; Rasheed et al. 2017) but also to conventional phenotyping, which has lagged behind genotyping, as acquisition of phenotypic records represents a laborious and time-consuming task is and is considered as a major bottleneck for genetic progress (Cobb et al. 2013; Araus and Cairns 2014; Kumar et al. 2016). These constrains might hamper the implementation of GS within breeding programs.

Alternatively, adoption of PS within breeding programs enables to infer genetic values of individuals using accurate and inexpensive tools. For instance, equipping harvesters with infrared imagers permits to save time and reduce the sample handling costs (Araus et al. 2018), which is also the case for portable devices that are adapted to field experiments. Yet, the use of

infrared spectra with a view to enhancing genetic gain per unit time and cost stems not only from their capability to optimize the economic model due to their affordability but also from their throughput capacity, which contributes to large-scale evaluation of the agronomic performance of selection candidates. Therefore, screening larger breeding populations helps to enhance selection intensity especially thanks to the recourse to automation and remote-sensing which are amenable to large-scale field trials contributing to the generation of an unprecedented amounts of data. In addition, recent technological advancements has allowed for the extension of the use of infrared spectroscopy to cover a myriad of sophisticated techniques such as hyperspectral imaging, which enables to capture spatial variation as it covers large geographic areas and subsequently account for G×E interaction (Araus et al. 2018). Beyond the genetic variance, infrared spectroscopy offers a valuable tool to apprehend the complexity of G×E interaction and thus predict the differential response of genotypes across a wide range of environmental conditions and rank them according to their agronomic behavior. Indeed, deployment of low-cost tools across environments enables to expand the training population, promote genetic diversity and optimize experimental design of field trials.

Beyond reference models, with a view to optimize prediction accuracy, we assessed models that take into account causal QTLs of the trait genetic architecture. Accordingly, unequal weights were attributed to SNPs. Within this regards, several empirical and simulation studies highlighted the usefulness of treating SNPs tightly linked to important QTL as covariates in GS models. For instance, RR-BLUP assumes equal marker variance regardless of the amount of variance explained by the QTLs underlying the trait, which is considered as an unrealistic assumption. Yet, within optimized models, emphasis has been put on top SNPs declared as significantly associated with the targeted traits. Therefore, QTL effect have a specific variance different from random genome-wide markers, which is in accordance with the trait genetic architecture (Bernardo 2014; Arruda et al. 2016). In addition, adding major genes to GS model is more advantageous as the amount of explained genetic variance increases ($R^2 = 50\%$) and the trait is highly heritable ($H^2 = 0.8$). Conversely, including QTLs accounting for $R^2 < 10\%$ is disadvantageous for prediction accuracy (Bernardo 2014). Beyond the accuracy gain, PA decays rapidly in later cycles when major genes are treated as random effects instead of fixed-effect covariates. Moreover, (Zhang et al. 2016) has demonstrated that accounting for locus-specific variance in GBLUP led to a higher PA in comparison with Bayes B and Bayes C, which apply marker selection.

Regarding PS models, coupling spectral and genomic information tightly linked to the targeted traits could provide an accurate framework for the estimation of breeding values, which is consistent with a wide range of simulation and empirical studies that highlighted the usefulness of the combination of several sources of biological information with the aim of screening selection candidates (Hayes et al. 2017b; Krause et al. 2019).

PS accuracy was higher for traits associated with fruit quality, compared to GS model. The spectral characteristics of infrared radiation change according to the chemical composition of the samples (Nicolaï et al. 2007). Nevertheless, for traits which are independent from the biochemical constitution of the analyzed samples such as disease-related traits, we recommend the use of large number of biological samples to cover a substantial range of phenotypic variation and increase repeatability in order to enhance PS accuracy and thus improve selection response for the targeted traits. Nevertheless, even with lower PA compared to GS, PS enables selection by means of cost-effective resources at a high throughput capacity and at an early stage, which is likely to translate into genetic gain.

CONCLUSION

PS is grounded on harnessing infrared spectroscopy to infer genotypic values and support selection decisions. Aimed at covering the phenotypic variance unaccounted for by GS, PS offers a cost-effective alternative to selection based on Endophenotypic markers. Thanks to its high throughput capacity, PS allows for improving the scale of field trials and thus enhancing selection intensity, which are synonym to optimizing genetic gain. Within this framework, our findings revealed that MIRS-based models are preferred over GS models in accurately predicting traits linked to fruit quality with reference to its biochemical composition. As for traits that are independent from the chemical constitution of the analyzed samples, NIRS characterization performed on leaves exhibited higher PA with respect to GS for RI and TA. Concerning disease-related traits, GS outperformed PS models in predicting P.mildew, B.rot and rust susceptibility and to a lesser extent canker sensitivity. More importantly, leveraging trait genetic architecture in PS models led to higher accuracies compared to reference models, MAS and GS models for the majority of the traits with the exception of those depicting already high PA using NIRS- and MIRS-based models.

To sum up, adoption of PS in apricot program holds a valuable potential toward accelerating genetic gain and unlocking new avenues for apricot selection and creation of agronomically superior varieties. Further, characterizing a wide range of environments through envirotyping

to complement genotyping and phenotyping and integrating environmental variability within PS models might offer a potential means of optimizing the accuracy of phenomics models.

LITERATURE CITED

Akdemir, D., and J. Isidro-Sánchez, 2019 Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports* 9 (1).

Alberch, P., 1991 From genes to phenotype: dynamical systems and evolvability. *Genetica* 84 (1):5-11.

Alqudah, A.M., A. Sallam, P. Stephen Baenziger, and A. Börner, 2020 GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley – A review. *Journal of Advanced Research* 22:119-135.

Araus, J.L., and J.E. Cairns, 2014 Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science* 19 (1):52-61.

Araus, J.L., S.C. Kefauver, M. Zaman-Allah, M.S. Olsen, and J.E. Cairns, 2018 Translating High-Throughput Phenotyping into Genetic Gain. *Trends in Plant Science* 23 (5):451-466.

Arruda, M.P., A.E. Lipka, P.J. Brown, A.M. Krill, C. Thurber *et al.*, 2016 Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (Triticum aestivum L.). *Molecular Breeding* 36 (7):84.

Barnes, R.J., M.S. Dhanoa, and S.J. Lister, 1989 Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy* 43 (5):772-777.

Bassi, F.M., A.R. Bentley, G. Charmet, R. Ortiz, and J. Crossa, 2016 Breeding schemes for the implementation of genomic selection in wheat (Triticum spp .). *Plant Science* 242:23-36.

Bates, D., M. Mächler, B. Bolker, and S. Walker, 2014 *Package Lme4: Linear Mixed-Effects Models Using Eigen and S4*.

Bernardo, R., 2014 Genomewide Selection when Major Genes Are Known. *Crop Science* 54 (1):68-75.

Bernardo, R., and J. Yu, 2007 Prospects for Genomewide Selection for Quantitative Traits in Maize. *Crop Science* 47.

Bonnafous, F., G. Fievet, N. Blanchet, M.C. Boniface, S. Carrère *et al.*, 2018 Comparison of GWAS models to identify non-additive genetic control of flowering time in sunflower hybrids. *Theor Appl Genet* 131 (2):319-332.

Bourguiba, H., J.-M. Audergon, L. Krichen, N. Trifi-Farah, A. Mamouni *et al.*, 2012 Loss of genetic diversity as a signature of apricot domestication and diffusion into the Mediterranean Basin. *BMC Plant Biology* 12 (1):49.

Bourguiba, H., I. Scotti, C. Sauvage, T. Zhebentyayeva, C. Ledbetter *et al.*, 2020 Genetic Structure of a Worldwide Germplasm Collection of Prunus armeniaca L. Reveals Three Major Diffusion Routes for Varieties Coming From the Species' Center of Origin. *Frontiers in Plant Science* 11 (638).

Bureau, S., D. Cozzolino, and C.J. Clark, 2019 Contributions of Fourier-transform mid infrared (FT-MIR) spectroscopy to the study of fruit and vegetables: A review. *Postharvest Biology and Technology* 148:1-14.

Calus, M.P.L., 2010 Genomic breeding value prediction: methods and procedures. *Animal* 4 (2):157-164.

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp, 2008 Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. *Genetics* 178 (1):553.

Cobb, J.N., G. DeClerck, A. Greenberg, R. Clark, and S. McCouch, 2013 Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theoretical and Applied Genetics* 126 (4):867-887.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín *et al.*, 2017 Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science* 22 (11):961-975.

Dekkers, J.C., 2004 Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J Anim Sci* 82 (328).
Dirlewanger, E., J. Quero-García, L. Le Dantec, P. Lambert, D. Ruiz *et al.*, 2012 Comparison of the genetic determinism of two key phenological traits, flowering and maturity dates, in three Prunus species: peach, apricot and sweet cherry. *Heredity* 109 (5):280-292.
Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto *et al.*, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE* 6 (5):e19379.
Endelman, J.B., 2011 Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* 4 (3).
Furbank, R.T., and M. Tester, 2011 Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science* 16 (12):635-644.
Gjuvsland, A.B., J.O. Vik, D.A. Beard, P.J. Hunter, and S.W. Omholt, 2013 Bridging the genotype-phenotype gap: what does it take? *The Journal of physiology* 591 (8):2055-2066.
Gonçalves, M.T.V., G. Morota, P.M.d.A. Costa, P.M.P. Vidigal, M.H.P. Barbosa *et al.*, 2021 Near-infrared spectroscopy outperforms genomics for predicting sugarcane feedstock quality traits. *PLOS ONE* 16 (3):e0236853.
Hayes, B.J., J. Panozzo, C.K. Walker, A.L. Choy, S. Kant *et al.*, 2017 Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theor Appl Genet* 130 (12):2505-2519.
Heffner, E.L., J.-L. Jannink, H. Iwata, E. Souza, and M.E. Sorrells, 2011 Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. *Crop Science* 51 (6):2597-2606.
Heffner, E.L., A.J. Lorenz, J.-L. Jannink, and M.E. Sorrells, 2010 Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. *Crop Science* 50 (5):1681-1690.
Houle, D., D.R. Govindaraju, and S. Omholt, 2010 Phenomics: the next challenge. *Nature Reviews Genetics* 11 (12):855-866.
Knoch, D., C.R. Werner, R.C. Meyer, D. Riewe, A. Abbadi *et al.*, 2021 Multi-omics-based prediction of hybrid performance in canola. *Theoretical and Applied Genetics* 134 (4):1147-1165.
Krause, M.R., L. González-Pérez, J. Crossa, P. Pérez-Rodríguez, O. Montesinos-López *et al.*, 2019 Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat. *G3: Genes|Genomes|Genetics* 9 (4):1231.
Kumar, S., M.C.A.M. Bink, R.K. Volz, V.G.M. Bus, and D. Chagné, 2012 Towards genomic selection in apple (Malus ×domestica Borkh.) breeding programmes: Prospects, challenges and strategies. *Tree Genetics & Genomes* 8 (1):1-14.
Kumar, S., D. Raju, R.N. Sahoo, and V. Chinnusamy, 2016 Phenomics: unlocking the hidden genetic variation for breaking the barriers in yield and stress tolerance. *Indian Journal of Plant Physiology* 21 (4):409-419.
Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14):1754-1760.
Mackay, T.F.C., E.A. Stone, and J.F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* 10 (8):565-577.
Makowsky, R., N.M. Pajewski, Y.C. Klimentidis, A.I. Vazquez, C.W. Duarte *et al.*, 2011 Beyond Missing Heritability: Prediction of Complex Traits. *PLOS Genetics* 7 (4):e1002051.
Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard, 2001 Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157 (4):1819-1829.
Nicolaï, B.M., K. Beullens, E. Bobelyn, A. Peirs, W. Saeys *et al.*, 2007 Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology* 46 (2):99-118.
Nsibi, M., B. Gouble, S. Bureau, T. Flutre, C. Sauvage *et al.*, 2020 Adoption and Optimization of Genomic Selection To Sustain Breeding for Apricot Fruit Quality. *G3: Genes|Genomes|Genetics* 10 (12):4513.
Omrani, M., M. Roth, G. Roch, A. Blanc, C.E. Morris *et al.*, 2019 Genome-wide association multi-locus and multi-variate linear mixed models reveal two linked loci with major effects on partial resistance of apricot to bacterial canker. *BMC Plant Biology* 19 (1):31.

Peterson, R.A., and J.E. Cavanaugh, 2020 Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics* 47 (13-15):2312-2327.
Pigliucci, M., 2005 Evolution of phenotypic plasticity: where are we going now? *Trends in Ecology & Evolution* 20 (9):481-486.
Rasheed, A., Y. Hao, X. Xia, A. Khan, Y. Xu *et al.*, 2017 Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. *Molecular Plant* 10 (8):1047-1064.
Rincent, R., A. Charcosset, and L. Moreau, 2018 Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics* 130 (11):2231-2247.
Ritchie, M.D., E.R. Holzinger, R. Li, S.A. Pendergrass, and D. Kim, 2015 Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* 16 (2):85-97.
Savitzky, A., and M.J.E. Golay, 1964 Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36 (8):1627-1639.
Segura, V., B.J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren *et al.*, 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44 (7):825-830.
Te Pas, M.F.W., O. Madsen, M.P.L. Calus, and M.A. Smits, 2017 The Importance of Endophenotypes to Evaluate the Relationship between Genotype and External Phenotype. *International journal of molecular sciences* 18 (2):472.
Xu, Y., X. Liu, J. Fu, H. Wang, J. Wang *et al.*, 2020 Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Communications* 1 (1):100005.
Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal, 2016 Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. *Frontiers in Genetics* 7 (151).
Zhu, J., P. Sova, Q. Xu, K.M. Dombek, E.Y. Xu *et al.*, 2012 Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. *PLOS Biology* 10 (4):e1001301.

Chapitre 6 : Discussion générale et perspectives

L'objet de cette section est de fournir un cadre de réflexion sur l'intégration de données haut débit dans le processus de sélection et de création variétale mais aussi de discuter du potentiel de valorisation de ces données en vue d'optimiser le schéma de sélection actuel de l'abricotier dans le but, à terme, de proposer des variétés toujours plus qualitatives, notamment (Figure 29). Plus particulièrement, nous visons à apporter des éléments de réponse à la question des opportunités offertes à la sélection chez cette espèce suite aux progrès technologiques accomplis en matière de séquençage et de connaissance du génome et le déploiement de méthodes de phénotypage à haut débit, telles la spectroscopie infrarouge (IR). Les perspectives de mise en œuvre de la sélection génomique et phénomique sont discutées sur la base des connaissances historiques acquises jusqu'à présent. L'enjeu est de taille puisque la filière abricot tend à passer d'une position d'observation et de sélection ciblée des caractères à celle de leur prédiction.

Deux designs expérimentaux contrastés en termes de structure génétique, de patrons de DL ainsi que de covariance entre individus ont été étudiés. Sur ces deux dispositifs, différentes variables phénotypiques ont été caractérisées par le biais des mêmes outils et protocoles, dans des conditions environnementales différentes, à la fois au regard des conditions climatiques locales et des années successives. L'évaluation des stratégies de sélection génomique et phénomique a été appliquée et optimisée en premier lieu sur le dispositif expérimental de la descendance biparentale Go×Mo, régulièrement phénotypée depuis 2004, en raison de son intérêt pour l'étude du déterminisme de la qualité des fruits. Nous avons appliqué cette démarche en deuxième lieu à un panel de diversité, afin d'évaluer le potentiel représenté par la mobilisation de ressources génomiques et spectroscopiques sur un matériel maximisant la diversité génétique et phénotypique travaillée dans le contexte du programme d'amélioration génétique de l'abricotier.

La performance des modèles de sélection génomique et phénomique a été examinée en appuyant la démarche sur un schéma de validation croisée, la qualité de la prédiction étant basée sur la déviation des valeurs prédites par rapport aux données expérimentales, qui a été quantifiée par le coefficient de corrélation de Pearson, entre différents modèles prédictifs.

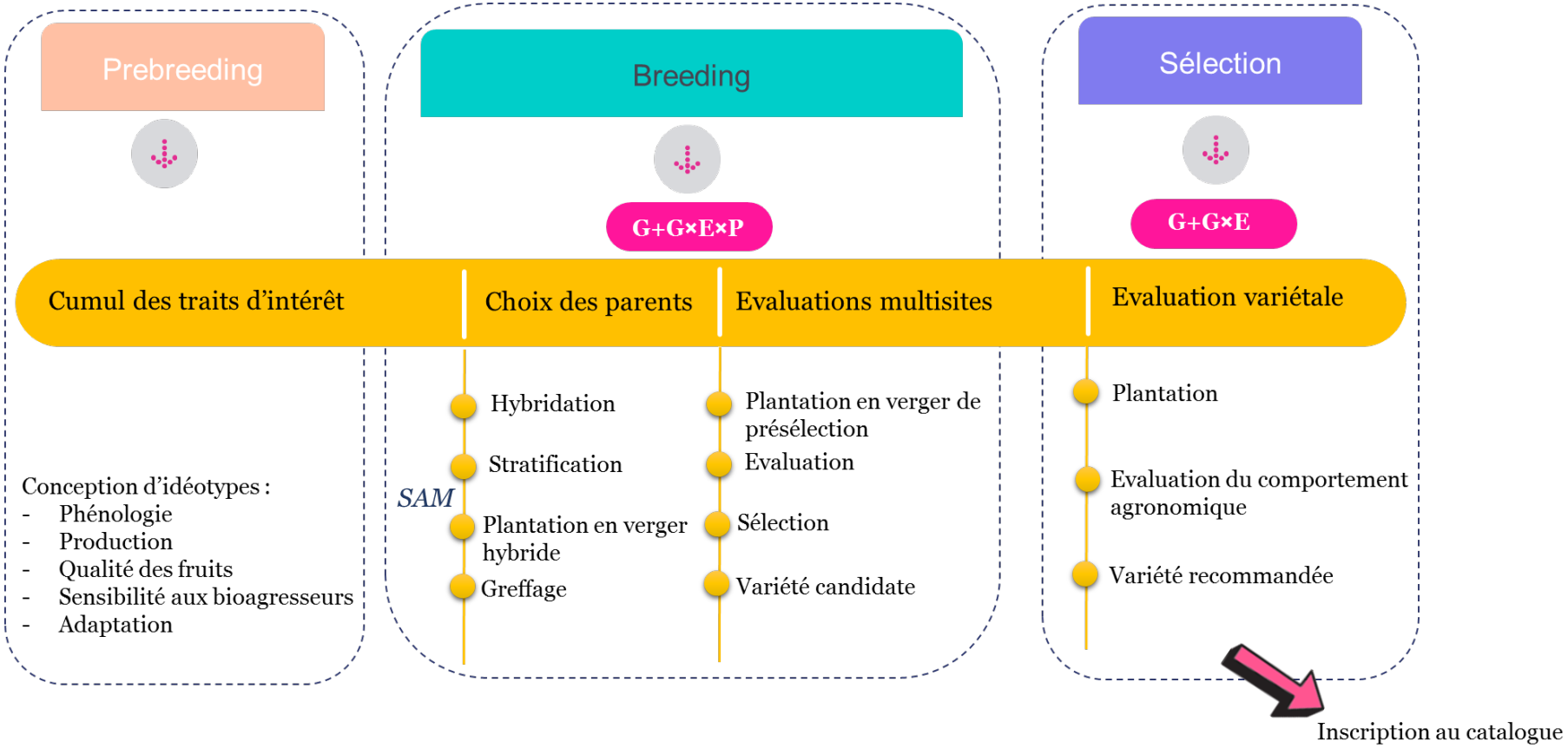


Figure 29: Etapes clés de l'amélioration génétique chez l'abricotier

6.1. Etude de l'architecture génétique des caractères cibles

Le déterminisme génétique de caractères de qualité des fruits et de phénologie a été étudié au sein de la descendance biparentale Go×Mo par cartographie d'intervalle composite, identifiant des QTLs responsables de 7.6 % à 51.2% de la variabilité phénotypique pour l'acidité titrable du fruit et la couleur du fond de l'épiderme, respectivement. Par la suite, le déterminisme génétique d'un sous-ensemble de critères intégrateurs de la qualité des fruits (mobilisés classiquement par les opérateurs de la filière) a été étudié par génétique d'association (GWAS) sur le panel de diversité, ainsi que divers caractères de sensibilité aux maladies, et de phénologie (dates de floraison et de maturité du fruit). Ces travaux ont permis de révéler les régions chromosomiques impliquées dans la variation de ces caractères cibles et d'en identifier ou d'en confirmer l'architecture génétique sous-jacente. Nous nous sommes efforcés de déterminer le contrôle génétique de ces caractères d'intérêt chez l'abricotier en vue d'interpréter les résultats obtenus par les modèles de sélection génomique et phénomique et, par la suite, d'optimiser la précision de prédiction en intégrant cette architecture comme cofacteur des modèles prédictifs.

Nos résultats constituent ainsi une base solide pour une recherche plus approfondie sur les caractères eux-mêmes. En effet, les intervalles de confiance réduits autour des marqueurs ainsi que la stabilité interannuelle par rapport au fond génétique mobilisé offrent l'opportunité d'appréhender le déterminisme des caractères investigués, ce qui ouvre la voie à une démarche gène – candidat. Dans ce but, rappelons que des analyses de liaison et d'association génétiques ont été réalisées afin d'appréhender l'architecture génétique des caractères intéressants pour la sélection, notamment la tolérance à différentes maladies. Par exemple, dans le cas du chancre bactérien, l'étude d'association phénotype – génotype a permis de mettre en évidence 22 QRLs (Quantitative Resistance Loci) dont deux majeurs identifiés sur les chromosomes 5 et 6 expliquant 26% et 43% de la variation phénotypique, respectivement (Omrani et al. 2019). Ainsi, ces QRLs orientent vers des mécanismes de résistance complémentaires qui nécessitent d'être consolidés par des études sur des panels plus larges.

