Localisation de régions du génome du pommier contrôlant la variation de caractères de qualité du fruit et de résistance aux maladies : signatures de sélection et génétique d'association

Diane Leforestier

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

université angers

Thèse de Doctorat

Diane LEFORESTIER

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université d'Angers
sous le label de L'Université Nantes Angers Le Mans

École doctorale : VENAM
Discipline : Biologie Cellulaire Végétale
Unité de recherche : Institut de Recherche en Horticulture et Semences (UMR 1345 IRHS)

Thèse N° 1478

Soutenue le 29 juin 2015

Localisation de régions du génome du pommier contrôlant la variation de caractères de qualité du fruit et de résistance aux maladies : signatures de sélection et génétique d'association

JURY

Rapporteurs :	**Brigitte COURTOIS**, Chercheur, CIRAD Montpellier **Yves VIGOUROUX**, DR2, IRD Montpellier
Examinateurs :	**Stéphanie MARIETTE,** CR1, INRA Bordeaux **Philippe SIMONEAUX**, Professeur, Université d'Angers
Directeur de Thèse :	**Charles-Eric DUREL**, DR2, INRA Angers

Remerciements

Et voilà, dernière partie à rédiger, enfin, les remerciements. Les quelques mots écrits ci-dessous ne sont bien évidemment pas représentatifs de toute la gratitude que j'éprouve envers toutes les personnes que j'ai côtoyées et qui m'ont aidée, de près ou de loin, pendant ma thèse.

Je tiens tout d'abord à remercier mon directeur de thèse, **Charles-Eric**, de m'avoir accueillie dans son équipe et de m'avoir amenée jusqu'ici. Merci de votre patience, d'avoir su m'aiguiller vers les bonnes personnes quand cela a été nécessaire, d'avoir pu vous rendre disponible pendant la rude période de la rédaction et pour les cours de stats express grâce auxquels je suis passée du niveau « un écart-ty… quoi ? » à un niveau, bien que basal, que je n'aurais jamais pensé atteindre.

Merci à **Brigitte Courtois** et **Yves Vigouroux** d'avoir accepté d'être rapporteurs de ma thèse, et à **Stéphanie Mariette** et **Philippe Simoneau** d'être mes examinateurs.

Merci à **Hélène**, arrivée au milieu de la bataille, mais qui a été d'une aide précieuse. Merci d'avoir partagé mes déboires et désillusions, d'avoir été aussi disponible et pour tes conseils et les relectures des différents manuscrits.

Merci à **Elisa**, Minitruc, mon chaton, et **Caro**, Charoline Denance, de m'avoir soutenue, aidée, fournie en chewing-gums, motivée, enfin bref d'avoir été là au quotidien pour supporter mes plaintes et d'avoir su me redonner courage !

Un grand merci à **Tatiana**, **Antoine** et **Christophe** d'avoir accepté de m'aider à un moment où je pensais que tout était fichu, de m'avoir initiée à la génétique des populations et d'avoir permis à mon article tant attendu de voir le jour.

Merci à **Vincent Segura**, **Stéphane Nicolas**, **Jérémy Clotault** et **Patrice This** d'avoir accepté de faire partie de mon Comité de Suivi de Thèse ainsi que pour tous leurs conseils et la vision extérieure qu'ils ont apportée à mon projet de thèse.

Merci à **Thibault** pour ses nombreux conseils et les discussions que nous avons eues tout au long de ma thèse et qui m'ont permis d'y voir un peu plus clair dans mes analyses de génétique des pops. Tu vois au final on se comprend pas si mal que ça !!

Merci à **Arnaud G.** et **Laurence Feugey** pour avoir mis à notre disposition les ressources génétiques de l'INRA. Merci Arnaud également pour ton aide concernant les données de qualité du fruit.

Merci à **Nicolas**, **Michel** et **Lysiane** pour le soin particulier qu'ils ont apporté à mes plantes en serre et grâce à qui les essais tavelure et feu se sont passés sans anicroche.

Merci à **Jean-Pierre**, pour ses nombreuses anecdotes sur sa vie au NIH, et quantité d'autres sujets passionnants qui ont animé les pauses !

Merci à **Jean-Marc** pour les pauses du vendredi, pour avoir joué le rôle du rabat-joie quand il s'agissait de commencer à écrire le manuscrit, pour la carte et pour m'avoir empêché d'ingurgiter trop de calories en me volant sans cesse mes gâteaux. Tu vas pouvoir manger les Granola que tu m'as dédicacés !

Merci à **Inès** entre autres pour ses scripts R, pour les pauses de l'après-midi, pour avoir toujours eu le sourire et pour m'avoir encouragée jusqu'au bout. On va pouvoir se la faire notre Ladie's Night !

Merci à **Sylvain G.**, mon héros, d'avoir répondu à mes questions sur la « merveilleuse » séquence du pommier et d'avoir résolu bon nombre de mes problèmes de scripts, toujours avec le sourire !

Merci aux gars de l'équipe VadiPom, **Mehdi**, **François** et **Bernard** pour m'avoir initiée aux joies de la pollinisation et du grattage de fleurs, mais surtout pour leur gentillesse et leur bonne humeur à toute épreuve et les nombreux after-works passés en leur charmante compagnie.

Merci aux membres de notre super équipe administrative, **Patricia V.**, **Patricia R.**, **Sylvie**, **Magali** et **Valérie** qui gèrent l'unité d'une main de maître. Merci également, bien entendu, à **Anne** sans qui ces trois années à l'INRA n'auraient pas été aussi mouvementées et agréables, malgré le vol de paires de chaussures et une voiture pleine de ballons! C'est vraiment plus pareil sans toi.

Merci à **Philippe Barre** d'avoir partagé avec moi son script magique et d'avoir pris le temps de me l'expliquer en détails.

Merci à **Roland** pour les préparations d'inoculum de feu et pour sa bonne humeur qui nous manque, enfin « je pense ».

Merci à l'ensemble des habitants du Bâtiment B, et notamment **Marie-Charlotte**, **Clément**, **Laurence**, **David**, **Philippe**, **Rémi**, **Agathe**, **Julien Jeauffre**, **Annie** et tous les autres pour leur gentillesse et tout particulièrement pour avoir servi de cobayes pour mes diverses recettes ! Je ne peux qu'espérer retrouver une ambiance de travail comme celle-là un jour !

Merci aux filles du bureau des doctorantes pour cette ambiance studieuse bien évidemment, mais pas que ! Merci donc à **Zindy**, le Kinder en chef, à **Marie**, sauf pour la poisse que tu as laissée dans le bureau en partant, à **Sandrine**, pour ses encouragements, ses critiques constructives sur mes pâtisseries, et d'une manière générale pour sa présence quotidienne. A nos déboires !! Merci à **Mathilde**, qui sous ses airs de gentille s'est révélée, à mon plus grand bonheur, être aussi vilaine que nous autres et à **Manue**, ma meilleure amie !! pour tous ces moments que l'on a partagés. C'est ton tour maintenant tu vas voir c'est trop fun !

Merci à mon collègue de thèse et ami **Arnaud**, pour tellement de choses que résumer ces presque 4 ans en quelques mots ne serait pas leur rendre justice. Merci d'avoir été là et d'être différent (dans tous les sens du terme !!). Merci à la lumière de l'univers, **Jean-François**, d'avoir éclairé bon nombre de mes journées et d'être juste toi (non je n'arrêterai pas !!), et à la huitième merveille du monde, **Benjamin** malgré tes tentatives répétées pour me voler Arnaud ! Merci à **Marie**, la mal aimable, de réussir à me faire passer pour quelqu'un d'avenant ! On se retrouve bientôt pour la boutique, prépare ton costume de muffin ! Merci également à **Christophe**, mon « amour », ma **Kékette** et **Piau** d'être toujours présents et de m'avoir soutenue.

Enfin, merci à ma famille et surtout à mes parents de m'avoir encouragée à aller le plus loin possible dans mes études. Merci **Murt** de faire semblant de m'écouter quand je te parle de déséquilibre de liaison et de génétique d'association, et de supporter mon fichu caractère. Merci à mon père pour ledit caractère et pour tout le reste au final. J'aurais vraiment aimé que tu puisses voir ça. C'est pas un prix Nobel mais j'espère que tu aurais quand même été content. Merci également à mes sœurs chéries **Charlotte** et **Sylve**, à leurs conjoints **Guillaume** et **Romain**, et à mes trois merveilleux neveux et nièce, **Louis**, **Thomas** et **Astrid** de toujours savoir me redonner le sourire même dans les moments les plus difficiles. Merci à mes deux cousins préférés **Eric**, pour m'avoir initiée aux joies du codage, pour m'avoir hébergée lors de mes séjours à Orsay, et **Marc**, pour l'ajout au script perl !

Table des matières

Liste des abréviations

ACP	analyse en composantes principales
ADN	acide désoxyribonucléique
ANOVA	analysis of variance (analyse de variance)
ARN	acide ribonucléique
AUDPC	area under the disease progress curve (aire sous la courbe de progression de la maladie)
BIC	bayesian information criterion (critère d'information Bayésien)
CC	core collection
cM	centimorgan
DL	déséquilibre de liaison
EWAS	epigenome-wide association study (étude d'association sur épigénome entier)
GWAS	genome-wide association study (étude d'association sur génome entier)
HR	hypersensitive response (réaction d'hypersensibilité)
INRA	Institut National de la Recherche Agronomique
FDR	false discovery rate (taux de faux positifs)
LG	linkage group (groupe de liaison)
LRR	leucine rich repeat (répétitions riches en leucine)
LOD	logarithm of odds ratio (logarithme des probabilités)
Mb	méga bases
NBS	nucleotide binding site (site de liaison à des nucléotides)
pb	paire de bases
QTL	quantitative trait locus (locus de caractère quantitatif)
ROS	reactive oxygen species (espèces réactives de l'oxygène)
SAM	sélection assistée par marqueurs
SAR	systemic acquired resistance (résistance systémique acquise)
SFS	site frequency spectrum (spectre de fréquences alléliques)
SNP	single nucleotide polymorphism (polymorphisme à un site nucléotidique)
SSR	single sequence repeat (séquence simple répétée, en général microsatellites)