Nos travaux ont permis de valider les QTLs impliqués dans les dates de floraison et de maturité chez le cerisier et l'abricotier sur le GL4 (Dirlewanger et al. 2012) dans le dispositif biparental ainsi que dans le panel de diversité, De même pour la qualité des fruits, nos résultats coïncident avec ceux obtenus par Ruiz et al. (2010) sur l'abricotier, portant sur deux descendances biparentales de 120 individus phénotypés deux années consécutives et génotypés avec des

marqueurs microsatellites, validant ainsi des QTLs identifiés sur le GL1 pour le poids du fruit, sur les GL2 et 3 pour les sucres et le GL6 pour l'acidité dans les deux dispositifs expérimentaux.

Avant l'introduction de la sélection génomique, au début des années 2000, la cartographie de QTLs avait permis la découverte de mutations causales liées à divers caractères d'importance agronomique. Cependant, son utilisation pratique dans les programmes d'amélioration génétique demeurait restreinte (Dekkers 2004; Bernardo and Yu 2007). Ainsi, en étudiant 150 programmes d'amélioration génétique d'espèces pérennes fruitières et ornementales, Byrne (2007) a montré que seulement 14% des essais utilisaient la SAM pour des objectifs de recherche, et seulement 3% mobilisaient activement des marqueurs pour orienter la sélection variétale. Parmi les raisons sous-jacentes à l'application limitée de la SAM, il convient de citer l'échelle réduite des programmes de sélection et le manque de marqueurs moléculaires disponibles. A cela s'ajoute, l'existence d'alternatives moins onéreuses telle la sélection phénotypique conventionnelle notamment pour la sensibilité à la tavelure (*Cladosporium carpophilum*) pour laquelle le phénotypage en serre s'appuyant sur des tests biologiques est efficace. Tel est le cas également de caractères mendéliens faciles à mesurer comme la couleur de la chair chez la pêche. Sur le plan économique, le coût supplémentaire associé à l'évaluation de la stabilité des QTLs dans plusieurs environnements et fonds génétiques peut représenter un obstacle majeur à la mise en œuvre de la SAM.

Ainsi, chez les arbres forestiers, les marqueurs identifiés dans les études d'association expliquent pour la plupart moins de 5 % de la variation totale du caractère, de sorte que pour les caractères complexes influencés par de nombreux QTLs à effet faible, la SAM n'est alors pas particulièrement utile ou rentable (Hospital 2009; Kainer et al. 2015). Par conséquent, le maintien de la sélection phénotypique a pu être justifié par un ratio coût – efficacité favorable en comparaison avec la SAM, car les ressources financières requises pour le développement de populations de cartographie d'effectifs suffisants sont importantes, ce qui limite l'adoption de cette technologie au niveau des programmes de sélection (Hospital 2009).

Dans un tel contexte la SAM est appliquée de manière systématique dans le cadre des programmes de prébreeding et de breeding INRAE pour introgresser les caractères de résistance à la sharka et l'autofertilité chez l'abricotier. Elle est en cours de validation à l'étape de prébreeding pour des caractères polygéniques intégrant les sensibilités au Chancre bactérien, au monilia sur fleurs et à la rouille, afin a minima de piloter le choix des géniteurs en croisements après caractérisation des ressources génétiques, ou d'implémenter la SAM en

prébreeding à des fins d'introgression ciblée de différents caractères (association de l'autofertilité, de la résistance à la sharka, au chancre bactérien, au monilia et à la rouille).

En outre, la SAM exploite le déséquilibre de liaison entre marqueurs et QTLs. Ainsi, il est recommandé d'augmenter le nombre de marqueurs afin d'augmenter la résolution de la méthode de détection et de diminuer les intervalles de confiance pour déterminer les gènes qui co-ségrégent avec les QTLs et éviter tout événement de recombinaison entre le marqueur en DL et le gène responsable de la variation phénotypique. Ceci est particulièrement pertinent chez l'abricotier qui présente une étendue de DL faible, qui décroit rapidement au-delà de 100 pb (Mariette et al. 2016; Omrani et al. 2019).

Plus généralement, la disponibilité de l'information génomique à grande échelle est cruciale pour analyser la génétique des caractères d'intérêt dans le contexte de la sélection variétale. Néanmoins, la valorisation de cet outil dans le cadre de la SAM semble être problématique notamment pour les caractères complexes. En effet, une faible proportion de l'héritabilité est représentée par les locus qui abritent les signaux d'association. De multiples causes sous-tendent le manque d'héritabilité expliquée par les QTLs identifiés. Parmi celles-ci, les interactions gène – gène, un dispositif expérimental de taille insuffisante ainsi que la présence de variants rares (Makowsky et al. 2011b). Par ailleurs, afin de réduire le biais associé à la détection de signaux faux positifs, seuls les QTLs hautement significatifs sont mobilisés dans le cadre de la SAM, ce qui biaise à la hausse les effets des QTLs choisis (Beavis 1994).

Par ailleurs, comparée à la SG, la SAM s'est montrée moins efficace en termes de progrès génétique. En effet, en comparant les gains génétiques par unité de temps issus de programmes de SG et de SAM conçus avec des budgets égaux chez deux espèces d'importance économique (blé tendre et maïs), (Heffner et al. 2010) ont montré que le gain dérivant de la SG est supérieur même avec des valeurs génétiques faibles, renforçant ainsi l'intérêt des filières pour l'intégration de cette approche en sélection variétale.

6.2. Quelle place pour la sélection génomique chez l'abricotier ?

Dans la population biparentale étudiée où il existe une proportion substantielle d'allèles en commun, la SG évaluée par le biais de la capacité prédictive de ses modèles s'est révélée prometteuse au vu d'une précision moyenne variant de 0.31 à 0.77, en fonction des caractères et du modèle de prédiction. La performance des modèles de sélection génomique s'explique non seulement par l'identité par descendance (IBD) mais également par l'étendue du DL entre SNPs et QTLs qui coségrégent dans les partitions d'entraînement et de validation (Crossa et al.

2014) de la population d'étude. Ainsi, dans l'étude de (Schopp et al. 2017) portant sur le potentiel de la SG dans des familles biparentales apparentées (FBA) renfermant des demi- et pleins-frères, ou non apparentées (FBN), les auteurs ont souligné que la précision de prédiction inter-familles pouvait atteindre des niveaux très faibles lorsque les variants causaux qui ne ségrégent pas dans la population d'entraînement représentent une proportion importante de la variance génétique par rapport à la population de validation. Ils ont également conclu que l'entraînement du modèle de SG dans des FBA entraîne généralement des précisions de prédiction stables alors qu'un niveau important d'incertitude est rencontré dans les prédictions inter-familles.

Outre la proximité génétique, l'étendue du DL présente une forte incidence sur la capacité prédictive des modèles de SG. Dans ce sens, la prédiction génomique dans des descendances biparentales présente un cadre plus favorable en comparaison avec des panels de diversité. Le nombre de marqueurs requis pour capturer l'information génétique utile dans des populations biparentales est par conséquent plus faible que celui des panels de diversité génétique. Ceci est tributaire d'un nombre limité d'évènements de recombinaison ayant lieu dans le cadre biparental où les individus partagent des régions génomiques de grande taille alors que les événements plus nombreux de recombinaison ancestrale capturés dans un panel de diversité nécessitent un marquage moléculaire plus dense pour atteindre une capacité prédictive équivalente.

Bien que les prédictions génomiques intra-familles soient plus prometteuses que les prédictions inter-familles, dans notre étude sur le panel de diversité de l'abricotier, où le pourcentage de segments chromosomiques partagés étant substantiellement plus faible que celui du dispositif biparental, la performance du modèle RR-BLUP en termes de précision moyenne a varié de 0.02 pour la sensibilité au chancre bactérien jusqu'à 0.71 pour la date de maturité du fruit, entre ces deux dispositifs. Ces résultats soulignent la capacité très variable du modèle de sélection génomique à prédire la performance agronomique d'individus supposés non phénotypés pour les caractères cibles.

Particulièrement pour le chancre, l'année de mesure (2020) a coïncidé avec des conditions climatiques non propices à l'expression des symptômes, engendrant une faible variabilité génétique et par conséquent une faible héritabilité. Ceci s'est traduit par une très faible précision du modèle de SG, soulignant l'importance de l'intégration de la variabilité environnementale dans le cadre de prédiction. Des évaluations multisites et pluriannuelles semblent être

essentielles à l'optimisation de la précision par le biais de l'introduction de covariables environnementales dans les modèles de SG.

Par ailleurs, comme déjà mentionné, la précision de la prédiction dépend de plusieurs facteurs, parmi lesquels le nombre de marqueurs moléculaires. La précision est conditionnée par une couverture suffisante du génome, avec comme situation idéale, un marqueur par 'haplobloc'. Une densité insuffisante de marqueurs se traduit par un déficit de précision du modèle, alors qu'une densité croissante de SNPs améliore sa performance, ceci jusqu'à un plateau de précision maximale, au-delà duquel la capacité prédictive se détériore suite à la surestimation des effets des marqueurs. En effet, la précision des modèles de SG est fortement influencée par le DL entre locus aux marqueurs et QTL. Ainsi, l'augmentation du nombre de marqueurs permet d'augmenter la probabilité que les QTL soient en DL avec les marqueurs flanquants. Cependant, le nombre maximal de marqueurs à partir duquel la précision de prédiction se dégrade peut être déterminé en fonction du nombre de segments chromosomiques indépendants (Goddard 2009; Daetwyler et al. 2010b).

6.3. Quel modèle pour la sélection génomique ?

Les recherches actuelles sur la base génomique des caractères complexes d'intérêt agronomique illustre l'importance des méthodes statistiques en génétique. En corollaire, étant donné que les modèles mathématiques présentent un élément clé de la génétique statistique (Charlesworth 2019), le choix du modèle de sélection génomique revêt une importance cruciale pour la modélisation des effets des marqueurs. En effet, leur capacité prédictive est fortement influencée par l'écart entre l'hypothèse du modèle de prédiction (portant sur la loi de distribution et la variance des effets des marqueurs) et l'architecture génétique du caractère cible. Dans la présente étude, pour divers caractères polygéniques de qualité des fruits, des QTLs expliquant un faible pourcentage de variance phénotypique (< 10%) ont été détectés. C'est le cas de l'acidité titrable (AT), et particulièrement de la teneur en glucose, caractère pour lequel un seul QTL a été identifié par analyse de liaison génétique, expliquant seulement 10.8% de la variabilité et laissant place à une forte proportion d'héritabilité manquante. Pour ces caractères complexes, aucune différence significative n'a été constatée entre les modèles bayésiens, tandis que le modèle RR-BLUP se montrait le plus robuste. En revanche, pour des architectures génétiques dépendant de QTLs à effets modérés à forts, le modèle Bayes B s'est révélé être supérieur, ce qui est en adéquation avec son hypothèse principale postulant qu'une proportion $(1 - \pi)$ de marqueurs est en DL avec les variants causaux. Nous avons montré que tel était aussi le cas de la couleur de fond et de la production éthylénique du fruit, caractères

pour lesquels la variance phénotypique expliquée par les QTLs détectés était de 51% et 43%, respectivement. Malgré son hypothèse jugée irréaliste par plusieurs auteurs, spécifiquement pour des caractères dont l'architecture génétique s'écarte considérablement de son postulat par rapport à une contribution égale des effets des marqueurs (Daetwyler et al. 2010b; Bernardo 2014), le modèle RR-BLUP s'est montré le plus performant, pour des caractères complexes pour lesquels le modèle infinitésimal semble être réaliste. Au-delà du modèle statistique, la précision de prédiction se décline en deux composantes. Une composante est due à l'apparentement entre les individus et la deuxième est liée au DL entre SNPs et QTLs (Habier et al. 2007). Par conséquent, le modèle le plus performant est celui qui parvient à capturer les deux composantes. Tel est le cas de Bayes B. Ces résultats démontrent clairement l'intérêt de tester différents modèles prédictifs et d'adapter celui-ci au caractère phénotypique étudié en vue de sa prédiction. Cela s'avère relativement lourd dans un schéma de sélection mais doit être mis en regard du gain génétique potentiel résultant de la mise en œuvre d'une approche plus fine et adaptée.

Au-delà de ce travail, il serait pertinent de perfectionner les modèles de SG, en s'appuyant sur la caractérisation phénotypique et génotypique d'un sous-ensemble de candidats à la sélection pour éviter la dégradation du pouvoir prédictif des équations de SG suite à la dégradation du DL au fil des générations. Rappelons que cette dégradation est due à l'accumulation d'évènements de recombinaison entre les populations d'entraînement et candidate à la sélection. De plus, une augmentation de la densité de marqueurs peut être intéressante dans le cadre du modèle RR-BLUP afin de préserver la précision de la prédiction sur le long terme, notamment lorsque le nombre de générations entre la population d'entraînement et la population candidate à la sélection augmente (Solberg et al. 2009). Concernant Bayes B, maximiser le DL entre SNPs et QTLs peut s'effectuer via l'utilisation d'haplotypes, afin de suivre la transmission des régions chromosomiques IBD au fil des générations.

6.4. Valorisation de l'architecture génétique dans le cadre de la SG

Nous nous sommes penchés sur l'optimisation des modèles de prédiction par le biais de l'exploitation des connaissances sur l'architecture génétique des traits sous-jacents, ainsi que par la modélisation multivariée. Pour la valorisation de l'information apportée par l'analyse de liaison génétique, nous avons mis l'accent sur les marqueurs significativement liés aux QTLs en leur attribuant plus de poids (variance) que les marqueurs à effet faible. La réponse des modèles de prédiction génomique à l'intégration de l'information a priori sur l'architecture génétique dépend fortement du pourcentage de variation phénotypique expliquée par les QTLs

inclus dans le modèle, comme cela était attendu. Le gain en précision est d'autant plus prononcé que les QTLs pris en compte en effets fixes expliquent une proportion substantielle de la variabilité. A titre d'exemple, c'est le cas de la couleur du fond du fruit (36% avec le modèle statistique BRR). Pour d'autres caractères, tel le poids du fruit et sa teneur en glucose, la pondération du modèle de prédiction par l'architecture génétique du caractère a entraîné une diminution de la précision de prédiction ; ces deux caractères pour lequel les QTLs identifiés ne déterminent qu'une faible proportion de la variance phénotypique (15.51% et 10.77%, respectivement), peuvent être considérés comme fortement polygéniques. Or il a été démontré qu'attribuer plus de poids à un QTL à effet faible a un impact négatif sur la précision de prédiction (Kainer et al. 2018), c'est pourquoi, une représentation détaillée de l'architecture génétique du caractère par les marqueurs intégrés dans le modèle est cruciale pour garantir des prédictions génomiques précises. Cela nécessite également de se projeter dans la mise en œuvre d'un processus pas à pas d'implémentation de la sélection génomique et d'atteinte de son plein potentiel.

6.5. Quel est l'intérêt de l'approche multivariée ?

Durant les phases initiales de ce travail, une question importante pour l'amélioration concomitante de divers caractères d'intérêt agronomique était la suivante : « Comment valoriser les ressources facilement accessibles et économiquement abordables pour améliorer des caractères complexes et difficiles à phénotyper ?». Pour répondre à cette question, nous avons exploré la modélisation multivariée de la sélection génomique et plus particulièrement les modèles bivariés. Nous avons démontré que les modèles génomiques ciblant deux caractères dotés d'une forte corrélation génétique fournissent des prédictions plus précises que celles obtenues par le biais de modèles univariés. Néanmoins, la modélisation bivariée s'est révélée défavorable lorsque la corrélation entre caractères était faible. C'est le cas des modèles utilisant la couleur de fond pour prédire les teneurs en sucres et en acides organiques. Nos résultats sont donc en adéquation avec ceux de plusieurs auteurs rapportant une optimisation de la précision de prédiction dans le cadre de la prédiction multivariée. Par exemple chez le blé tendre, Michel et al. (2018) ont montré que la mobilisation d'informations préalables sur la teneur en protéines a entraîné une amélioration de la précision de prédiction pour l'évaluation des caractéristiques rhéologiques de la pâte, qui implique des opérations de phénotypage consommatrices de main-d'œuvre et coûteuses. Le gain en précision est d'autant plus important que le caractère à prédire est moins fortement héritable que le proxy. Ceci est illustré dans le cas de l'abricot par le trait phénotypique de la teneur en saccharose (H^2=0.56) qui a pu être prédit par le biais du proxy

indice de réfraction (H^2=0.9). Le gain de précision issu de la mobilisation de l'IR pour prédire le saccharose a été de 25% et 23% pour les modèles bivariés basés sur les valeurs phénotypiques et génomiques, respectivement. De même, pour les teneurs en glucose et fructose de l'abricot, le gain a atteint 9% et 1% avec les valeurs phénotypiques et 16% et 8%, respectivement avec les valeurs génomiques du proxy pour lequel les corrélations génétiques étaient de 0.7 pour le glucose et de 0.9 pour le fructose. Tel est le cas aussi pour le caractère AT, utilisé comme proxy pour informer les modèles de prédiction univariés de l'acide citrique, pour lequel le gain en précision a atteint 2% avec les valeurs phénotypiques du proxy et 25% avec ses valeurs génomiques. Quant à l'acide malique, bien que la corrélation génétique avec l'AT soit négative et forte (-0.7), elle s'est traduite par une perte en précision à la suite de l'intégration de ce proxy dans le modèle prédictif de l'acide malique.

Il convient de souligner que même des valeurs de corrélation génétique fortement négatives entraînent une optimisation de la prédiction des caractères focaux. Ceci peut être illustré par le cas du glucose et l'AT de l'abricot, qui présentent une corrélation de -0.9, pour lesquels un gain en précision de 6% et 7% a été observé en utilisant les valeurs phénotypiques et génomiques du proxy, respectivement. A l'inverse, de faibles corrélations génétiques entre les caractères focal et proxy résultent en une faible variation de la précision, tel qu'attendu. C'est le cas du caractère lié à la production éthylénique pour lequel la mobilisation de la couleur du fond comme proxy s'est traduite par un gain de précision de 2% seulement, pour une valeur de corrélation génétique égale à 0.1.

Par ailleurs, nous avons pu démontrer que le modèle bivarié basé sur un indice de sélection intégrant la valeur génomique du proxy au lieu de sa valeur phénotypique s'est révélé plus performant. Ceci est en accord avec les résultats de Michel et al. (2018) qui ont souligné également que la méthode de l'indice de sélection est moins exigeante en termes de calcul. De plus, elle n'est pas sujette aux problèmes de convergence qui se produisent souvent avec les modèles multivariés non structurés. L'utilisation de cette méthode s'est traduite par un gain de précision moyenne estimé à 10.25%.

La forte corrélation entre les caractères peut être attribuée à la liaison génétique entre les QTLs qui sous-tendent ces caractères, ou par des effets de pléiotropie. Dans ce sens, nous avons pu identifier par cartographie génétique sur la descendance Go×Mo, plusieurs colocalisations entre QTLs, par exemple entre ceux détectés pour l'IR et la teneur en saccharose, et ceux mis en

évidence pour l'AT et la teneur acide citrique, ce qui consolide nos résultats par rapport à l'approche bivariée.

Une telle stratégie d'optimisation pourrait être implémentée pour prédire des caractères coûteux ou complexes à mesurer, afin d'aider les décisions de sélection. Par ailleurs, cette approche permettrait d'élargir la population d'entraînement en SG, en phénotypant spécifiquement les caractères les plus accessibles en phénotypage haut-débit, permettant par conséquent d'augmenter l'intensité de sélection. Ce cadre de prédiction offrirait également l'opportunité d'optimiser la capacité prédictive des modèles en incluant des caractères corrélés.

6.6. Quelle place pour la sélection phénomique chez l'abricotier ?

L'une des questions majeures que soulève l'utilisation de la SP est de savoir « Comment concevoir des stratégies d'amélioration capables d'explorer la carte génotype – phénotype ? », notamment dans le contexte du changement climatique qui ajoute un degré de complexité supplémentaire.

La carte qui lie le génotype au phénotype permet de combler ce qui les sépare, en appréhendant la complexité des interactions entre la génétique et les réponses physiologiques. En effet, le phénotype est déterminé par la contribution conjointe de nombreux facteurs causaux et représente donc le résultat de dynamiques complexes incluant des dimensions génétiques et environnementales (Pigliucci 2010; Gjuvsland et al. 2013). Par conséquent, contrairement à la SG où l'information génétique occupe une place privilégiée dans l'interprétation de la variation phénotypique, le concept de carte génotype-phénotype repose sur la contribution de différentes données biologiques de grande dimension (-omiques), notamment la génomique, la dominance intralocus et l'épistasie interlocus. Une telle perspective permettrait d'améliorer notre compréhension des mécanismes qui sous-tendent les phénotypes et d'appréhender les interactions entre la génétique et l'environnement (Alberch 1991; Houle et al. 2010; Te Pas et al. 2017).

La spectroscopie infrarouge semble être un outil performant permettant de prendre en compte de manière relativement rapide et peu coûteuse les contributions au phénotype de la génétique, de l'environnement et de leurs interactions. Particulièrement, les spectres dans le PIR se sont révélés capables de capturer indirectement la variation endophénotypique ainsi que la covariance génétique entre individus. En corollaire, ils permettent de capturer l'échantillonnage mendélien responsable de la variation des caractères quantitatifs ainsi que le DL, qui présentent les piliers de la prédiction dans le cadre de la SG.

L'avènement des technologies de génotypage des SNP à haut débit, par exemple les puces SNP, le génotypage par séquençage (GBS) et le reséquençage du génome entier, a depuis plusieurs années facilité l'accès au génotypage à haute densité et à moindre coût. Néanmoins, malgré la baisse spectaculaire des coûts de génotypage, ces derniers restent élevés et dépassent les coûts d'acquisition de spectres de réflectance sur les échantillons végétaux, notamment quand ils sont issus de programmes de sélection multi-sites. Comme le phénotypage reste considéré comme un goulot d'étranglement majeur pour le progrès génétique (Cobb et al. 2013; Araus and Cairns 2014; Kumar et al. 2016), la SP peut offrir une alternative à la SG en vue de l'accélérer. Dans notre étude, l'évaluation des modèles de SP appliquée à la descendance biparentale Go×Mo, basée sur l'exploitation de 2307 spectres acquis dans le PIR et de 1736 spectres dans le MIR a montré la supériorité de la SP par rapport à la SG pour la majorité des caractères étudiés, qu'ils soient ou non liés à la composition biochimique des échantillons analysés (fruits). Plus particulièrement, le modèle fondé sur les spectres MIR obtenus sur purées de fruits s'est révélé supérieur en capacité prédictive à celui basé sur les spectres PIR ou la SG, pour 7 caractères sur 12 portants sur la qualité des fruits et sur les caractères de phénologie. Concernant les prédictions relatives au panel de diversité de l'abricotier, la SP s'est montrée supérieure à la SG pour la prédiction de l'IR et l'AT avec un niveau important de précision très élevé pour les deux régions spectrales : PIR sur feuilles ou MIR sur pulpes de fruits. Cependant, pour les caractères liés à la sensibilité aux maladies, le modèle de SG a présenté une performance prédictive supérieure à celle de la SP, étant donné que l'interaction $G \times E$ n'a pas été intégrée dans le cas des caractères de sensibilité. Cette information est cruciale pour les modèles phénomiques vu que le phénome représente l'interaction entre les différents niveaux biologiques modulée par la contribution des facteurs environnementaux. Ainsi, la compréhension de la variation des traits complexes nécessite des connaissances non seulement sur les variations génomiques, mais aussi sur les effets environnementaux qui affectent l'expression du génome (Te Pas et al. 2017).

Dans notre travail, la performance du modèle de SP s'est avérée tributaire de la composante $G \times E$, laquelle a été inclue dans le cadre des prédictions dans la descendance biparentale, ce qui n'a pas pu être le cas pour le panel de diversité, expliquant ainsi la moindre précision des modèles de prédiction dans les deux dispositifs expérimentaux où ce panel était étudié. Il serait donc pertinent d'élargir la gamme des conditions environnementales dans lesquelles les génotypes sont caractérisés, et de faire appel à des dispositifs expérimentaux pluriannuels et multi-sites afin de capturer un large éventail de variabilité génétique afin d'intégrer l'interaction $G \times E$ dans les prédictions. Par ailleurs, la gestion des pratiques

culturales est cruciale pour optimiser la diversité prise en compte dans les itinéraires techniques conçus pour répondre aux différents enjeux de sélection. Ainsi, l'intégration des interactions $Génotype \times Environnement \times Pratique\ culturale$ dans le schéma de sélection permettrait de comprendre et, si possible, de prédire le comportement agronomique des variétés en réponse au changement des conditions pédoclimatiques et des itinéraires techniques. A titre d'exemple, l'éclaircissage des fruits a un impact sur la qualité des fruits produits et sur la variabilité génétique, donc sur la précision de la prédiction et la réponse à la sélection pour ce caractère.

Par ailleurs, l'optimisation de la SP grâce à l'intégration de cofacteurs de l'architecture génétique des caractères s'est montrée efficace pour la plupart des caractères dans les deux designs expérimentaux supports dans le cadre de la thèse. Par conséquent, il apparait important de complémenter l'information spectrale par des connaissances sur le contrôle génétique des caractères en vue d'optimiser la précision de la sélection.

6.7. Place potentielle de la SG et SP

La plus-value apportée par l'intégration d'une stratégie de sélection haut-débit basée sur les marqueurs moléculaires ou sur les spectres dans le schéma de sélection de l'abricotier ayant été discutée, nos résultats, en référence à la moyenne et à la variation de la précision de prédiction des stratégies de SG et SP appliquées à des dispositifs contrastés, peuvent interroger sur les démarches de sélection en vue de leur optimisation. Ces interrogations devraient notamment conduire à réfléchir sur :

- l'allocation des ressources entre les partitions d'entraînement et de validation,

- la conception de populations d'entraînement incluant une large gamme de diversité génétique afin de garantir la robustesse et la stabilité des prédictions génomiques et phénomiques,

- le déploiement d'un schéma de sélection fondé sur une approche intégrative des deux stratégies.

Dans cette optique, la SG et SP peuvent être implémentées en amont du processus de sélection dans le cadre du prébreeding, tel que suit (Figure 30) :

- Pour caractériser et gérer les ressources génétiques :
    - Dans des approches ciblées portant en même temps sur plusieurs caractères.
    - En vue de la recherche de combinaisons alléliques favorables à introgresser dans la population candidate à la sélection en fonction des connaissances sur le contrôle génétique des caractères cibles, le degré d'apparentement et l'étendue du DL au sein

du panel de diversité. Prenons l'exemple de la sensibilité aux maladies, il s'avère crucial d'identifier et d'intégrer des composantes de résistance au cortège de bioagresseurs les plus préoccupants en termes de dégâts occasionnés affectant le potentiel de production. Pour ce faire, il sera indispensable de mettre en place des dispositifs expérimentaux multi-sites afin d'intégrer la composante d'interaction G×E et d'évaluer la stabilité et donc la durabilité de la résistance. Dans ce contexte, la caractérisation de la sensibilité variétale pourra être effectuée par le biais de la spectroscopie infrarouge, qui offre un cadre pratique grâce à la facilité d'acquisition des données phénomiques.

- Pour enrichir la diversité génétique du germplasm afin d'alimenter l'implémentation de la SG et de la PS, par introduction de matériel exotique (espèces apparentées) et éviter les effets de consanguinité.
- Pour rechercher les déterminants génétiques des traits d'intérêt méconnus avant leur déploiement en collection ou dans le cadre de programmes d'introgression

Les variétés retenues devront cumuler le plus de caractères d'intérêt y compris la résistance à un cortège d'agents pathogènes, la qualité des fruits et l'auto-fertilité.

Au-delà du prébreeding, la SG et SP doivent être adoptées à un stade de développement précoce des candidats à la sélection permettant ainsi de raccourcir la longueur du cycle de breeding en les évaluant par le biais du génotypage ou la spectroscopie infrarouge au stade plantule, par exemple.

Compte tenu de la facilité d'acquisition des données spectrales, la détermination de la fréquence de réestimation des valeurs génétiques des candidats à la sélection pourrait être effectuée à un stade précoce du développement de l'arbre. En effet, une sélection précoce est susceptible de permettre une réduction de la durée du cycle de sélection. Par conséquent, la SP pourrait être intégrée dans les programmes de sélection comme une activité de routine pour cribler les candidats à la sélection sans investissements importants.

Nous recommandons l'utilisation du cadre multi-trait pour sélectionner pour des caractères complexes. Ainsi celle des caractères proxies tels que l'IR pour sélectionner pour la teneur en sucres (saccharose, glucose et fructose), ou l'AT pour la teneur en acides organiques (citrique et malique), pourrait optimiser la réponse à la sélection. Nous recommandons également, l'intégration des prédictions univariées des proxies sous la forme d'un indice de sélection pour prédire les caractères cibles. Plus important encore, la valorisation de la spectroscopie pour mesurer les caractères proxies tels que l'IR et l'AT dont les modèles de prédiction basés sur le

PIR et le MIR se sont montrés performants avec des précisions moyennes qui s'approchent de 1 (0.95 et 0.72 pour la descendance biparentale et 0.89 et 0.81 pour le panel, respectivement). En corollaire, l'utilisation conjointe de la spectroscopie infrarouge avec l'approche multivariée semble être prometteuse notamment pour IR et AT qui se sont révélés être des proxies performants dans les deux domaines spectraux de l'infrarouge ainsi que dans les deux dispositifs expérimentaux (population biparentale et panel de diversité).

Les enjeux de la filière abricot préconisent un changement de paradigme des méthodes conventionnelles vers des schémas axés sur la prédiction des performances agronomiques tirant parti des avancées technologiques génomiques et phénomiques. Il est donc potentiellement possible d'appréhender la complexité des dispositifs intégrant des interactions Génotype × Environnement × Pratique culturale. Ces interactions représentent une source majeure de variabilité du comportement agronomique des génotypes que la filière se doit de comprendre, décrire et utiliser afin de garantir et optimiser les rendements et la qualité des produits agronomiques tout en permettant une réduction sensible de l'usage des pesticides. Avec des outils moléculaires et spectraux haut-débit, il est également possible d'assurer une gestion optimisée du verger dans le cadre d'une diversité de contextes pédoclimatiques et de sélectionner les combinaisons G × E × P les plus favorables dans le but d'accélérer le progrès génétique chez l'abricotier. Dans ce sens, ces outils peuvent offrir la possibilité de mettre en place des expérimentations dans plusieurs sites incluant une large variabilité de conditions représentatives des principales zones de production d'abricot. Par ailleurs, ces outils permettront également d'intégrer une combinaison de plusieurs performances agronomiques d'intérêt et de concevoir des génotypes élites, après exploration de la réponse des individus à une combinaison de facteurs dans des itinéraires techniques diversifiés et élimination du matériel végétal ne comportant pas les caractères intéressants pour la sélection. En outre, le déploiement des approches génomique et phénomique haut-débit permettrait de répondre à de nouveaux enjeux par le biais de l'intégration de nouveaux caractères d'intérêt agronomique via la caractérisation des ressources génétiques dans des environnements sujets aux stress abiotiques (hydrique, thermique et azoté) susceptibles d'impacter le potentiel de production et la performance des génotypes. De surcroît, des caractères complexes faiblement héritables ainsi que des caractères s'exprimant tardivement au cours du cycle de sélection pourraient également être intégrés dans le schéma de sélection de l'abricotier pour répondre aux défis posés par le contexte climatique afin de satisfaire à terme les attentes des différents acteurs de la filière.

Cependant, un déploiement raisonnable des deux approches génomique et phénomique est fortement recommandé afin d'éviter toute érosion de la diversité génétique du germplasm. En effet, l'intensification des pratiques culturales peut générer des pressions de sélection susceptibles d'engendrer une perte rapide de la variabilité, et par conséquent la perte d'allèles rares et le développement de la dépression de consanguinité, ce qui est préjudiciable au progrès génétique. Les risques liés à l'érosion génétique seront d'autant plus importants que la longueur du cycle de sélection sera faible. En corollaire, il sera fortement recommandé de mettre en place un suivi constant de la diversité génétique et, le cas échéant, d'apporter de la diversité aux programmes de sélection en exploitant le patrimoine génétique inhérent à l'espèce. Dans ce sens, afin d'éviter de compromettre le gain génétique sur le long terme, des croisements de génotypes élites avec des ressources génétiques présentant des caractéristiques phénotypiques contrastées seront indispensables.

Conceptuellement, la SP semble offrir des avantages économiques prometteurs par rapport à la SG et à la SAM, car les coûts de l'acquisition spectrale sont inférieurs à ceux du génotypage. Il est donc primordial d'évaluer la performance économique des deux stratégies pour un large éventail de scénarios de sélection impliquant la variation d'outils de prédiction, de caractères et de matériel génétique.

Il serait enfin intéressant d'évaluer la pertinence de l'adossement de la sélection phénomique aux progrès d'autres technologies de phénotypage utilisant les capteurs, notamment l'imagerie multi- ou hyperspectrale afin d'intégrer la variabilité spatio-temporelle des conditions environnementales et d'optimiser le débit et la résolution des entrées (input) des modèles phénomiques. Ceci rentre dans le cadre de l'envirotypage reposant sur la caractérisation détaillée des environnements par différents facteurs non génétiques (climat, sol et facteurs biotiques) qui représentent une forte incidence sur la variabilité phénotypique. La nécessité de l'envirotypage s'impose en plus du phénotypage et du génotypage pour prédire non seulement les phénotypes d'intérêt mais aussi les interactions G×E.

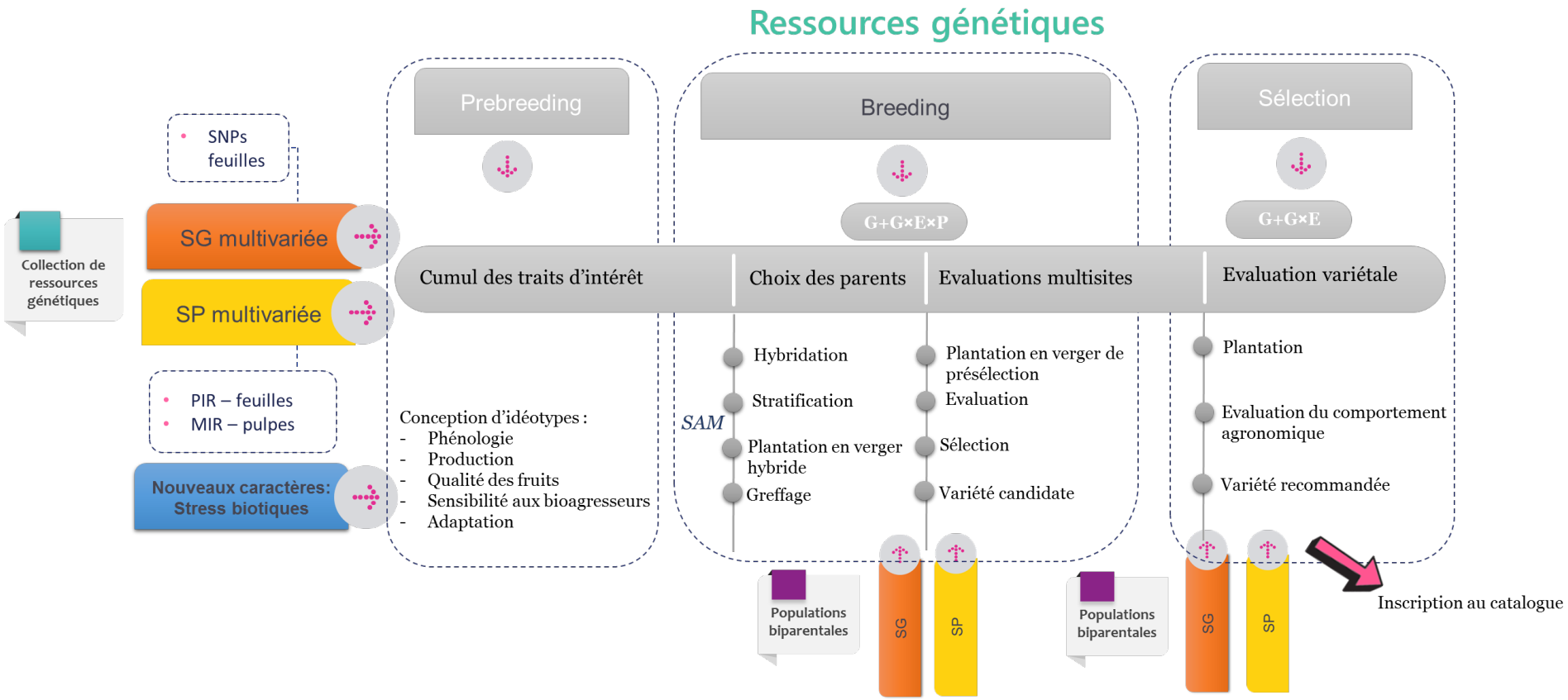


Figure 30: Place potentielle des deux stratégies de sélection phénomique et génomique

Références bibliographiques

Adami, M., P. De Franceschi, F. Brandi, A. Liverani, D. Giovannini *et al.*, 2013 Identifying a Carotenoid Cleavage Dioxygenase (ccd4) Gene Controlling Yellow/White Fruit Flesh Color of Peach. *Plant Molecular Biology Reporter* 31 (5):1166-1175.

Agreste, 2020 Infos rapides Abricot, Agreste Conjoncture n°2020-091.

Akdemir, D., W. Beavis, R. Fritsche-Neto, A.K. Singh, and J. Isidro-Sánchez, 2018 Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity* 122:672-683.

Akdemir, D., and J. Isidro-Sánchez, 2019 Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports* 9 (1).

Alberch, P., 1991 From genes to phenotype: dynamical systems and evolvability. *Genetica* 84 (1):5-11.

Alqudah, A.M., A. Sallam, P. Stephen Baenziger, and A. Börner, 2020 GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley – A review. *Journal of Advanced Research* 22:119-135.

Anderson, J.A., R.W. Stack, S. Liu, B.L. Waldron, A.D. Fjeld *et al.*, 2001 DNA markers for Fusarium head blight resistance QTLs in two wheat populations. *Theoretical and Applied Genetics* 102 (8):1164-1168.

Aranzana, M.J., V. Decroocq, E. Dirlewanger, I. Eduardo, Z.S. Gao *et al.*, 2019a Prunus genetics and applications after de novo genome sequencing: achievements and prospects. *Horticulture Research* 6 (1):58.

Aranzana, M.J., V. Decroocq, E. Dirlewanger, I. Eduardo, Z.S. Gao *et al.*, 2019b Prunus genetics and applications after de novo genome sequencing: achievements and prospects. *Horticulture Research* 6 (1).

Araus, J.L., and J.E. Cairns, 2014 Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science* 19 (1):52-61.

Araus, J.L., S.C. Kefauver, M. Zaman-Allah, M.S. Olsen, and J.E. Cairns, 2018 Translating High-Throughput Phenotyping into Genetic Gain. *Trends in Plant Science* 23 (5):451-466.

Arruda, M.P., A.E. Lipka, P.J. Brown, A.M. Krill, C. Thurber *et al.*, 2016 Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (Triticum aestivum L.). *Molecular Breeding* 36 (7):84.

Arumuganathan, K., and E.D. Earle, 1991 Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* 9 (4):208-218.

Audergon, J.M., C. Castelain, G. Morvan, and M.G. Chastelliere, 1989 Apricot varietal sensibility and genetic variability to apricot chlorotic leaf roll disease, pp. 205-214. International Society for Horticultural Science (ISHS), Leuven, Belgium.

Balding, D.J., 2006 A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7 (10):781-791.

Banik, D., 2019 Achieving Food Security in a Sustainable Development Era. *Food Ethics* 4 (2):117-121.

Barbin, D., A.L. Felicio, D.-W. Sun, S. Nixdorf, and E. Hirooka, 2014 Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview. *Food Research International* 61.

Barnes, R.J., M.S. Dhanoa, and S.J. Lister, 1989 Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy* 43 (5):772-777.

Bartholomé, J., M. Bink, J. Heerwaarden, E. Chancerel, C. Boury *et al.*, 2016 Linkage and association mapping for two major traits used in the maritime pine breeding program: height growth and stem straightness. *PloS one* 11:e0171439.

Barton, N.H., 2017 How does epistasis influence the response to selection? *Heredity* 118 (1):96-109.

Bassi, D., F. Bartolozzi, and E. Muzzi, 2006 Patterns and heritability of carboxylic acids and soluble sugars in fruits of apricot (*Prunus armeniaca* L.). *Plant Breeding* 115:67-70.

Bassi, F.M., A.R. Bentley, G. Charmet, R. Ortiz, and J. Crossa, 2016 Breeding schemes for the implementation of genomic selection in wheat (Triticum spp .). *Plant Science* 242:23-36.

Bates, D., M. Mächler, B. Bolker, and S. Walker, 2014a Fitting linear mixed-effects models using lme4. *ArXiv e-prints* arXiv:1406:1-48.

Bates, D., M. Mächler, B. Bolker, and S. Walker, 2014b *Package Lme4: Linear Mixed-Effects Models Using Eigen and S4*.

Batista-Silva, W., V.L. Nascimento, D.B. Medeiros, A. Nunes-Nesi, D.M. Ribeiro *et al.*, 2018 Modifications in Organic Acid Profiles During Fruit Development and Ripening: Correlation or Causation? *Frontiers in Plant Science* 9:1689.

Beaulieu, J., T. Doerksen, S. Clément, J. MacKay, and J. Bousquet, 2014 Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity* 113 (4):343-352.

Beavis, W., 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies, pp. 266 in *Proceedings of the forty-ninth annual corn and sorghum industry research conference*. Chicago, IL.

Bernardo, R., 2014 Genomewide Selection when Major Genes Are Known. *Crop Science* 54 (1):68-75.

Bernardo, R., 2020 Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity* 125 (6):375-385.

Bernardo, R., and J. Yu, 2007 Prospects for Genomewide Selection for Quantitative Traits in Maize. *Crop Science* 47.

Bhatta, M., L. Gutierrez, L. Cammarota, F. Cardozo, S. Germán *et al.*, 2020 Multi-trait Genomic Prediction Model Increased the Predictive Ability for Agronomic and Malting Quality Traits in Barley (Hordeum vulgare L.). *G3: Genes|Genomes|Genetics* 10 (3):1113.

Blanco, M., and I. Villarroya, 2002 NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry* 21 (4):240-250.

Bonnafous, F., G. Fievet, N. Blanchet, M.C. Boniface, S. Carrère *et al.*, 2018 Comparison of GWAS models to identify non-additive genetic control of flowering time in sunflower hybrids. *Theor Appl Genet* 131 (2):319-332.

Bourguiba, H., J.-M. Audergon, L. Krichen, N. Trifi-Farah, A. Mamouni *et al.*, 2012 Loss of genetic diversity as a signature of apricot domestication and diffusion into the Mediterranean Basin. *BMC Plant Biology* 12 (1):49.

Bourguiba, H., I. Scotti, C. Sauvage, T. Zhebentyayeva, C. Ledbetter *et al.*, 2020 Genetic Structure of a Worldwide Germplasm Collection of Prunus armeniaca L. Reveals Three Major Diffusion Routes for Varieties Coming From the Species' Center of Origin. *Frontiers in Plant Science* 11 (638).

Broman, K., H. Wu, S. Sen, and G. Churchill, 2003a R/QTL: QTL mapping in experimental crosses. *Bioinformatics* 19:889-890.

Broman, K.W., H. Wu, S. Sen, and G.A. Churchill, 2003b R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19 (7):889-890.

Brummell, D.A., V. Dal Cin, C.H. Crisosto, and J.M. Labavitch, 2004 Cell wall metabolism during maturation, ripening and senescence of peach fruit. *J Exp Bot* 55 (405):2029-2039.

Buddenbaum, H., and M. Steffens, 2012 The Effects of Spectral Pretreatments on Chemometric Analyses of Soil Profiles Using Laboratory Imaging Spectroscopy. *Applied and Environmental Soil Science* 2012:274903.

Bureau, S., D. Cozzolino, and C.J. Clark, 2019 Contributions of Fourier-transform mid infrared (FT-MIR) spectroscopy to the study of fruit and vegetables: A review. *Postharvest Biology and Technology* 148:1-14.

Bureau, S., D. Ruiz, M. Reich, B. Gouble, D. Bertrand *et al.*, 2009a Rapid and non-destructive analysis of apricot fruit quality using FT-near-infrared spectroscopy. *Food Chemistry* 113:1323-1328.

Bureau, S., D. Ruiz, M. Reich, B. Gouble, D. Bertrand *et al.*, 2009b Application of ATR-FTIR for a rapid and simultaneous determination of sugars and organic acids in apricot fruit. *Food Chemistry* 115 (3):1133-1140.

Byrne, D., 2007 Molecular marker use in perennial plant breeding. *Acta Horticulturae*:163-167.

Calus, M.P.L., 2010 Genomic breeding value prediction: methods and procedures. *Animal* 4 (2):157-164.

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp, 2008a Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* 124:362-368.

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp, 2008b Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. *Genetics* 178 (1):553.

Calus, M.P.L., and R.F. Veerkamp, 2011 Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution* 43:26.

Carraut, A., and P. Crossa-Raynaud, 1981 CARACTERISTIQUES DE LA CULTURE ET ORIENTATIONS DES RECHERCHES SUR L'ABRICOTIER EN TUNISIE, pp. 683-691. International Society for Horticultural Science (ISHS), Leuven, Belgium.

Chagné, D., C. Krieger, M. Rassam, M. Sullivan, J. Fraser *et al.*, 2012 QTL and candidate gene mapping for polyphenolic composition in apple fruit. *BMC Plant Biology* 12 (1):12.

Chambroy, Y., M. Souty, G. Jacquemin, R.-M. Gomez, and J.-M. Audergon, 1995 Research on the suitability of modified atmosphere packaging for shelf-life and quality improvement of apricot fruit. *Acta Hort.* 384:633-638.

Charlesworth, B., 2019 In defence of doing sums in genetics. *Heredity* 123 (1):44-49.

Chen, Z.-Q., J. Baison, J. Pan, B. Karlsson, B. Andersson *et al.*, 2018 Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. *BMC genomics* 19 (1).

Cheriton, D., and R.E. Tarjan, 1976 Finding Minimum Spanning Trees. *SIAM Journal on Computing* 5 (4):724-742.

Cobb, J.N., G. DeClerck, A. Greenberg, R. Clark, and S. McCouch, 2013 Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theoretical and Applied Genetics* 126 (4):867-887.

Cookson, W., L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, 2009 Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* 10 (3):184-194.

Covarrubias-Pazaran, G., 2016 Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PloS one* 11:1-15.

Covarrubias-Pazaran, G., 2018 Software update: Moving the R package sommer to multivariate mixed models for genome-assisted prediction. *biorxv*.

Covarrubias-Pazaran, G., B. Schlautman, L. Diaz-Garcia, E. Grygleski, J. Polashock *et al.*, 2018a Multivariate GBLUP Improves Accuracy of Genomic Selection for Yield and Fruit Weight in Biparental Populations of Vaccinium macrocarpon Ait. *Frontiers in Plant Science* 9 (1310).

Covarrubias-Pazaran, G., B. Schlautman, L. Diaz-Garcia, E. Grygleski, J. Polashock *et al.*, 2018b Multivariate GBLUP Improves Accuracy of Genomic Selection for Yield and Fruit Weight in Biparental Populations of Vaccinium macrocarpon Ait. *Frontiers in Plant Science* 9:1310.

Cozzolino, D., W.U. Cynkar, N. Shah, and P. Smith, 2011 Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality. *Food Research International* 44 (7):1888-1896.

Crossa, J., P. Perez-Rodriguez, J. Cuevas, O. Montesinos-Lopez, D. Jarquin *et al.*, 2017a Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci* 22 (11):961-975.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín *et al.*, 2017b Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science* 22 (11):961-975.

Crossa, J., P. Pérez, J. Hickey, J. Burgueño, L. Ornella *et al.*, 2014 Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112 (1):48-60.

Daetwyler, H.D., M.P. Calus, R. Pong-Wong, G. de Los Campos, and J.M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193 (2):347-365.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J. Woolliams, 2010a The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics* 185:1021-1031.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams, 2010b The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185 (3):1021-1031.

Daetwyler, H.D., B. Villanueva, and J. Woolliams, 2008 Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PloS one* 3:e3395.

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327-345.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics* 182 (1):375.

de Roos, A.P.W., B.J. Hayes, and M.E. Goddard, 2009 Reliability of Genomic Predictions Across Multiple Populations. *Genetics* 183:1545-1553.

Defilippi, B.G., A.M. Dandekar, and A.A. Kader, 2004 Impact of suppression of ethylene action or biosynthesis on flavor metabolites in apple (*Malus domestica* Borkh) fruits. *Journal of agricultural and food chemistry* 52:5694-5701.

Dekkers, J.C., 2004 Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J Anim Sci* 82 (328).

Desta, Z.A., and R. Ortiz, 2014 Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci* 19 (9):592-601.

Dirlewanger, E., E. Graziano, T. Joobeur, F. Garriga-Calderé, P. Cosson *et al.*, 2004 Comparative mapping and marker-assisted selection in Rosaceae fruit crops. *Proceedings of the National Academy of Sciences of the United States of America* 101 (26):9891.

Dirlewanger, E., J. Quero-García, L. Le Dantec, P. Lambert, D. Ruiz *et al.*, 2012 Comparison of the genetic determinism of two key phenological traits, flowering and maturity dates, in three Prunus species: peach, apricot and sweet cherry. *Heredity* 109 (5):280-292.

Donald, C.M., 1968 The breeding of crop ideotypes. *Euphytica* 17 (3):385-403.

Eduardo, I., I. Pacheco, G. Chietera, D. Bassi, C. Pozzi *et al.*, 2011 QTL analysis of fruit quality traits in two peach intraspecific populations and importance of maturity date pleiotropic effect. *Tree Genetics & Genomes* 7:323-335.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto *et al.*, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE* 6 (5):e19379.

Endelman, J.B., 2011a Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* 4 (3).

Endelman, J.B., 2011b Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4:250-255.

Endelman, J.B., 2011c Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* 4:250-255.

Etienne, A., M. Genard, P. Lobit, A.M.D. Mbeguie, and C. Bugaud, 2013 What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells. *J Exp Bot* 64 (6):1451-1469.

Fan, X., M.B. Sylvia, and P.M. James, 1999 1-Methylcyclopropene Inhibits Apple Ripening. *Journal of the American Society for Horticultural Science jashs* 124 (6):690-695.

Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu *et al.*, 2017 Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genetics Selection Evolution* 49:44.

FAO, F., 2018 FAOSTAT - Crops.

Faust, M., D. Surányi, and F. Nyujtó, 1998 Origin and dissemination of apricot. *Hort. Rev.* 22:225–266.

Fernandes, S.B., K.O.G. Dias, D.F. Ferreira, and P.J. Brown, 2018 Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theoretical and Applied Genetics* 131 (3):747-755.

Fisher, R.A., 1918 The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* 52 (2):399-433.

FISHER, R.A., 1941 Average excess and average effect of a gene substitution. *Annals of Eugenics* 11 (1):53-63.

Flutre, T., 2019 rutilstimflutre: Timothee Flutre's personal R.

Fodor, A., V. Segura, M. Denis, S. Neuenschwander, A. Fournier-Level *et al.*, 2014a Genome-Wide Prediction Methods in Highly Diverse and Heterozygous Species: Proof-of-Concept through Simulation in Grapevine. *PloS one* 9:e110436.

Fodor, A., V. Segura, M. Denis, S. Neuenschwander, A. Fournier-Level *et al.*, 2014b Genome-Wide Prediction Methods in Highly Diverse and Heterozygous Species: Proof-of-Concept through Simulation in Grapevine. *PLOS ONE* 9 (11):e110436.

FranceAgriMer, F., 2019 Les chiffres-clés de la filière Fruits & Légumes frais et transformés.

FranceAgriMer, F., 2020 Les chiffres-clés de la filière Fruits & Légumes frais et transformés en 2019.

Frett, T.J., G.L. Reighard, W.R. Okie, and K. Gasic, 2014 Mapping quantitative trait loci associated with blush in peach [Prunus persica (L.) Batsch]. *Tree Genetics & Genomes* 10:367-381.

Furbank, R.T., and M. Tester, 2011 Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science* 16 (12):635-644.

Gallais, A., 2011 *Méthodes de création de variétés en amélioration des plantes*. Versailles: Éd. Quae.

Gao, H.Y., B.Z. Zhu, H.L. Zhu, Y.L. Zhang, Y.H. Xie *et al.*, 2007 Effect of suppression of ethylene biosynthesis on flavor products in tomato fruits. *Russ. J. Plant Physiol.* 54:80-88.

García-Gómez, B.E., J.A. Salazar, L. Dondini, P. Martínez-Gomez, and D. Ruiz, 2019 Identification of QTLs linked to fruit quality traits in apricot (*Prunus armeniaca* L.) and biological validation through gene expression analysis using qPCR. *Molecular Breeding* 39.

Gatti, E., B.G. Defilippi, S. Predieri, and R. Infante, 2009 Apricot (*Prunus armeniaca* L.) quality and breeding perspectives. *Journal of Food, Agriculture and Environment* 7:573-580.

Gautam, R., S. Vanga, F. Ariese, and S. Umapathy, 2015 Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation* 2 (1):8.

Gautier, H., A. Rocci, M. Buret, D. Grasselly, and M. Causse, 2005 Fruit load or fruit position alters response to temperature and subsequent cherry tomato quality. *Journal of the Science of Food and Agriculture* 85:1009-1016.

Gebreselassie, M.N., K. Ader, N. Boizot, F. Millier, J.-P. Charpentier *et al.*, 2017 Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from Populus nigra (L.) natural populations. *Industrial Crops and Products* 107:159-171.

Gegas, V., A. Gay, A. Camargo, and J. Doonan, 2014 Challenges of Crop Phenomics in the Post-genomic Era, pp. 142-171.

Geladi, P., and E. Dåbakk, 1995 An Overview of Chemometrics Applications in near Infrared Spectrometry. *Journal of Near Infrared Spectroscopy* 3 (3):119-132.

Gianola, D., M. Pérez-Enciso, and M.A. Toro, 2003 On marker-assisted prediction of genetic value: Beyond the ridge. *Genetics* 163:347-365.

Gjuvsland, A.B., J.O. Vik, D.A. Beard, P.J. Hunter, and S.W. Omholt, 2013 Bridging the genotype-phenotype gap: what does it take? *The Journal of physiology* 591 (8):2055-2066.

Goddard, M., 2008 Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136:245-257.

Goddard, M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136 (2):245-257.

Goddard, M.E., and B.J. Hayes, 2007 Genomic selection. *J Anim Breed Genet* 124 (6):323-330.

Gonçalves, M.T.V., G. Morota, P.M.d.A. Costa, P.M.P. Vidigal, M.H.P. Barbosa *et al.*, 2021 Near-infrared spectroscopy outperforms genomics for predicting sugarcane feedstock quality traits. *PLOS ONE* 16 (3):e0236853.

Grattapaglia, D., F.L. Bertolucci, and R.R. Sederoff, 1995 Genetic mapping of QTLs controlling vegetative propagation in Eucalyptus grandis and E. urophylla using a pseudo-testcross strategy and RAPD markers. *Theoretical and Applied Genetics* 90 (7):933-947.

Grattapaglia, D., and M.D.V. Resende, 2011 Genomic selection in forest tree breeding. *Tree Genetics & Genomes* 7:241-255.

Grattapaglia, D., and R. Sederoff, 1994 Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* 137 (4):1121-1137.

Groppi, A., 2021 Following the Adaptive Path of Fruit Tree Domestication using Population Genomics: the Case of Apricots (submitted) in *Nature Communications*.

Guo, G., F. Zhao, Y. Wang, Y. Zhang, L. Du *et al.*, 2014 Comparison of single-trait and multiple-trait genomic prediction models. *BMC genetics* 15:30.

Guo, Z., M.M. Magwire, C.J. Basten, Z. Xu, and D. Wang, 2016 Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor Appl Genet* 129 (12):2413-2427.

Gurrieri, F., J.-M. Audergon, G. Albagnac, and M. Reich, 2001 Soluble sugars and carboxylic acids in ripe apricot fruit as parameters for distinguishing different cultivars. *Euphytica* 117:183-189.

Habier, D., R.L. Fernando, and J.C.M. Dekkers, 2007 The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* 177 (4):2389.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12 (186):1471-2105.

Hastie, T., R. Tibshirani, and J.H. Friedman, 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer.

Hayes, B., J. Panozzo, C. Walker, A. Choy, S. Kant *et al.*, 2017a Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theoretical and Applied Genetics* 130:1-15.

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard, 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution* 41 (1):51.

Hayes, B.J., and M.E. Goddard, 2008 Technical note: prediction of breeding values using marker-derived relationship matrices. *J Anim Sci* 86 (9):2089-2092.

Hayes, B.J., J. Panozzo, C.K. Walker, A.L. Choy, S. Kant *et al.*, 2017b Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theor Appl Genet* 130 (12):2505-2519.

Hayes, B.J., J. Pryce, A.J. Chamberlain, P.J. Bowman, and M.E. Goddard, 2010 Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. *PLOS Genetics* 6 (9):e1001139.

Heffner, E., M. Sorrells, and J.-L. Jannink, 2009 Genomic Selection for Crop Improvement. *Crop Science* 49.

Heffner, E.L., J.-L. Jannink, H. Iwata, E. Souza, and M.E. Sorrells, 2011 Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. *Crop Science* 51 (6):2597-2606.

Heffner, E.L., A.J. Lorenz, J.-L. Jannink, and M.E. Sorrells, 2010 Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. *Crop Science* 50 (5):1681-1690.

Henderson, C.R., 1975 Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31 (2):423-447.

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink, 2012 Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science* 56:146-160.

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu *et al.*, 2014 Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science* 54 (4):1476-1488.

Hoerl, A.E., and R.W. Kennard, 1970 Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12 (1):55-67.

Hospital, F., 2009 Challenges for effective marker-assisted selection in plants. *Genetica* 136 (2):303-310.

Houle, D., D.R. Govindaraju, and S. Omholt, 2010 Phenomics: the next challenge. *Nature Reviews Genetics* 11 (12):855-866.

Huang, H., J.U. Qureshi, S. Liu, Z. Sun, C. Zhang *et al.*, 2020 Hyperspectral Imaging as a Potential Online Detection Method of Microplastics. *Bulletin of Environmental Contamination and Toxicology*.

Huet, G., 1961 La sélection clonale de la variété rouge du Roussillon, pp. 7 p. in *Journées nationales de l'abricotier*, Perpignan, France.

Infante, R., P. Martínez-Gomez, and S. Predieri, 2008 Quality oriented fruit breeding: Peach [Prunus persica (L.) Batsch]. *Journal of Food, Agriculture and Environment* 6:342-356.

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2014 Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics* 128:145–158.

Jennings, H.S., 1917 The Numerical Results of Diverse Systems of Breeding, with Respect to Two Pairs of Characters, Linked or Independent, with Special Relation to the Effects of Linkage. *Genetics* 2 (2):97-154.

Jia, Y., and J.-L. Jannink, 2012 Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. *Genetics* 192:1513–1522.

Jiang, F., J. Zhang, S. Wang, L. Yang, Y. Luo *et al.*, 2019 The apricot (Prunus armeniaca L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis. *Horticulture Research* 6 (1):128.

Kainer, D., R. Lanfear, W.J. Foley, and C. Külheim, 2015 Genomic approaches to selection in outcrossing perennials: focus on essential oil crops. *Theor Appl Genet* 128 (12):2351-2365.

Kainer, D., E.A. Stone, A. Padovan, W.J. Foley, and C. Külheim, 2018 Accuracy of Genomic Prediction for Foliar Terpene Traits in Eucalyptus polybractea. *G3* 8 (8):2573-2583.

Karaman, E.M.S., M.T. Lund, M. Anche, L. Janss, and G. Su, 2018 Genomic prediction using multi-trait weighted GBLUP accounting for heterogeneous variances and covariances across the genome. *Genes Genomes Genetics* 8:3549-3558.

Knoch, D., C.R. Werner, R.C. Meyer, D. Riewe, A. Abbadi *et al.*, 2021 Multi-omics-based prediction of hybrid performance in canola. *Theoretical and Applied Genetics* 134 (4):1147-1165.

Kostina, K.F., 1969 The use of varietal resources of apricots for breeding. *Trud. nikit. bot. Sad.* 40:45-63.

Krause, M.R., L. González-Pérez, J. Crossa, P. Pérez-Rodríguez, O. Montesinos-López *et al.*, 2019 Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat. *G3: Genes|Genomes|Genetics* 9 (4):1231.

Kumar, S., M.C.A.M. Bink, R.K. Volz, V.G.M. Bus, and D. Chagné, 2012a Towards genomic selection in apple (Malus ×domestica Borkh.) breeding programmes: Prospects, challenges and strategies. *Tree Genetics & Genomes* 8 (1):1-14.

Kumar, S., D. Chagné, M.C.A.M. Bink, R.K. Volz, C. Whitworth *et al.*, 2012b Genomic selection for fruit quality traits in apple (*Malus* × *domestica* Borkh.). *PloS one* 7:e36674.

Kumar, S., C. Kirk, C.H. Deng, A. Shirtliff, C. Wiedow *et al.*, 2019 Marker-trait associations and genomic predictions of interspecific pear (Pyrus) fruit characteristics. *Scientific Reports* 9 (1):9072.

Kumar, S., D. Raju, R.N. Sahoo, and V. Chinnusamy, 2016 Phenomics: unlocking the hidden genetic variation for breaking the barriers in yield and stress tolerance. *Indian Journal of Plant Physiology* 21 (4):409-419.

Lado, B., D. Vázquez, M. Quincke, P. Silva, I. Aguilar *et al.*, 2018 Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality. *Theor Appl Genet* 131 (12):2719-2731.

Lamine, C., S. Simon, J. Audergon, S. Penvern, G. Clauzel *et al.*, 2017 Réalités et perspectives de l'écologisation en arboriculture fruitière - Pour une approche intégrée à partir du cas des vergers de pêchers et d'abricotiers en Rhône-Alpes.

Lande, R., and R. Thompson, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124 (3):743.

Lane, H., S. Murray, O. Montesinos-López, A. Montesinos‐López, J. Crossa *et al.*, 2020 Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. 3.

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen, 2008 Performance of Genomic Selection in Mice. *Genetics* 180:611-618.

Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan *et al.*, 2014 Usefulness of Multiparental Populations of Maize (Zea mays L.) for Genome-Based Prediction. *Genetics* 198:3-16.

Lenz, P.R.N., J. Beaulieu, S.D. Mansfield, S. Clément, M. Desponts *et al.*, 2017 Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). *BMC Genomics* 18:335.

Lewontin, R.C., 1964 THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS; HETEROTIC MODELS. *Genetics* 49 (1):49.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14):1754-1760.

Lichou, J., and M. Jay, 2012 *Abricot*.

Liu, H., H. Zhou, Y. Wu, X. Li, J. Zhao *et al.*, 2015a The Impact of Genetic Relationship and Linkage Disequilibrium on Genomic Selection. *PLOS ONE* 10 (7):e0132379.

Liu, H., H. Zhou, Y. Wu, X. Li, J. Zhao *et al.*, 2015b The Impact of Genetic Relationship and Linkage Disequilibrium on Genomic Selection. *PloS one* 10:e0132379.

Liu, S., A. Cornille, S. Decroocq, D. Tricon, A. Chague *et al.*, 2019a The complex evolutionary history of apricots: Species divergence, gene flow and multiple domestication events. *Molecular Ecology* 28 (24):5299-5314.

Liu, X., H. Wang, H. Xiaojiao, K. Li, Z. Liu *et al.*, 2019b Improving Genomic Selection With Quantitative Trait Loci and Nonadditive Effects Revealed by Empirical Evidence in Maize. *Frontiers in Plant Science* 10:1129.

Lopes, M.S., H. Bovenhuis, M. van Son, Ø. Nordbø, E. Grindflek *et al.*, 2017 Using markers with large effect in genetic and genomic predictions. *Journal of Animal Science* 95:59-71.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi *et al.*, 2011 Chapter Two - Genomic Selection in Plant Breeding: Knowledge and Prospects, pp. 77-123 in *Advances in Agronomy*, edited by D.L. Sparks. Academic Press.

Lorenz, A.J., and K.P. Smith, 2015 Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. *Crop Science* 55:2657.

Lorenzana, R.E., and R. Bernardo, 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics* 120:151-161.

Lozada, D.N., R.E. Mason, J.M. Sarinelli, and G. Brown-Guedira, 2019 Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC genetics* 20:82.

Mackay, T.F.C., E.A. Stone, and J.F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* 10 (8):565-577.

Maier, R., G. Moser, G.B. Chen, S. Ripke, W. Coryell *et al.*, 2015 Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet* 96 (2):283-294.

Makowsky, R., N.M. Pajewski, Y.C. Klimentidis, A.I. Vazquez, C.W. Duarte *et al.*, 2011 Beyond Missing Heritability: Prediction of Complex Traits. *PLoS genetics* 7:e1002051.

Mariette, S., F. Wong Jun Tai, G. Roch, A. Barre, A. Chague *et al.*, 2016 Genome-wide association links candidate genes to resistance to Plum Pox Virus in apricot (Prunus armeniaca). *New Phytol* 209 (2):773-784.

Marty, I., B. Sylvie, G. Sarkissian, B. Gouble, J.-M. Audergon *et al.*, 2005 Ethylene regulation of carotenoid accumulation and carotenogenic gene expression in colour-contrasted apricot varieties (*Prunus armeniaca*). *Journal of experimental botany* 56:1877-1886.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20 (9):1297-1303.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (4):1819-1829.

Michel, S., C. Kummer, M. Gallee, J. Hellinger, C. Ametz *et al.*, 2018 Improving the baking quality of bread wheat by genomic selection in early generations. *Theor Appl Genet* 131 (2):477-493.

Michel, S., C. Kummer, M. Gallee, J. Hellinger, C. Ametz *et al.*, 2017 Improving the baking quality of bread wheat by genomic selection in early generations. *Theoretical and Applied Genetics* 131:1-17.

Minamikawa, M., K. Nonaka, E. Kaminuma, H. Kajiya-Kanegae, A. Onogi *et al.*, 2017 Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits. *Scientific Reports* 7:1–13.

Misra, A., 2014 Climate change and challenges of water and food security. *International Journal of Sustainable Built Environment* 3.

Montes, J.M., A.E. Melchinger, and J.C. Reif, 2007 Novel throughput phenotyping platforms in plant genetic studies. *Trends in Plant Science* 12 (10):433-436.

Montesinos-López, O.A., A. Montesinos-López, J. Crossa, D. Gianola, C.M. Hernández-Suárez *et al.*, 2018 Multi-trait, Multi-environment Deep Learning Modeling for Genomic-Enabled Prediction of Plant Traits. *G3 Genes|Genomes|Genetics* 8 (12):3829-3840.

Morgante, F., W. Huang, C. Maltecca, and T.F.C. Mackay, 2018 Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. *Heredity* 120:500–514.

Muir, W., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics* 124:342-355.

Munkvold, J.D., J. Tanaka, D. Benscher, and M.E. Sorrells, 2009 Mapping quantitative trait loci for preharvest sprouting resistance in white wheat. *Theor Appl Genet* 119 (7):1223-1235.

Munoz-Sanz, J.V., E. Zuriaga, I. Lopez, M.L. Badenes, and C. Romero, 2017 Self-(in)compatibility in apricot germplasm is controlled by two major loci, S and M. *BMC Plant Biol* 17 (1):82.

Muranty, H., M. Troggio, b.s. Ines, M. Rifaï, A. Auwerkerken *et al.*, 2015 Accuracy and responses of genomic selection on key traits in apple breeding. *Horticulture Research* 2:15060.

Nicolaï, B.M., K. Beullens, E. Bobelyn, A. Peirs, W. Saeys *et al.*, 2007 Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology* 46 (2):99-118.

Nsibi, M., B. Gouble, S. Bureau, T. Flutre, C. Sauvage *et al.*, 2020 Adoption and Optimization of Genomic Selection To Sustain Breeding for Apricot Fruit Quality. *G3: Genes|Genomes|Genetics* 10 (12):4513.

Obenchain, V., M. Lawrence, V. Carey, S. Gogarten, P. Shannon *et al.*, 2014 VariantAnnotation : a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* 30 (14):2076-2078.

Omrani, M., 2018 Caractérisation des déterminants génétiques et moléculaires liés à la résistance au dépérissement bactérien chez l'abricotier et analyse des risques associés.

Omrani, M., M. Roth, G. Roch, A. Blanc, C.E. Morris *et al.*, 2019 Genome-wide association multi-locus and multi-variate linear mixed models reveal two linked loci with major effects on partial resistance of apricot to bacterial canker. *BMC Plant Biology* 19 (1):31.

Onogi, A., 2020 Connecting mathematical models to genomes: Joint estimation of model parameters and genome-wide marker effects on these parameters. *Bioinformatics*.

Ozaki, Y., S. Morita, and Y. Du, 2006 Spectral Analysis, pp. 47-72.

Park, T., and G. Casella, 2008 The Bayesian Lasso. *Journal of the American Statistical Association* 103 (482):681-686.

Paul, V., R. Pandey, and G.C. Srivastava, 2012 The fading distinctions between classical patterns of ripening in climacteric and non-climacteric fruit and the ubiquity of ethylene-An overview. *Journal of food science and technology* 49:1-21.

Pedryc, A., S. Ruthner, R. Hermán, B. Krska, A. Hegedűs *et al.*, 2009 Genetic diversity of apricot revealed by a set of SSR markers from linkage group G1. *Scientia Horticulturae* 121 (1):19-26.

Pérez, P., and G. de los Campos, 2014 Genome-Wide Regression & Prediction with the BGLR Statistical Package. *Genetics* 198:483–495.

Peterson, R.A., and J.E. Cavanaugh, 2020 Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics* 47 (13-15):2312-2327.

Piepho, H.P., J. Möhring, A.E. Melchinger, and A. Büchse, 2008 BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161 (1):209-228.

Pigliucci, M., 2005 Evolution of phenotypic plasticity: where are we going now? *Trends in Ecology & Evolution* 20 (9):481-486.

Pigliucci, M., 2010 Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365 (1540):557-566.

Pizarro, C., I. Esteban-Díez, A.-J. Nistal, and J.-M.a. González-Sáiz, 2004 Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Analytica Chimica Acta* 509 (2):217-227.

Porep, J., D. Kammerer, and R. Carle, 2015 On-line application of near infrared (NIR) spectroscopy in food production. *Trends in Food Science & Technology* 46:211–230.

Prunier, J.-P., J.-P. Jullian, R. Minodier, G. Clauzel, and J. Martins, 2005 Chancre bactérien sur abricotier: Eviter les dégâts par la mise en oeuvre de pratiques simples et raisonnées. *Arboriculture Fruitière* 8:23-30.

R Core Team, R., 2018 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rajsic, P., A. Weersink, A. Navabi, and K. Peter Pauls, 2016 Economics of genomic selection: the role of prediction accuracy and relative genotyping costs. *Euphytica* 210 (2):259-276.

Ramstein, G.P., S.E. Jensen, and E.S. Buckler, 2019 Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theoretical and Applied Genetics* 132:559–567.

Rana, M., A. Sood, W. Hussain, R. Kaldate, T. Sharma *et al.*, 2019 Gene Pyramiding and Multiple Character Breeding, pp. 83-124 in *Lentils: Potential Resources for Enhancing Genetic Gains*. edited by M. Singh. Academic Press, London.

Rasheed, A., Y. Hao, X. Xia, A. Khan, Y. Xu *et al.*, 2017 Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. *Molecular Plant* 10 (8):1047-1064.

Ratcliffe, B., O.G. El-Dien, E.P. Cappa, I. Porth, J. Klápště *et al.*, 2017 Single-Step BLUP with Varying Genotyping Effort in Open-Pollinated <em>Picea glauca</em&gt. *G3: Genes|Genomes|Genetics* 7 (3):935.

Ren, D., L. An, B. Li, L. Qiao, and W. Liu, 2020 Efficient weighting methods for genomic best linear-unbiased prediction (BLUP) adapted to the genetic architectures of quantitative traits. *Heredity*.

Resende, M.D.V., M.F.R. Resende Jr, C.P. Sansaloni, C.D. Petroli, A.A. Missiaggia *et al.*, 2012a Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist* 194 (1):116-128.

Resende, M.F.R., Jr., P. Muñoz, J.J. Acosta, G.F. Peter, J.M. Davis *et al.*, 2012b Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol* 193 (3):617-624.

Resende, M.F.R., P. Muñoz, M.D.V. Resende, D.J. Garrick, R.L. Fernando *et al.*, 2012c Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine. *Genetics* 190 (4):1503.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow *et al.*, 2012 Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics* 44 (2):217-220.

Riedelsheimer, C., J.B. Endelman, M. Stange, M.E. Sorrells, J.-L. Jannink *et al.*, 2013 Genomic Predictability of Interconnected Biparental Maize Populations. *Genetics* 194:493-503.

Rincent, R., A. Charcosset, and L. Moreau, 2018a Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics* 130 (11):2231-2247.

Rincent, R., J.-P. Charpentier, P. Faivre-Rampant, E. Paux, J. Le Gouis *et al.*, 2018b Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar. *G3: Genes|Genomes|Genetics* 8 (12):3961.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel *et al.*, 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.

Rinnan, Å., F.v.d. Berg, and S.B. Engelsen, 2009 Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* 28 (10):1201-1222.

Ritchie, M.D., E.R. Holzinger, R. Li, S.A. Pendergrass, and D. Kim, 2015 Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* 16 (2):85-97.

Romero, C., A. Pedryc, V. Muñoz, G. Llácer, and M.L. Badenes, 2003 Genetic diversity of different apricot geographical groups determined by SSR markers. *Genome* 46 (2):244-252.

Roth, M., H. Muranty, M. Di Guardo, W. Guerra, A. Patocchi *et al.*, 2020 Genomic prediction of fruit texture and training population optimization towards the application of genomic selection in apple. *Horticulture Research* 7 (1):148.

Roussos, P.A., V. Sefferou, N.-K. Denaxa, E. Tsantili, and V. Stathis, 2011 Apricot (*Prunus armeniaca* L.) fruit quality attributes and phytochemicals under different crop load. *Scientia Horticulturae* 129:472-478.

Ruiz, D., P. Lambert, J.-M. Audergon, L. Dondini, S. Tartarini *et al.*, 2010 Identification of QTLs for fruit quality traits in apricot. *Acta Horticulturae* 862:587-592.

Ruiz, D., M. Reich, S. Bureau, C.M. Renard, and J.M. Audergon, 2008 Application of reflectance colorimeter measurements and infrared spectroscopy methods to rapid and nondestructive evaluation of carotenoids content in apricot (Prunus armeniaca L.). *J Agric Food Chem* 56 (13):4916-4922.

Salazar, J., D. Ruiz, J. Campoy, S. Tartarini, L. Dondini *et al.*, 2016 Inheritance of reproductive phenology traits and related QTL identification in apricot. *Tree Genetics & Genomes* 12.

Salazar, J.A., I. Pacheco, P. Shinya, P. Zapata, C. Silva *et al.*, 2017 Genotyping by Sequencing for SNP-Based Linkage Analysis and Identification of QTLs Linked to Fruit Quality Traits in Japanese Plum (*Prunus salicina* Lindl.). *Frontiers in Plant Science* 8:476.

Savitzky, A., and M.J.E. Golay, 1964 Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36 (8):1627-1639.

Scandella, D., and X. Vernin, 2019 *Évolution du marché de l'abricot: Perception et attentes de la filière et des consommateurs*: CTIFL.

Schefers, J.M., and K.A. Weigel, 2012 Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. *Animal Frontiers* 2 (1):4-9.

Schopp, P., D. Müller, Y.C.J. Wientjes, and A.E. Melchinger, 2017 Genomic Prediction Within and Across Biparental Families: Means and Variances of Prediction Accuracy and Usefulness of Deterministic Equations. *G3: Genes|Genomes|Genetics* 7 (11):3571.

Segura, V., B.J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren *et al.*, 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44 (7):825-830.

Shirasawa, K., K. Isuzugawa, M. Ikenaga, Y. Saito, T. Yamamoto *et al.*, 2017 The genome sequence of sweet cherry (Prunus avium) for use in genomics-assisted breeding. *DNA Research* 24 (5):499-508.

Socquet-Juglard, D., D. Christen, G. Devènes, C. Gessler, B. Duffy *et al.*, 2013 Mapping Architectural, Phenological, and Fruit Quality QTLs in Apricot. *Plant Molecular Biology Reporter* 31:387-397.

Solberg, T.R., A.K. Sonesson, J.A. Woolliams, and T.H. Meuwissen, 2009 Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution* 41:29.

Spindel, J.E., H. Begum, D. Akdemir, B. Collard, E. Redoña *et al.*, 2016 Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116:395–408.

Stevens, A., and L. Ramirez-Lopez, 2020 An introduction to the prospectr package.

Tan, B., D. Grattapaglia, G.S. Martins, K.Z. Ferreira, B. Sundberg *et al.*, 2017 Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biology* 17:110.

Taylor, J., and D. Butler, 2017 R Package ASMap: Efficient Genetic Linkage Map Construction and Diagnosis. *Journal of Statistical Software* 79:1-29.

Te Pas, M.F.W., O. Madsen, M.P.L. Calus, and M.A. Smits, 2017 The Importance of Endophenotypes to Evaluate the Relationship between Genotype and External Phenotype. *International journal of molecular sciences* 18 (2):472.

Testolin, R., 2011 Kiwifruit breeding: From the phenotypic analysis of parents to the genomic estimation of their breeding value (GEBV). *Acta Horticulturae* 913:123-130.

Thompson, R., and K. Meyer, 1986 A review of theoretical aspects in the estimation of breeding values for multi-trait selection. *Livestock Production Science* 15 (4):299-313.

Tibshirani, R., 1996 Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1):267-288.

Valdés, H., M. Pizarro, R. Campos-Vargas, R. Infante, and B.G. Defilippi, 2009 Effect of Ethylene Inhibitors on Quality Attributes of Apricot cv. Modesto and Patterson during Storage. *Chilean Journal of Agricultural Research* 69:134-144.

Van Ghelder, C., 2019 The Ma resistance locus to root-knot nematodes in Prunus : Structural originality and evolution within the NBS-LRR gene family in plants

VanRaden, P.M., 2008 Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91 (11):4414-4423.

Verde, I., N. Bassil, S. Scalabrin, B. Gilmore, C.T. Lawley *et al.*, 2012 Development and Evaluation of a 9K SNP Array for Peach by Internationally Coordinated SNP Detection and Validation in Breeding Germplasm. *PLOS ONE* 7 (4):e35668.

Verde, I., A.G. Abbott, S. Scalabrin, S. Jung, S. Shu *et al.*, 2013 The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics* 45 (5):487-494.

Verde, I., J. Jenkins, L. Dondini, S. Micali, G. Pagliarani *et al.*, 2017 The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC genomics* 18 (1):225-225.

Vilanova, S., M.L. Badenes, L. Burgos, J. Martínez-Calvo, G. Llácer *et al.*, 2006 Self-Compatibility of Two Apricot Selections Is Associated with Two Pollen-Part Mutations of Different Nature. *Plant Physiology* 142 (2):629.

Visscher, P.M., S.E. Medland, M.A.R. Ferreira, K.I. Morley, G. Zhu *et al.*, 2006 Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLoS genetics* 2:e41.

Voorrips, R.E., 2002 MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *The Journal of heredity* 93:77-78.

Walsh, K.B., J. Blasco, M. Zude-Sasse, and X. Sun, 2020 Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use. *Postharvest Biology and Technology* 168:111246.

Wang, S.Y., and M.J. Camp, 2000 Temperatures after bloom affect plant growth and fruit quality of strawberry. *Scientia Horticulturae* 85:183-199.

Wang, X., and G. Zhou, 2011 Study on Pretreatment Algorithm of Near Infrared Spectroscopy, pp. 623-632 in *Computer and Computing Technologies in Agriculture IV*, edited by D. Li, Y. Liu and Y. Chen. Springer Berlin Heidelberg, Berlin, Heidelberg.

White, J., P. Andrade-Sanchez, M. Gore, K. Bronson, T. Coffelt *et al.*, 2012 Field-based phenomics for plant genetics research. *Field Crops Research* 133:101–112.

Whittaker, J.C., R. Thompson, and M.C. Denham, 2000 Marker-assisted selection using ridge regression. *Genetical research* 75:249-252.

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus, 2013 The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics* 193:621-631.

Wu, R., H. Wei, Y. Cui, H. Li, T. Liu *et al.*, 2007 Modeling the Genetic Architecture of Complex Traits With Molecular Markers. *Recent patents on nanotechnology* 1:41-49.

Xu, Y., X. Liu, J. Fu, H. Wang, J. Wang *et al.*, 2020 Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Communications* 1 (1):100005.

Zhang, Q., W. Chen, L. Sun, F. Zhao, B. Huang *et al.*, 2012 The genome of Prunus mume. *Nature Communications* 3 (1):1318.

Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal, 2016 Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. *Frontiers in Genetics* 7 (151).

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao *et al.*, 2014 Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. *PLoS One* 9 (3):e93017.

Zhong, S., J.C.M. Dekkers, R.L. Fernando, and J.-L. Jannink, 2009 Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. *Genetics* 182:355-364.

Zhu, J., P. Sova, Q. Xu, K.M. Dombek, E.Y. Xu *et al.*, 2012 Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. *PLOS Biology* 10 (4):e1001301.

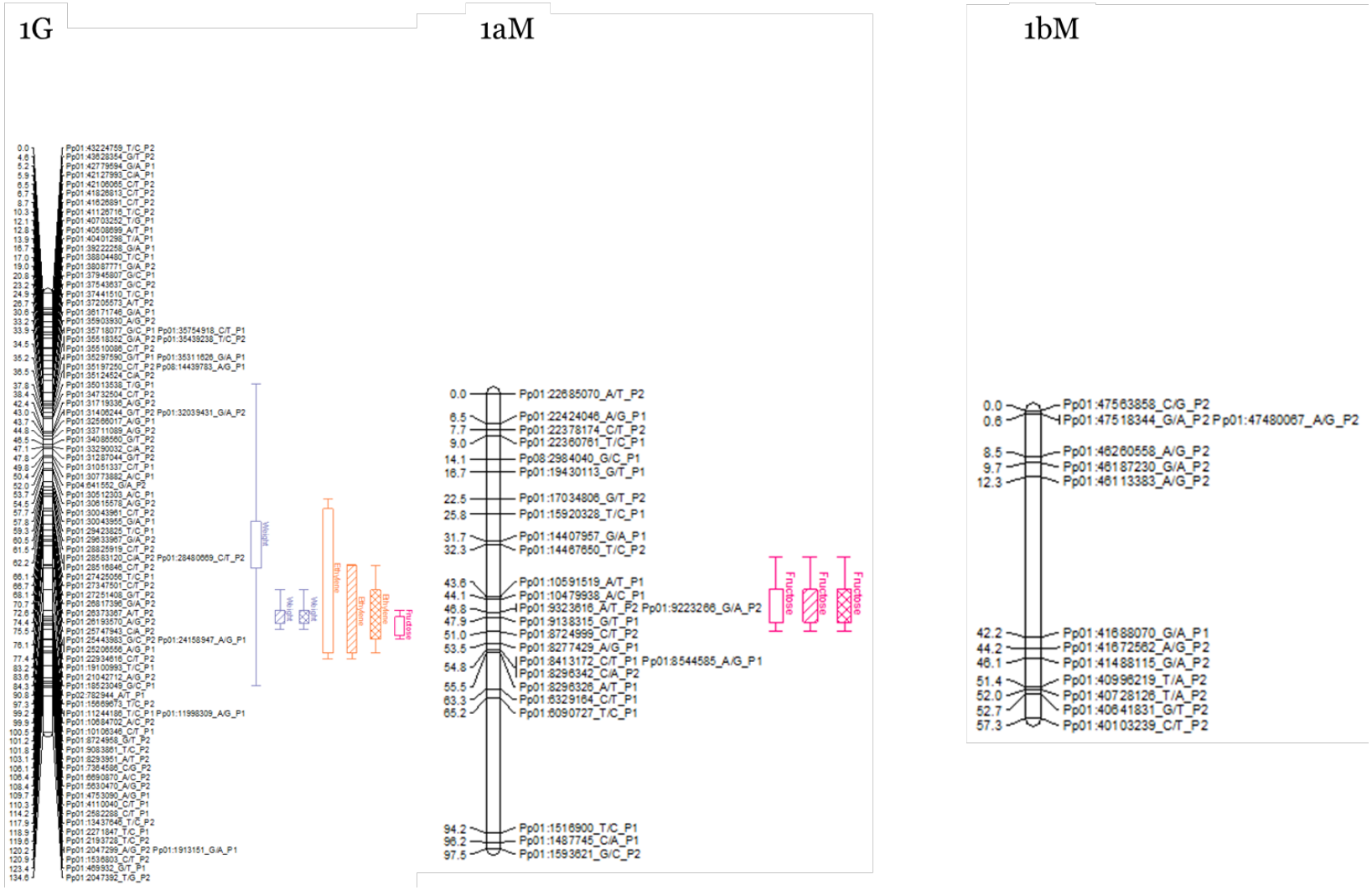
Table des annexes

Annexe I : Matériel supplémentaire du chapitre 3

Descriptive statistics of phenotypic data

TRAITS	UNIT	ABBREVIATIONS	MIN	MEDIAN	MAX	MEAN	SD	%CV
FRUIT WEIGHT	g	F.Weight	30.60	58.90	97.80	59.61	12.71	21.32
HUE GROUND	degrees	Hue.g	61.51	75.64	98.55	76.25	6.49	8.51
ETHYLENE PRODUCTION	nmol kg-1h-1	Ethylene	1.95	6.43	9.60	6.14	1.83	29.74
REFRACTIVE INDEX	% Brix	RI	9.60	13.80	20.60	14.00	1.85	13.22
SUCROSE CONTENT	g 100 g-1 of fresh weight	Sucrose	3.29	6.39	11.83	6.46	1.27	19.71
GLUCOSE CONTENT	g 100 g-1 of fresh weight	Glucose	0.39	1.60	3.10	1.62	0.34	20.67
FRUCTOSE CONTENT	g 100 g-1 of fresh weight	Fructose	0.15	0.78	3.44	0.80	0.24	30.41
TITRATABLE ACIDITY	meq 100 g-1 of fresh weight	TA	11.27	24.13	41.05	24.54	5.15	21.00
CITRIC ACID CONTENT	meq 100 g-1 of fresh weight	Citric.A.	8.01	20.71	45.26	21.10	5.07	24.04
MALIC ACID CONTENT	meq 100 g-1 of fresh weight	Malic.A.	4.09	7.81	17.92	8.09	2.15	26.61

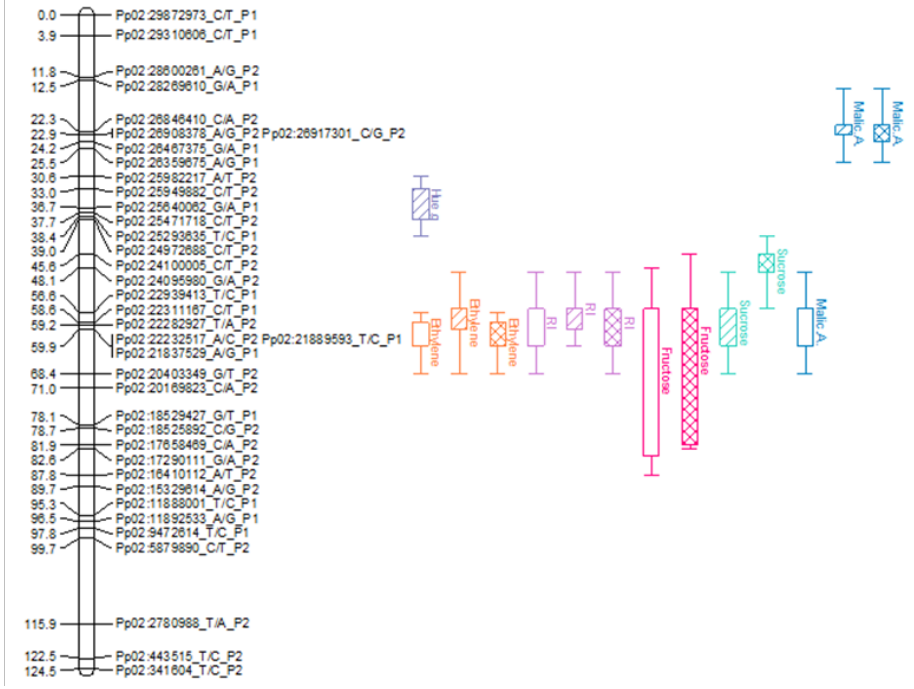
Parental genetic maps constructed for the Go X Mo progeny



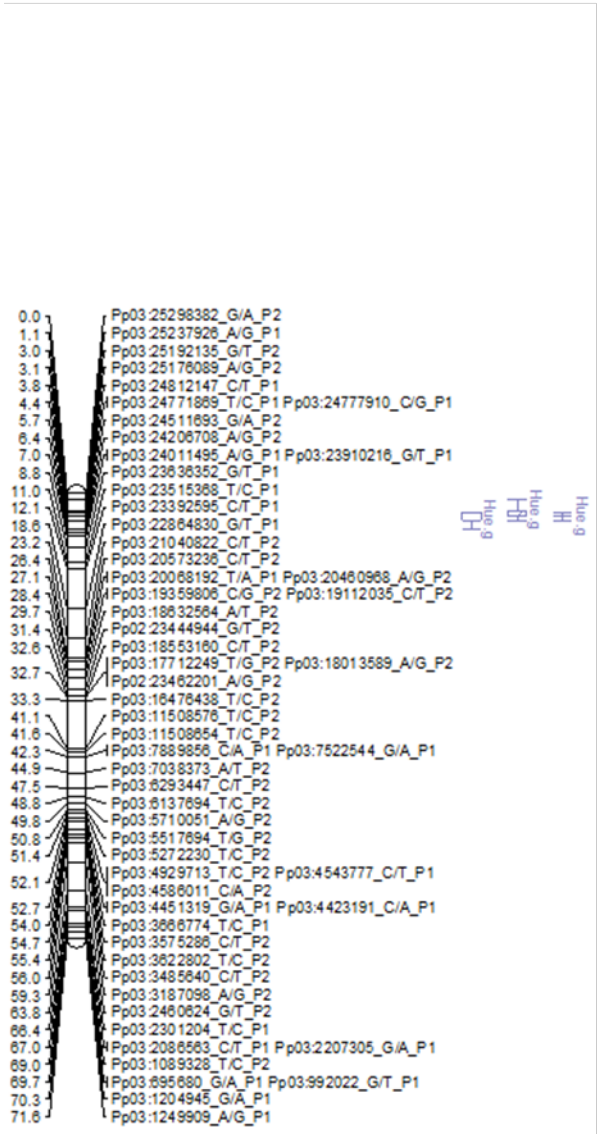
2G


0.0 Pp02:29941497_A/T_P2
0.6 Pp02:30117837_A/T_P2
2.0 Pp02:29416978_A/G_P2
4.9 Pp02:28898871_C/G_P1
5.2 Pp02:29021649_C/T_P2
6.5 Pp02:28893418_T/C_P2
7.2 Pp02:28685810_G/A_P2
7.8 Pp02:28456165_G/A_P2 Pp02:28478480_A/G_P1
9.1 Pp02:28333335_T/C_P1
10.4 Pp02:28209742_G/A_P1 Pp02:28169924_T/C_P1
12.4 Pp02:27660529_C/A_P1
13.5 Pp02:27781020_A/G_P2
15.1 Pp02:27132358_T/C_P2
20.8 Pp02:27097409_T/G_P1
25.3 Pp02:27018113_G/T_P1
26.0 Pp02:26846496_A/T_P1
27.3 Pp02:26342242_T/A_P2
29.8 Pp02:26008875_G/A_P2
32.7 Pp02:19786089_G/A_P1
34.3 Pp02:21721333_C/T_P2
Pp02:23164976_T/C_P2 Pp02:20285970_C/G_P2
35.0 Pp02:17818416_C/A_P2 Pp02:21144208_C/T_P1
Pp02:19781073_G/A_P2
39.6 Pp02:15328394_T/G_P2
41.5 Pp02:14401261_T/C_P2
45.2 Pp02:12107428_G/A_P2
45.6 Pp02:12005153_G/T_P2 Pp02:12005068_G/A_P2
47.5 Pp02:6759166_C/T_P2
48.2 Pp02:5038218_G/A_P1 Pp02:4451025_G/A_P2
48.8 Pp02:4451056_C/A_P2
49.9 Pp04:3435209_G/A_P2
51.3 Pp02:4173108_C/G_P2
52.4 Pp05:14777810_C/A_P1
54.9 Pp02:3308517_T/A_P1
56.8 Pp02:2758055_G/A_P2
57.4 Pp02:2451974_T/A_P2
65.3 Pp02:1052851_A/C_P1
66.6 Pp02:961721_A/G_P2
67.3 Pp02:928905_C/T_P1 Pp02:929028_T/G_P1
69.1 Pp02:518164_G/T_P2
70.9 Pp02:204539_A/C_P2
Ethylene
Ethylene
Ethylene
TA
TA
TA
Citric.A
Citric.A
Citric.A
Malic.A
Malic.A
Malic.A

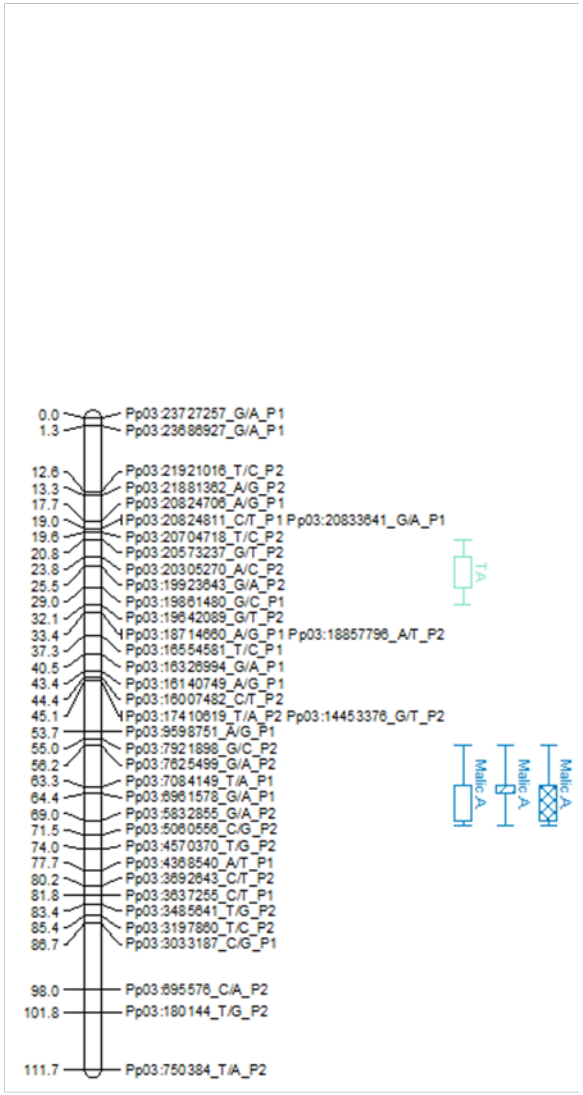
2M

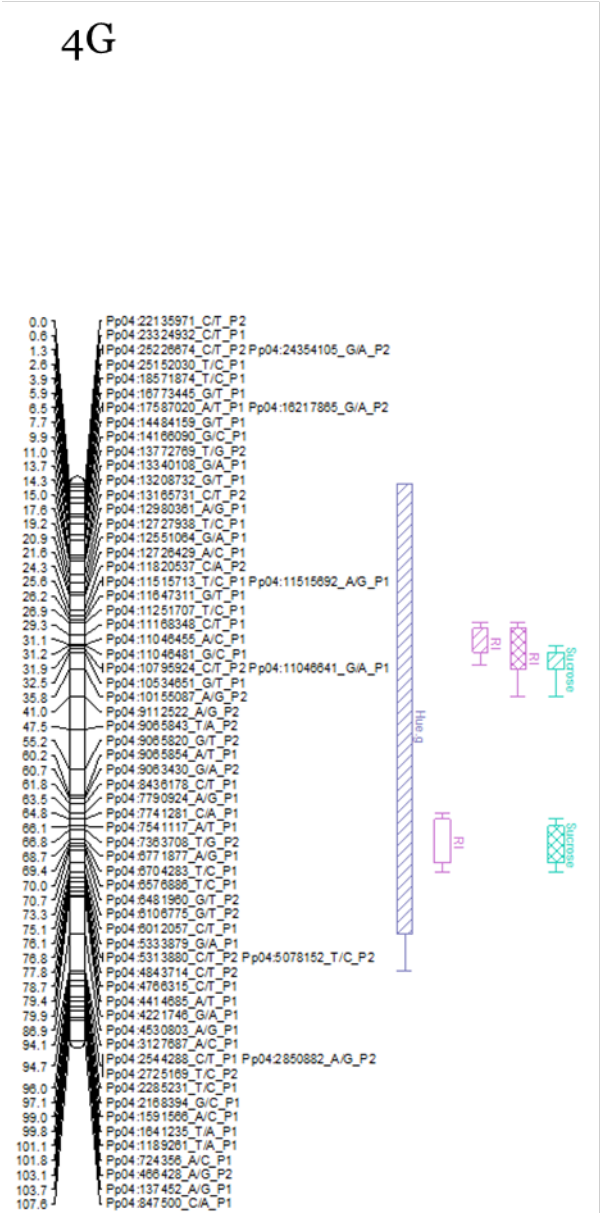

0.0 Pp02:29872973_C/T_P1
3.9 Pp02:29310606_C/T_P1
11.8 Pp02:28600261_A/G_P2
12.5 Pp02:28269610_G/A_P1
22.3 Pp02:26846410_C/A_P2
22.9 Pp02:26908378_A/G_P2 Pp02:26917301_C/G_P2
24.2 Pp02:26467375_G/A_P1
25.5 Pp02:26359675_A/G_P1
30.6 Pp02:25982217_A/T_P2
33.0 Pp02:25949882_C/T_P2
36.7 Pp02:25640062_G/A_P1
37.7 Pp02:25471718_C/T_P2
38.4 Pp02:25293635_T/C_P1
39.0 Pp02:24972688_C/T_P2
45.6 Pp02:24100005_C/T_P2
48.1 Pp02:24095980_G/A_P2
56.6 Pp02:22939413_T/C_P1
58.6 Pp02:22311167_C/T_P1
59.2 Pp02:22282927_T/A_P2
Pp02:22232517_A/C_P2 Pp02:21889593_T/C_P1
59.9 Pp02:21837529_A/G_P1
68.4 Pp02:20403349_G/T_P2
71.0 Pp02:20169823_C/A_P2
78.1 Pp02:18529427_G/T_P1
78.7 Pp02:18525892_C/G_P2
81.9 Pp02:17658469_C/A_P2
82.6 Pp02:17290111_G/A_P2
87.8 Pp02:16410112_A/T_P2
89.7 Pp02:15329614_A/G_P2
95.3 Pp02:11888001_T/C_P1
96.5 Pp02:11892533_A/G_P1
97.8 Pp02:9472614_T/C_P1
99.7 Pp02:5879890_C/T_P2
115.9 Pp02:2780988_T/A_P2
122.5 Pp02:443515_T/C_P2
124.5 Pp02:341604_T/C_P2
Hue.g
Ethylene
Ethylene
RI
RI
RI
Fructose
Fructose
Sucrose
Sucrose
Malic.A
Malic.A
Malic.A

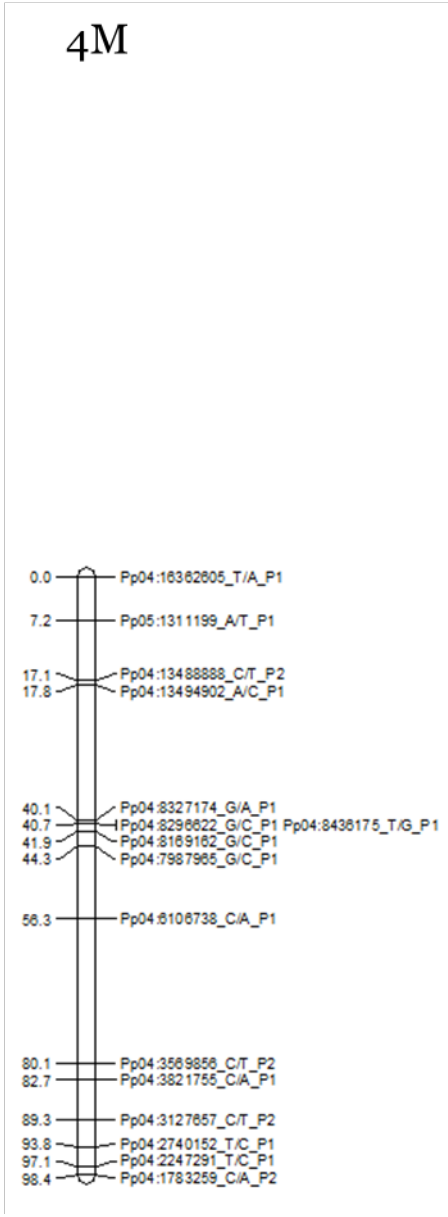
3G


0.0 Pp03:25298382_G/A_P2
1.1 Pp03:25237926_A/G_P1
3.0 Pp03:25192135_G/T_P2
3.1 Pp03:25176089_A/G_P2
3.8 Pp03:24812147_C/T_P1
4.4 Pp03:24771889_T/C_P1 Pp03:24777910_C/G_P1
5.7 Pp03:24511693_G/A_P2
6.4 Pp03:24206708_A/G_P2
7.0 Pp03:24011495_A/G_P1 Pp03:23910216_G/T_P1
8.8 Pp03:23636352_G/T_P1
11.0 Pp03:23515368_T/C_P1
12.1 Pp03:23392595_C/T_P1
18.6 Pp03:22864830_G/T_P1
23.2 Pp03:21040822_C/T_P2
26.4 Pp03:20573236_C/T_P2
27.1 Pp03:20068192_T/A_P1 Pp03:20460968_A/G_P2
28.4 Pp03:19359806_C/G_P2 Pp03:19112035_C/T_P2
29.7 Pp03:18632564_A/T_P2
31.4 Pp02:23444944_G/T_P2
32.6 Pp03:18553160_C/T_P2
32.7 Pp03:17712249_T/G_P2 Pp03:18013589_A/G_P2
Pp02:23462201_A/G_P2
33.3 Pp03:16476438_T/C_P2
41.1 Pp03:11508576_T/C_P2
41.6 Pp03:11508654_T/C_P2
42.3 Pp03:7889856_C/A_P1 Pp03:7522544_G/A_P1
44.9 Pp03:7038373_A/T_P2
47.5 Pp03:6293447_C/T_P2
48.8 Pp03:6137694_T/C_P2
49.8 Pp03:5710051_A/G_P2
50.8 Pp03:5517694_T/G_P2
51.4 Pp03:5272230_T/C_P2
52.1 Pp03:4929713_T/C_P2 Pp03:4543777_C/T_P1
Pp03:4586011_C/A_P2
52.7 Pp03:4451319_G/A_P1 Pp03:4423191_C/A_P1
54.0 Pp03:3666774_T/C_P1
54.7 Pp03:3575286_C/T_P2
55.4 Pp03:3622802_T/C_P2
56.0 Pp03:3485640_C/T_P2
59.3 Pp03:3187098_A/G_P2
63.8 Pp03:2460624_G/T_P2
66.4 Pp03:2301204_T/C_P1
67.0 Pp03:2086563_C/T_P1 Pp03:2207305_G/A_P1
69.0 Pp03:1089328_T/C_P2
69.7 Pp03:695680_G/A_P1 Pp03:992022_G/T_P1
70.3 Pp03:1204945_G/A_P1
71.6 Pp03:1249909_A/G_P1
Hue.g
Hue.g
Hue.g

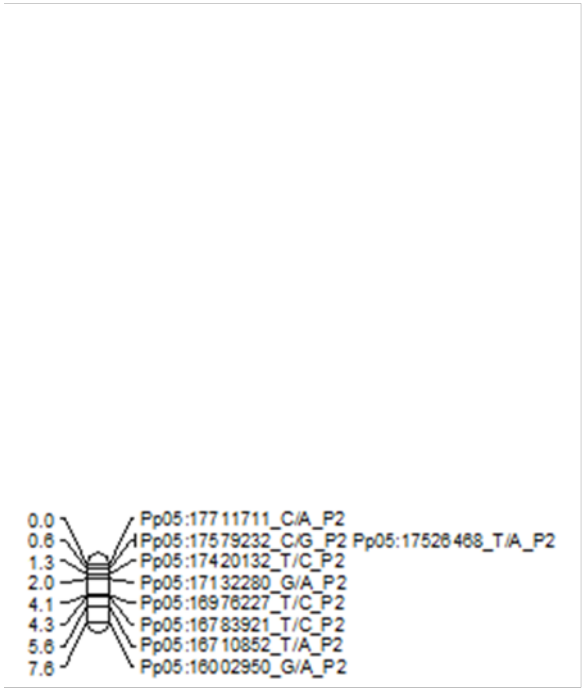
3M


0.0 Pp03:23727257_G/A_P1
1.3 Pp03:23686927_G/A_P1
12.6 Pp03:21921016_T/C_P2
13.3 Pp03:21881362_A/G_P2
17.7 Pp03:20824706_A/G_P1
19.0 Pp03:20824811_C/T_P1 Pp03:20833641_G/A_P1
19.6 Pp03:20704718_T/C_P2
20.8 Pp03:20573237_G/T_P2
23.8 Pp03:20305270_A/C_P2
25.5 Pp03:19923643_G/A_P2
29.0 Pp03:19861480_G/C_P1
32.1 Pp03:19642089_G/T_P2
33.4 Pp03:18714660_A/G_P1 Pp03:18857796_A/T_P2
37.3 Pp03:16554581_T/C_P1
40.5 Pp03:16326994_G/A_P1
43.4 Pp03:16140749_A/G_P1
44.4 Pp03:16007482_C/T_P2
45.1 Pp03:17410619_T/A_P2 Pp03:14453376_G/T_P2
53.7 Pp03:9598751_A/G_P1
55.0 Pp03:7921898_G/C_P2
56.2 Pp03:7625499_G/A_P2
63.3 Pp03:7084149_T/A_P1
64.4 Pp03:6961578_G/A_P1
69.0 Pp03:5832855_G/A_P2
71.5 Pp03:5060556_C/G_P2
74.0 Pp03:4570370_T/G_P2
77.7 Pp03:4366540_A/T_P1
80.2 Pp03:3692643_C/T_P2
81.8 Pp03:3637255_C/T_P1
83.4 Pp03:3485641_T/G_P2
85.4 Pp03:3197860_T/C_P2
86.7 Pp03:3033187_C/G_P1
98.0 Pp03:695576_C/A_P2
101.8 Pp03:180144_T/G_P2
111.7 Pp03:750384_T/A_P2
TA
Malic A
Malic A
Malic A

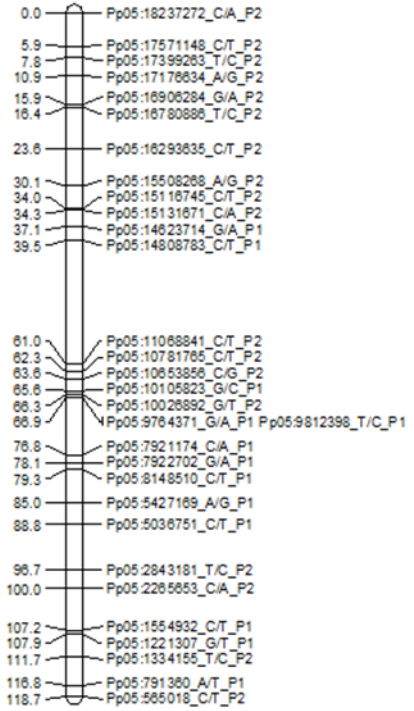

4G
0.0 Pp04:22135971_C/T_P2
0.6 Pp04:23324932_C/T_P1
1.3 Pp04:25226874_C/T_P2 Pp04:24354105_G/A_P2
2.6 Pp04:25152030_T/C_P1
3.9 Pp04:18571874_T/C_P1
5.9 Pp04:16773445_G/T_P1
6.5 Pp04:17587020_A/T_P1 Pp04:16217865_G/A_P2
7.7 Pp04:14484159_G/T_P1
9.9 Pp04:14166090_G/C_P1
11.0 Pp04:13772769_T/G_P2
13.7 Pp04:13340108_G/A_P1
14.3 Pp04:13208732_G/T_P1
15.0 Pp04:13165731_C/T_P2
17.6 Pp04:12980361_A/G_P1
19.2 Pp04:12727938_T/C_P1
20.9 Pp04:12551064_G/A_P1
21.6 Pp04:12726429_A/C_P1
24.3 Pp04:11820537_C/A_P2
25.6 Pp04:11515713_T/C_P1 Pp04:11515692_A/G_P1
26.2 Pp04:11647311_G/T_P1
26.9 Pp04:11251707_T/C_P1
29.3 Pp04:11168348_C/T_P1
31.1 Pp04:11046455_A/C_P1
31.2 Pp04:11046481_G/C_P1
31.9 Pp04:10795924_C/T_P2 Pp04:11046641_G/A_P1
32.5 Pp04:10534651_G/T_P1
35.8 Pp04:10155087_A/G_P2
41.0 Pp04:9112522_A/G_P2
47.5 Pp04:9065843_T/A_P2
55.2 Pp04:9065820_G/T_P2
60.2 Pp04:9065854_A/T_P1
60.7 Pp04:9063430_G/A_P2
61.8 Pp04:8436178_C/T_P1
63.5 Pp04:7790924_A/G_P1
64.8 Pp04:7741281_C/A_P1
66.1 Pp04:7541117_A/T_P1
66.8 Pp04:7363708_T/G_P2
68.7 Pp04:6771877_A/G_P1
69.4 Pp04:6704283_T/C_P1
70.0 Pp04:6576886_T/C_P1
70.7 Pp04:6481960_G/T_P2
73.3 Pp04:6106775_G/T_P2
75.1 Pp04:6012057_C/T_P1
76.1 Pp04:5333879_G/A_P1
76.8 Pp04:5313880_C/T_P2 Pp04:5078152_T/C_P2
77.8 Pp04:4843714_C/T_P2
78.7 Pp04:4766315_C/T_P1
79.4 Pp04:4414685_A/T_P1
79.9 Pp04:4221746_G/A_P1
86.9 Pp04:4530803_A/G_P1
94.1 Pp04:3127687_A/C_P1
94.7 Pp04:2544288_C/T_P1 Pp04:2850882_A/G_P2
96.0 Pp04:2725169_T/C_P2
97.1 Pp04:2285231_T/C_P1
99.0 Pp04:2168394_G/C_P1
99.8 Pp04:1591566_A/C_P1
101.1 Pp04:1841235_T/A_P1
101.8 Pp04:1189261_T/A_P1
103.1 Pp04:724356_A/C_P1
103.7 Pp04:486428_A/G_P2
107.6 Pp04:137452_A/G_P1
Pp04:847500_C/A_P1
Hue_g
RI
RI
RI
Sucrose
Sucrose


4M
0.0 Pp04:16362605_T/A_P1
7.2 Pp05:1311199_A/T_P1
17.1 Pp04:13488888_C/T_P2
17.8 Pp04:13494902_A/C_P1
40.1 Pp04:8327174_G/A_P1
40.7 Pp04:8296622_G/C_P1 Pp04:8436175_T/G_P1
41.9 Pp04:8169162_G/C_P1
44.3 Pp04:7987965_G/C_P1
56.3 Pp04:6106738_C/A_P1
80.1 Pp04:3569856_C/T_P2
82.7 Pp04:3821755_C/A_P1
89.3 Pp04:3127657_C/T_P2
93.8 Pp04:2740152_T/C_P1
97.1 Pp04:2247291_T/C_P1
98.4 Pp04:1783259_C/A_P2

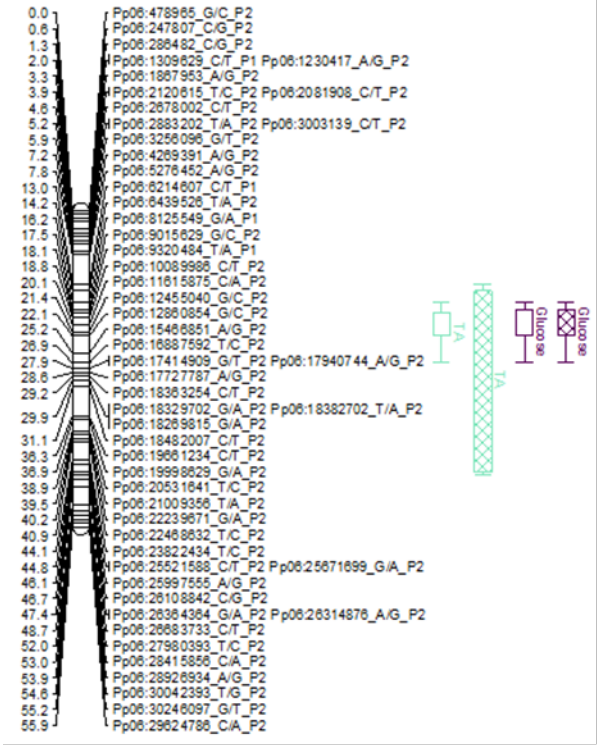
5G


0.0 Pp05:17711711_C/A_P2
0.6 Pp05:17579232_C/G_P2 Pp05:17526468_T/A_P2
1.3 Pp05:17420132_T/C_P2
2.0 Pp05:17132280_G/A_P2
4.1 Pp05:16976227_T/C_P2
4.3 Pp05:16783921_T/C_P2
5.6 Pp05:16710852_T/A_P2
7.6 Pp05:16002950_G/A_P2

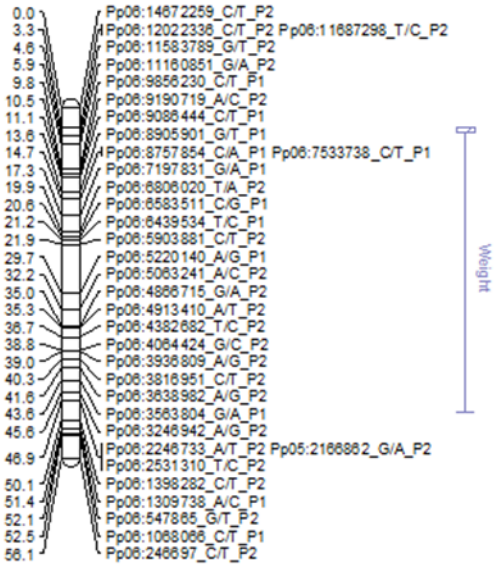
5M

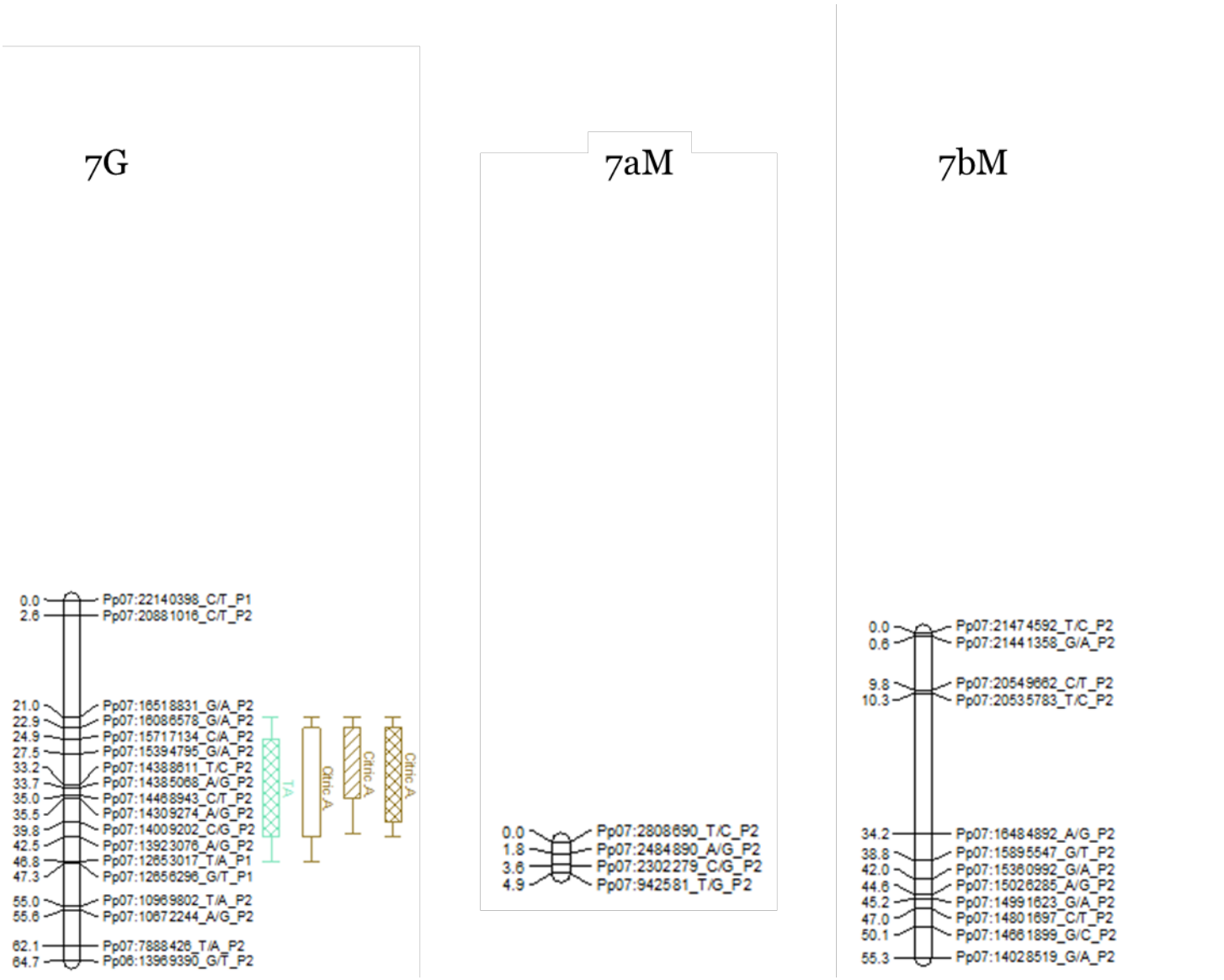

0.0 Pp05:18237272_C/A_P2
5.9 Pp05:17571148_C/T_P2
7.8 Pp05:17399263_T/C_P2
10.9 Pp05:17176634_A/G_P2
15.9 Pp05:16906284_G/A_P2
16.4 Pp05:16780886_T/C_P2
23.6 Pp05:16293635_C/T_P2
30.1 Pp05:15508268_A/G_P2
34.0 Pp05:15116745_C/T_P2
34.3 Pp05:15131671_C/A_P2
37.1 Pp05:14623714_G/A_P1
39.5 Pp05:14808783_C/T_P1
61.0 Pp05:11068841_C/T_P2
62.3 Pp05:10781765_C/T_P2
63.6 Pp05:10653856_C/G_P2
65.6 Pp05:10105823_G/C_P1
66.3 Pp05:10026892_G/T_P2
66.9 Pp05:9764371_G/A_P1 Pp05:9812398_T/C_P1
76.8 Pp05:7921174_C/A_P1
78.1 Pp05:7922702_G/A_P1
79.3 Pp05:8148510_C/T_P1
85.0 Pp05:5427169_A/G_P1
88.8 Pp05:5036751_C/T_P1
96.7 Pp05:2843181_T/C_P2
100.0 Pp05:2265653_C/A_P2
107.2 Pp05:1554932_C/T_P1
107.9 Pp05:1221307_G/T_P1
111.7 Pp05:1334155_T/C_P2
116.8 Pp05:791360_A/T_P1
118.7 Pp05:565018_C/T_P2

6G

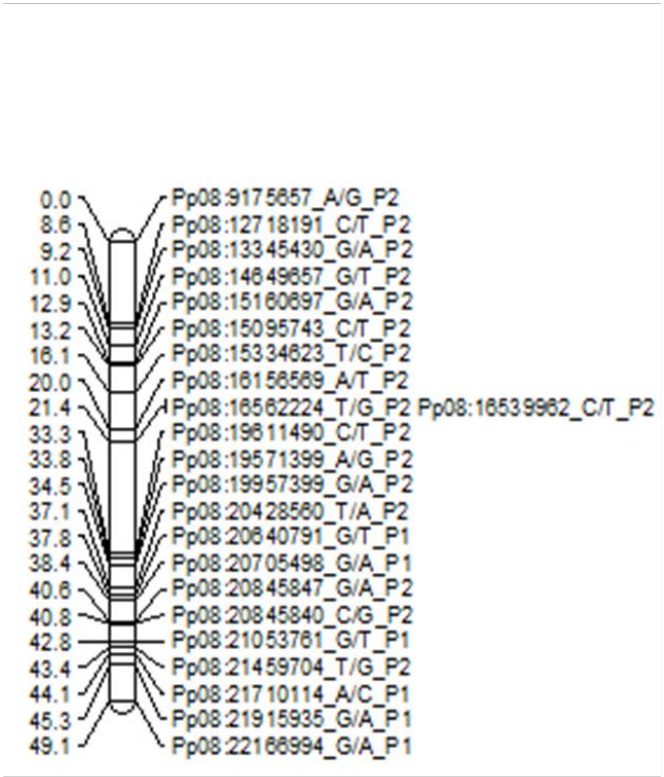

0.0 Pp06:478965_G/C_P2
0.6 Pp06:247807_C/G_P2
1.3 Pp06:286482_C/G_P2
2.0 Pp06:1309629_C/T_P1 Pp06:1230417_A/G_P2
3.3 Pp06:1887953_A/G_P2
3.9 Pp06:2120615_T/C_P2 Pp06:2081908_C/T_P2
4.6 Pp06:2678002_C/T_P2
5.2 Pp06:2883202_T/A_P2 Pp06:3003139_C/T_P2
5.9 Pp06:3256096_G/T_P2
7.2 Pp06:4269391_A/G_P2
7.8 Pp06:5276452_A/G_P2
13.0 Pp06:6214607_C/T_P1
14.2 Pp06:6439526_T/A_P2
16.2 Pp06:8125549_G/A_P1
17.5 Pp06:9015629_G/C_P2
18.1 Pp06:9320484_T/A_P1
18.8 Pp06:10089986_C/T_P2
20.1 Pp06:11615875_C/A_P2
21.4 Pp06:12455040_G/C_P2
22.1 Pp06:12860854_G/C_P2
25.2 Pp06:15466851_A/G_P2
26.9 Pp06:16887592_T/C_P2
27.9 Pp06:17414909_G/T_P2 Pp06:17940744_A/G_P2
28.6 Pp06:17727787_A/G_P2
29.2 Pp06:18363254_C/T_P2
29.9 Pp06:18329702_G/A_P2 Pp06:18382702_T/A_P2
31.1 Pp06:18269815_G/A_P2
36.3 Pp06:18482007_C/T_P2
36.9 Pp06:19661234_C/T_P2
38.9 Pp06:19998629_G/A_P2
39.5 Pp06:20531641_T/C_P2
40.2 Pp06:21009356_T/A_P2
40.9 Pp06:22239671_G/A_P2
44.1 Pp06:22468632_T/C_P2
44.8 Pp06:23822434_T/C_P2
46.1 Pp06:25521588_C/T_P2 Pp06:25671699_G/A_P2
46.7 Pp06:25997555_A/G_P2
47.4 Pp06:26108842_C/G_P2
48.7 Pp06:26364364_G/A_P2 Pp06:26314876_A/G_P2
52.0 Pp06:26683733_C/T_P2
53.0 Pp06:27980393_T/C_P2
53.9 Pp06:28415856_C/A_P2
54.6 Pp06:28926934_A/G_P2
55.2 Pp06:30042393_T/G_P2
55.9 Pp06:30246097_G/T_P2
Pp06:29624786_C/A_P2
TA
TA
Glucose
Glucose

6M

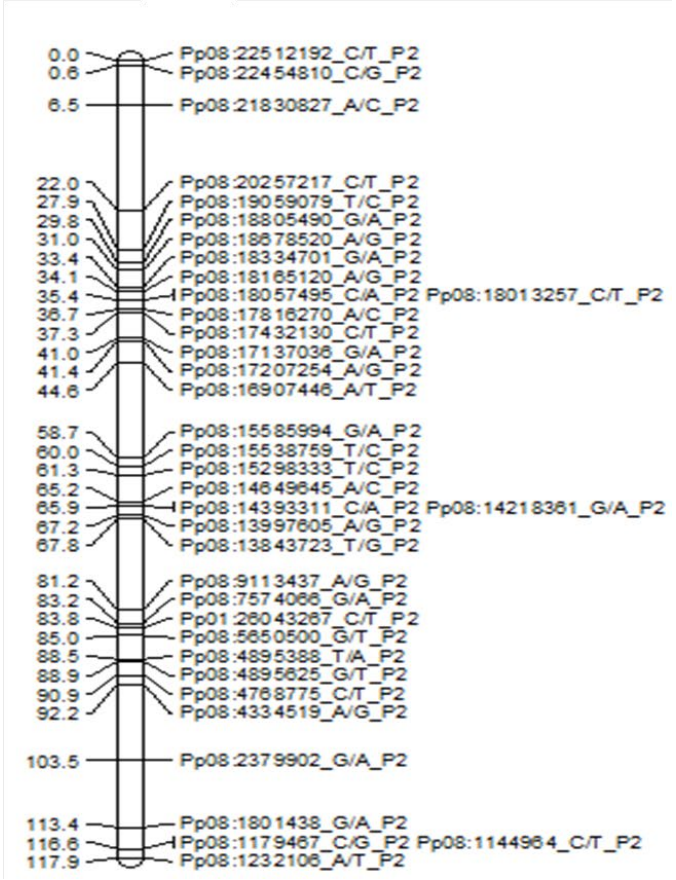

0.0 Pp06:14672259_C/T_P2
3.3 Pp06:12022336_C/T_P2 Pp06:11687298_T/C_P2
4.6 Pp06:11583789_G/T_P2
5.9 Pp06:11160851_G/A_P2
9.8 Pp06:9856230_C/T_P1
10.5 Pp06:9190719_A/C_P2
11.1 Pp06:9086444_C/T_P1
13.6 Pp06:8905901_G/T_P1
14.7 Pp06:8757854_C/A_P1 Pp06:7533738_C/T_P1
17.3 Pp06:7197831_G/A_P1
19.9 Pp06:6806020_T/A_P2
20.6 Pp06:6583511_C/G_P1
21.2 Pp06:6439534_T/C_P1
21.9 Pp06:5903881_C/T_P2
29.7 Pp06:5220140_A/G_P1
32.2 Pp06:5063241_A/C_P2
35.0 Pp06:4866715_G/A_P2
35.3 Pp06:4913410_A/T_P2
36.7 Pp06:4382682_T/C_P2
38.8 Pp06:4064424_G/C_P2
39.0 Pp06:3936809_A/G_P2
40.3 Pp06:3816951_C/T_P2
41.6 Pp06:3638982_A/G_P2
43.6 Pp06:3563804_G/A_P1
45.6 Pp06:3246942_A/G_P2
46.9 Pp06:2246733_A/T_P2 Pp05:2166862_G/A_P2
Pp06:2531310_T/C_P2
50.1 Pp06:1398282_C/T_P2
51.4 Pp06:1309738_A/C_P1
52.1 Pp06:547865_G/T_P2
52.5 Pp06:1068066_C/T_P1
56.1 Pp06:246697_C/T_P2
Weight


7G
0.0 Pp07:22140398_C/T_P1
2.6 Pp07:20881016_C/T_P2
21.0 Pp07:16518831_G/A_P2
22.9 Pp07:16086578_G/A_P2
24.9 Pp07:15717134_C/A_P2
27.5 Pp07:15394795_G/A_P2
33.2 Pp07:14388611_T/C_P2
33.7 Pp07:14385068_A/G_P2
35.0 Pp07:14468943_C/T_P2
35.5 Pp07:14309274_A/G_P2
39.8 Pp07:14009202_C/G_P2
42.5 Pp07:13923076_A/G_P2
46.8 Pp07:12653017_T/A_P1
47.3 Pp07:12656296_G/T_P1
55.0 Pp07:10969802_T/A_P2
55.6 Pp07:10672244_A/G_P2
62.1 Pp07:7888426_T/A_P2
64.7 Pp06:13969390_G/T_P2
T/A
Citric A
Citric A
Citric A
7aM
0.0 Pp07:2808690_T/C_P2
1.8 Pp07:2484890_A/G_P2
3.6 Pp07:2302279_C/G_P2
4.9 Pp07:942581_T/G_P2
7bM
0.0 Pp07:21474592_T/C_P2
0.6 Pp07:21441358_G/A_P2
9.8 Pp07:20549662_C/T_P2
10.3 Pp07:20535783_T/C_P2
34.2 Pp07:16484892_A/G_P2
38.8 Pp07:15895547_G/T_P2
42.0 Pp07:15360992_G/A_P2
44.6 Pp07:15026285_A/G_P2
45.2 Pp07:14991623_G/A_P2
47.0 Pp07:14801697_C/T_P2
50.1 Pp07:14661899_G/C_P2
55.3 Pp07:14028519_G/A_P2

8G

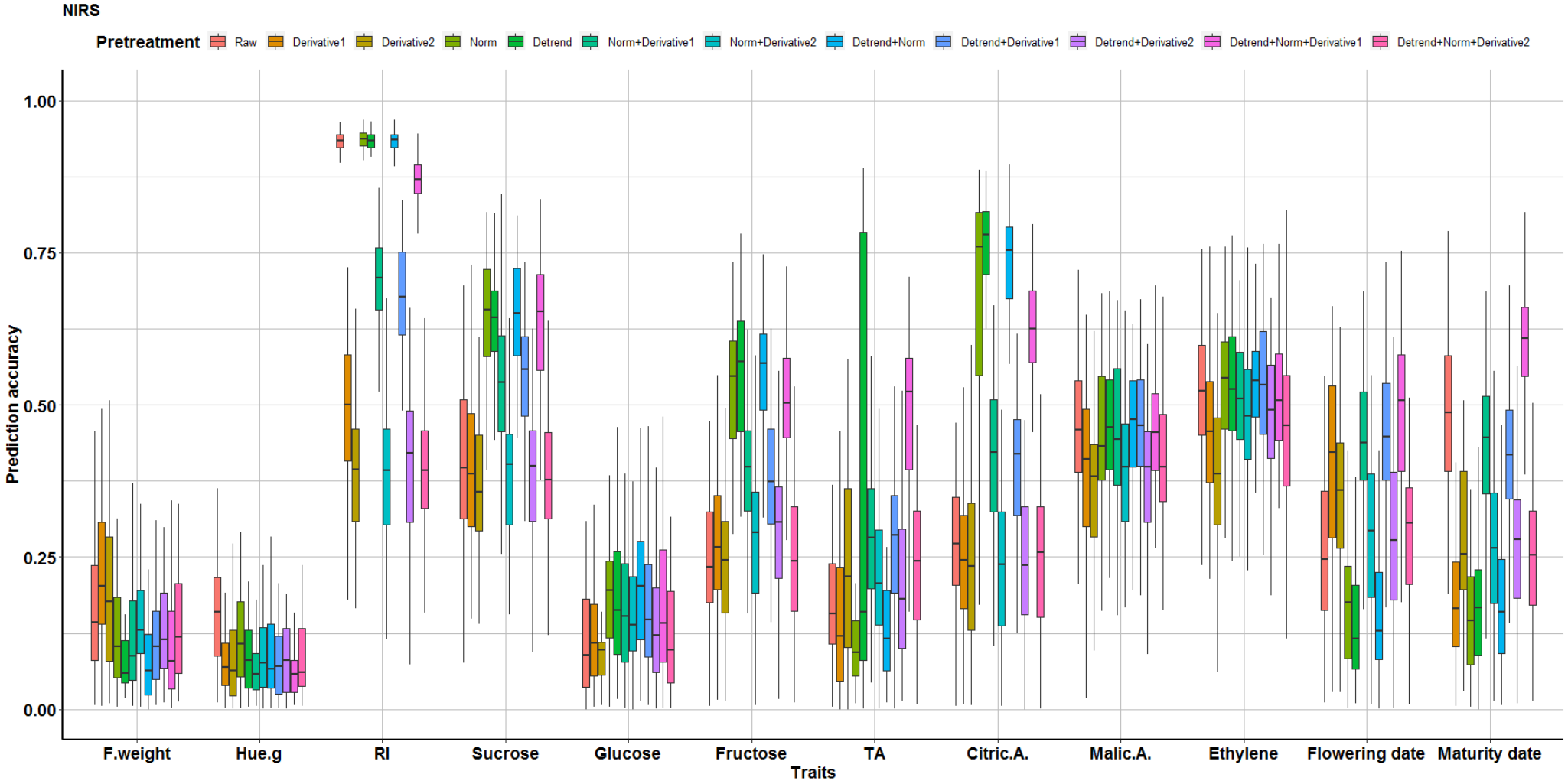

0.0 Pp08:9175657_A/G_P2
8.6 Pp08:12718191_C/T_P2
9.2 Pp08:13345430_G/A_P2
11.0 Pp08:14649657_G/T_P2
12.9 Pp08:15160697_G/A_P2
13.2 Pp08:15095743_C/T_P2
16.1 Pp08:15334623_T/C_P2
20.0 Pp08:16156569_A/T_P2
21.4 Pp08:16562224_T/G_P2 Pp08:16539962_C/T_P2
33.3 Pp08:19611490_C/T_P2
33.8 Pp08:19571399_A/G_P2
34.5 Pp08:19957399_G/A_P2
37.1 Pp08:20428560_T/A_P2
37.8 Pp08:20640791_G/T_P1
38.4 Pp08:20705498_G/A_P1
40.6 Pp08:20845847_G/A_P2
40.8 Pp08:20845840_C/G_P2
42.8 Pp08:21053761_G/T_P1
43.4 Pp08:21459704_T/G_P2
44.1 Pp08:21710114_A/C_P1
45.3 Pp08:21915935_G/A_P1
49.1 Pp08:22166994_G/A_P1

8M

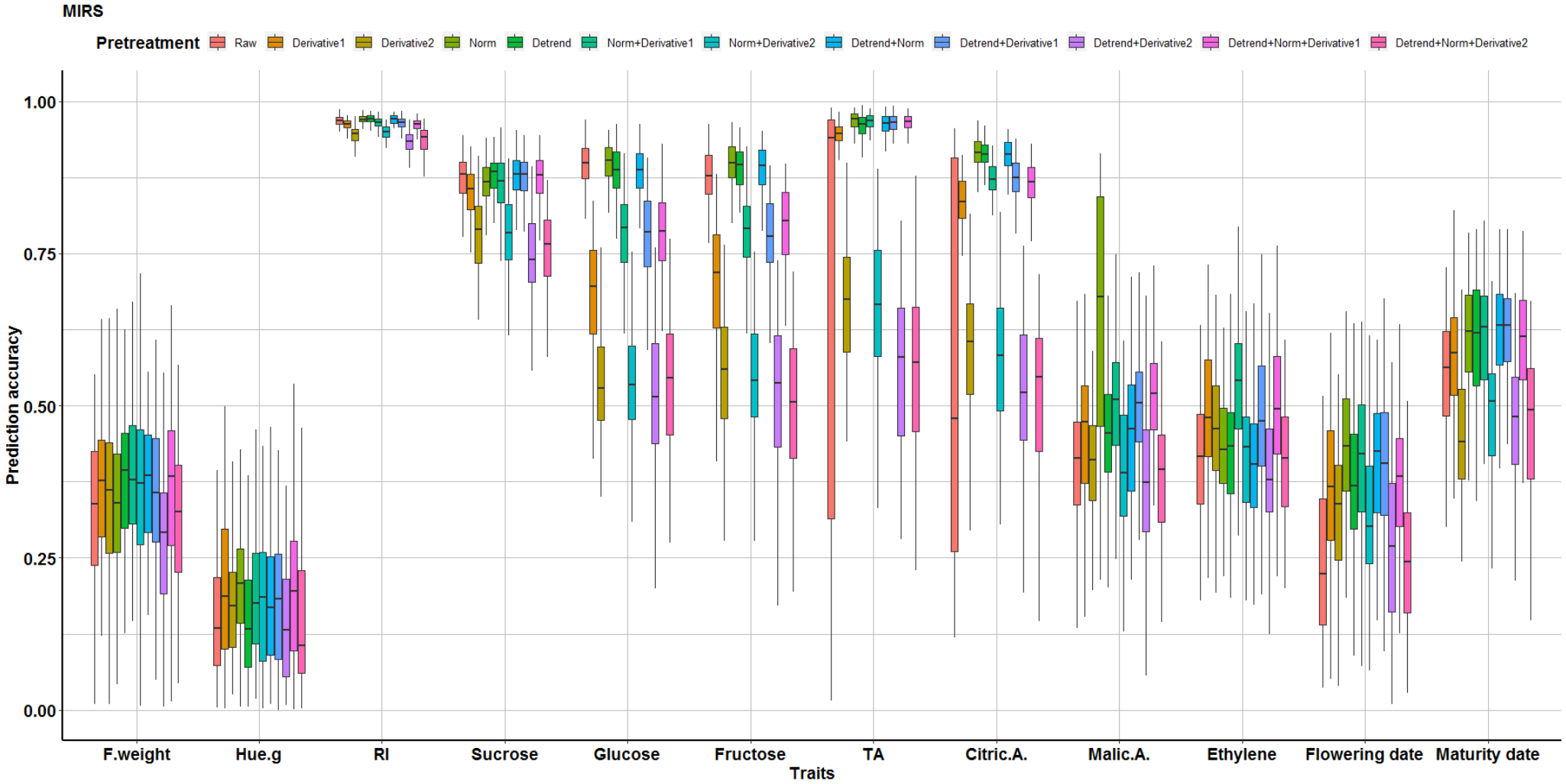

0.0 Pp08:22512192_C/T_P2
0.6 Pp08:22454810_C/G_P2
6.5 Pp08:21830827_A/C_P2
22.0 Pp08:20257217_C/T_P2
27.9 Pp08:19059079_T/C_P2
29.8 Pp08:18805490_G/A_P2
31.0 Pp08:18678520_A/G_P2
33.4 Pp08:18334701_G/A_P2
34.1 Pp08:18165120_A/G_P2
35.4 Pp08:18057495_C/A_P2 Pp08:18013257_C/T_P2
36.7 Pp08:17816270_A/C_P2
37.3 Pp08:17432130_C/T_P2
41.0 Pp08:17137036_G/A_P2
41.4 Pp08:17207254_A/G_P2
44.6 Pp08:16907446_A/T_P2
58.7 Pp08:15585994_G/A_P2
60.0 Pp08:15538759_T/C_P2
61.3 Pp08:15298333_T/C_P2
65.2 Pp08:14649645_A/C_P2
65.9 Pp08:14393311_C/A_P2 Pp08:14218361_G/A_P2
67.2 Pp08:13997605_A/G_P2
67.8 Pp08:13843723_T/G_P2
81.2 Pp08:9113437_A/G_P2
83.2 Pp08:7574066_G/A_P2
83.8 Pp01:26043267_C/T_P2
85.0 Pp08:5650500_G/T_P2
88.5 Pp08:4895388_T/A_P2
88.9 Pp08:4895625_G/T_P2
90.9 Pp08:4768775_C/T_P2
92.2 Pp08:4334519_A/G_P2
103.5 Pp08:2379902_G/A_P2
113.4 Pp08:1801438_G/A_P2
116.6 Pp08:1179467_C/G_P2 Pp08:1144964_C/T_P2
117.9 Pp08:1232106_A/T_P2

Annexe II : Matériel supplémentaire du chapitre 4

Assessement of prediction performance of NIRS-BLUP models for 12 agronomic traits linked to apricot fruit quality and phenology according to the spectral preprocessing



Assessement of prediction performance of MIRS-BLUP models for 12 agronomic traits linked to apricot fruit quality and phenology according to the spectral preprocessing



Annexe III: Matériel supplémentaire du chapitre 5

Plan du dispositif expérimental relatif à la core-collection implantée à Gotheron

	Clones	A	B	C	D	F	G	H	I	J	K	L	M
Bloc 1	1	6239	2348	4589	2346	1731	2243	2669	2361	2458	2067	5347	4415
	2	2069	2821	4079	5665	4294	380	2343	76	3976	5304	4516	3902
	3	2218	2360	4316	4892	3967	1811	4576	300	3950	2219	4804	4076
	4	157	4373	4034	3597	5129	2662	1314	74	2388	4806	4374	1727
	5	5328	5446	1343	1453	1915	1814	4049	2311	2265	4852	4800	1601
	6	4705	5006	2217	4896	2924	1714	2310	4445	3521	4710	4586	862
	7	1711	5271	5810	1793	2914	2382	682	2335	4394	4811	4360	1693
	8	3862	3698	4002	4423	1939	2205	2215	2928	2137	2410	1145	4650
	9	2129	1352	4322	2312	892	6240	2843	544	2635	4456	2089	1601
	10	3845	4617	4728	2481	1236	1813	2712	1729	2340	4711	2635	3075
	11	3408	500	6039	2241	660	804	8	2715	1330	2490	2994	3865
Bloc 2	12	4811	1727	4589	2310	2243	665	2894	2358	2156	634	4793	4167
	13	5304	3545	2069	2205	1793	2360	3075	2218	3865	2669	2914	1809
	14	3902	3844	4853	2241	2712	2928	3698	4166	4167	74	2156	3936
	15	5271	4710	4705	4516	2924	1693	3862	4449	4034	2358	4079	3967
	16	2662	4589	4617	2894	6239	2312	4456	2204	2388	1711	2217	8
	17	4423	1939	2067	3976	4076	544	4576	4373	4316	2410	2458	539
	18	4711	5810	1714	4804	1813	4445	4650	1601	157	2715	2481	6240
	19	4656	5446	3408	4892	1731	5347	5328	4793	2219	2635	1145	2346
	20	2343	1915	5006	4852	5665	3845	804	4394	2340	660	2129	2137
	21	3521	4728	4360	4806	2348	4843	2382	4374	544	3597	2311	4002
	22	500	3950	1811	6039	4800	1814	1729	4049	2843	2335	1453	39
	23	4322	1343	300	1236	4896	682	1352	892	4294	2821	2089	76
Bloc 3	24	4896	4360	2914	1314	3950	2265	2490	862	2734	4415	4586	804
	25	4076	4650	3865	4516	3976	3862	5129	2215	2137	3408	1330	2205
	26	3075	1727	3902	2067	2712	2635	1601	2335	2343	4322	2924	2458
	27	2928	3845	3936	2715	4456	1793	4374	2310	4415	5129	2481	4423
	28	2241	665	3967	4079	4294	1811	2156	4373	4316	4167	1731	4586
	29	157	300	2219	682	2311	2217	1809	2361	4049	2360	2129	1352
	30	1729	634	1814	3545	4800	76	2204	4394	2089	6239	5006	500
	31	2340	544	2490	4656	1453	660	4711	4710	2358	5304	1714	1145
	32	2410	862	4576	892	6240	2312	2821	2382	4449	2894	1939	1236
	33	539	2218	4166	4804	4445	4034	4793	2734	5328	4806	1711	4843
	34	1314	74	3844	4728	2069	380	4811	2388	3521	2348	2669	4892
	35	8	2243	2994	380	3698	2346	5665	2662	4853	1915	5271	6039
	36	4589	1813	5347	2843	3597	1343	39	4705	5810	5446	1729	4617
Bloc 4	37	4079	4650	1601	3936	3902	2712	4394	4456	4167	3698	2358	2361
	38	2215	4793	1314	3967	2137	2924	3597	4896	2067	3408	300	1236
	39	6039	5665	4445	4360	1814	1314	3862	4711	2348	1813	544	3845
	40	1793	4853	5810	2734	1731	2458	4374	4076	4811	2346	4516	2312
	41	5304	4576	1693	2311	2241	2156	1915	4034	2994	1729	4316	1809
	42	5446	4728	4049	3950	2481	3521	2928	1330	1453	4166	3844	660
	43	3075	5271	4449	4002	2129	1714	3976	665	500	892	2335	2243
	44	2205	2382	4322	2388	74	1939	4294	39	2410	6239	2894	157
	45	4892	4843	8	1145	2662	4373	4586	3408	634	2217	4705	4806
	46	2843	76	2490	1343	4710	4617	3545	1727	2914	2360	1711	5006
	47	2821	4852	1352	2218	2343	4800	2715	2219	6240	2089	5129	804
	48	380	2340	1811	2310	4804	862	4656	3865	539	2265	2069	4423
Bloc 5	49	2217	2360	4294	1727	1915	4002	4415	2669	2204	682	5347	5328
	50	39	4711	8	4449	4576	4166	1731	2069	2067	2388	660	2924
	51	4656	2156	665	4373	4076	2343	2129	2361	2137	2312	4002	4728
	52	1236	2265	1330	4804	1813	3902	2218	1693	3936	3950	4360	300
	53	4892	2243	4853	804	2348	76	4589	3844	2928	2346	4374	4852
	54	4516	4322	74	1343	2241	4806	2340	3976	4034	1145	2843	5129
	55	157	3862	2358	2635	2410	3698	1793	2458	3597	1711	2712	4793
	56	4710	4586	1809	2715	3545	2219	1939	2205	6239	3545	4617	2734
	57	5271	5328	4705	682	539	2382	4316	2089	1693	4843	4079	1352
	58	5446	1453	2204	4049	862	2481	634	4456	4800	5006	4445	3967
	59	2662	2994	3865	3075	2669	3521	2310	2490	4415	1714	4853	4650
	60	5810	5665	4394	5347	500	4811	2311	1814	2894	5304	4167	4852
	61	6039	4896	3845	2335	2821	1811	892	380	2914	6240	4423	2215

N

0000 : Pas de production en 2019

Première feuille 2019 (greffage 2018)

première feuille 2021 (Greffage 2020)

Première feuille 2020 (greffage 2019)

Liste des accessions d'abricotier du panel de diversité.

Accession	Dénomination	Provenance	Accession	Dénomination	Provenance	Accession	Dénomination	Provenance	Accession	Dénomination	Provenance
A1145	Stark Early Orange	T5A	A2243	Ivresse	T5A	A3865	Flavorella	T5A	A4793	Marvinka	T6A
A1236	Manicot	T5A	A2265	Gâterie	GOTHERON	A0039	Précoce Ampuis	T5A	A4800	Sublime	T6A
A1314	Arrogante	T5A	A2310	Modesto	T5A	A4034	115x1267 (95A65)	T6A	A4804	Greta	T6A
A1330	Rouge de Mauves	T5A	A2311	Lambertin n°1	T5A	A4076	Late Cot	T5A	A4806	Apribang	T6A
A1343	Canino	T5A	A2312	Flamingold	T5A	A4079	NJA82	T5A	A4892	Red Spring	T6A
A1352	Polonais	T5A	A2340	Saturn	T5A	A4166	115x3325-28	T5A	A4896	Mediabel	T6A
A1453	Perfection	T5A	A2343	Olimp	T5A	A4167	115x3325-31	T5A	A0500	Moniqui	T5A
A0157	Rouge du Roussillon	T5A	A2346	TIMPURII KITINAU	T5A	A4294	Incomparable de Malissard	T6A	A5006	OT1	T6A
A1711	Avikaline	T5A	A2348	Everani	T5A	A4316	Augusta 3	T6A	A5129	Major	T6A
A1793	Tardif de Bordaneil Type1	T5A	A2358	Héléna du Roussillon	T5A	A4322	Priboto	T5A	A5304	Pricia	T6A
A1809	Précoce de Tyrinthe	T5A	A2388	11N25	T5A	A4373	Bo 90 610 010	T5A	A5328	05F097	T6D
A1811	Harcot	T5A	A2458	Royal Roussillon	T5A	A4374	Murciana	T6A	A0544	Cafona	T5A
A1813	Laycot	T5A	A2490	Tardif de Tain	T5A	A4415	Wonder Cot	T6A	A0660	Bergeron	T5A
A1814	Hargrand	T5A	A2662	Alfred	T5A	A4423	Congat	T6A	A0682	Morden 604	T5A
A1939	Fantasme	T5A	A2734	Boucheran Boutard	T5A	A4445	Farbaly	T6A	A0074	Jaubert Foulon	T5A
A2067	Marouch n°14	T5A	A2821	Frisson	T5A	A4456	3576x4293-58	T5A	A0076	Pêche de Nancy	T5A

A2089	Bebeco	T5A	A2894	Orangered Standard	T5A	A4516	Gilgat	T6A	A0008	Colomer	T5A
A2129	Rouge de Fournes	T5A	A2924	Goldbar	T5A	A4586	Primaya	T6A	A0862	Pseudo Royal	T5A
A2137	Bakour	T5A	A2928	Early Blush	T5A	A4589	Farhial	T6A			
A2156	Veecot	T5A	A3408	n29 Ajouc	T5A	A4617	2914x2265-48	T5A			
A2204	Bebeco	T5A	A3521	G1 2122-11	T5A	A4650	F02A094	GOTHERON			
A2205	Ansdwee	T5A	A3597	2241x2218-55	T5A	A4656	Liligat	T6A			
A2217	Sunglo	T5A	A3698	Vestar	T5A	A4705	Latica	T6A			
A2218	Goldrich	T5A	A3844	Ravicille	T6A	A4710	Select 98	T6A			
A2219	Vivagold	T5A	A3862	Robada	T5A	A4728	2914x2265-48	T5A			

Effect of spectral pretreatment on phenomic selection accuracy using first derivative (Der1), second derivative (Der2), de-trending transformation (Detrend), normalization (Norm), normalization coupled to first derivative (Norm+Der1) and normalization coupled to second derivative (Norm+Der2) compared to raw spectral data based on NIR and MIR.