Table des illustrations

Table des tableaux

Table des annexes

Introduction générale

L'exploration des bases génétiques contrôlant la variation des caractères phénotypiques, concernant par exemple des maladies chez l'homme (Altshuler, Daly, and Lander 2008) ou des traits d'intérêt agronomique chez les animaux et les plantes (Holland 2007), s'est révélée être d'une importance capitale pour une meilleure compréhension des mécanismes mis en jeu lors de l'expression de ces caractères. Un grand nombre d'études ont ainsi permis la localisation de gènes, et, dans une certaine mesure, l'identification des voies métaboliques impliquées suite au clonage de ces gènes (Salvi and Tuberosa 2005). Bien que de grandes avancées dans le domaine de la sélection génomique (Goddard and Hayes 2007) aient été réalisées et aient propulsé en avant la sélection de certaines espèces comme les bovins laitiers, sans connaissance précise de tous les déterminants génétiques des caractères ciblés (Hayes et al. 2009), une meilleure connaissance de l'architecture génétique des caractères reste nécessaire, tant du point de vue académique (connaissance de la diversité génétique, compréhension des mécanismes physiologiques sous-jacents), que du point de vue pratique (exploitation de la variabilité génétique, sélection assistée par marqueurs sur des régions ciblées du génome, construction assistée par marqueurs de combinaisons génotypiques particulières).

Chez les plantes, et notamment les plantes pérennes, les études de cartographie sont longues et coûteuses à mettre en place. Les études, réalisées sur des descendances F1 ou dans des populations de pedigree, ne permettent d'explorer qu'une faible partie de la diversité génétique disponible. De plus, les effectifs réduits de ces descendances résultent en une puissance de détection et une précision de la localisation des facteurs génétiques recherchés relativement faibles (Collard et al. 2005). L'approche de la génétique d'association (Balding 2006; Yu and Buckler 2006; Ingvarsson and Street 2011), développée lors des deux dernières décennies, permet de s'affranchir de bon nombre des limites imposées par la cartographie en descendances en rendant possible l'étude d'individus non apparentés, et donc potentiellement porteurs d'une large diversité génétique.

Le pommier, *Malus domestica*, est une espèce importante des points de vue économique et culturel. De nombreux programmes de sélection, dont certains se déroulent à l'INRA d'Angers, ont pour but de sélectionner des nouvelles variétés de pommiers présentant de bonnes qualités gustatives mais également des niveaux de résistance aux maladies élevés. Des études de cartographie en descendances sur cette espèce ont permis l'identification de bon nombre de facteurs génétiques contrôlant la variation de ces caractères d'intérêt, tout en étant limitées à l'étude de quelques allèles.

L'objectif de cette thèse était donc d'étudier une large diversité génétique chez des variétés anciennes de pommiers cultivé afin de rechercher de nouvelles régions du génome contrôlant la variation de caractères agronomiques importants en sélection, la qualité du fruit et la résistance à la tavelure et au feu bactérien, mais également de chercher à préciser l'emplacement sur le génome de facteurs génétiques identifiés auparavant. Ces nouvelles informations sur le déterminisme génétique de ces caractères devraient être utiles pour de futurs programmes de sélection, et notamment ceux utilisant la sélection assistée par marqueurs.

Le premier chapitre de ce manuscrit est consacré dans un premier temps à la description des différentes méthodes permettant l'identification et la localisation de facteurs génétiques responsables des variations phénotypiques. Dans un second temps, nous décrirons le modèle d'étude utilisé ici, le pommier, ainsi que les différents traits phénotypiques d'intérêt majeur dans les programmes de sélection récents. Une troisième partie fera un bref inventaire des études de cartographie génétique ayant permis l'identification de QTLs et gènes gouvernant ces traits.

Un second chapitre, dédié aux matériels et méthodes employés, décrira les core collections étudiées dans ce travail de thèse, les techniques de génotypage utilisées ainsi que les divers tests statistiques appliqués aux données.

Un troisième chapitre s'attachera à étudier les bases génétiques des différences phénotypiques entre les pommes à cidre et les pommes à couteau. Deux core collections de pommiers à cidre et à couteau, génotypées à l'aide d'une puce de génotypage faible densité, ont été étudiées. Cette étude a fait l'objet d'un article accepté pour publication dans la revue *Evolutionnary Applications*.

Un quatrième chapitre concernera l'étude du déséquilibre de liaison, nécessaire à toute étude de génétique d'association, et la comparaison de plusieurs techniques de génotypage pour l'estimation de son étendue, au niveau du génome entier et dans des régions spécifiques (dans le cadre d'une AIP Bio-ressources).

Enfin, un cinquième chapitre sera consacré à l'étude de génétique d'association, réalisée dans le cadre du projet européen FruitBreedomics, sur une core collection de variétés anciennes de pommes à couteau dans le but d'identifier de nouvelles sources génétiques de résistance aux maladies et de qualité du fruit.

Chapitre 1

Synthèse bibliographique

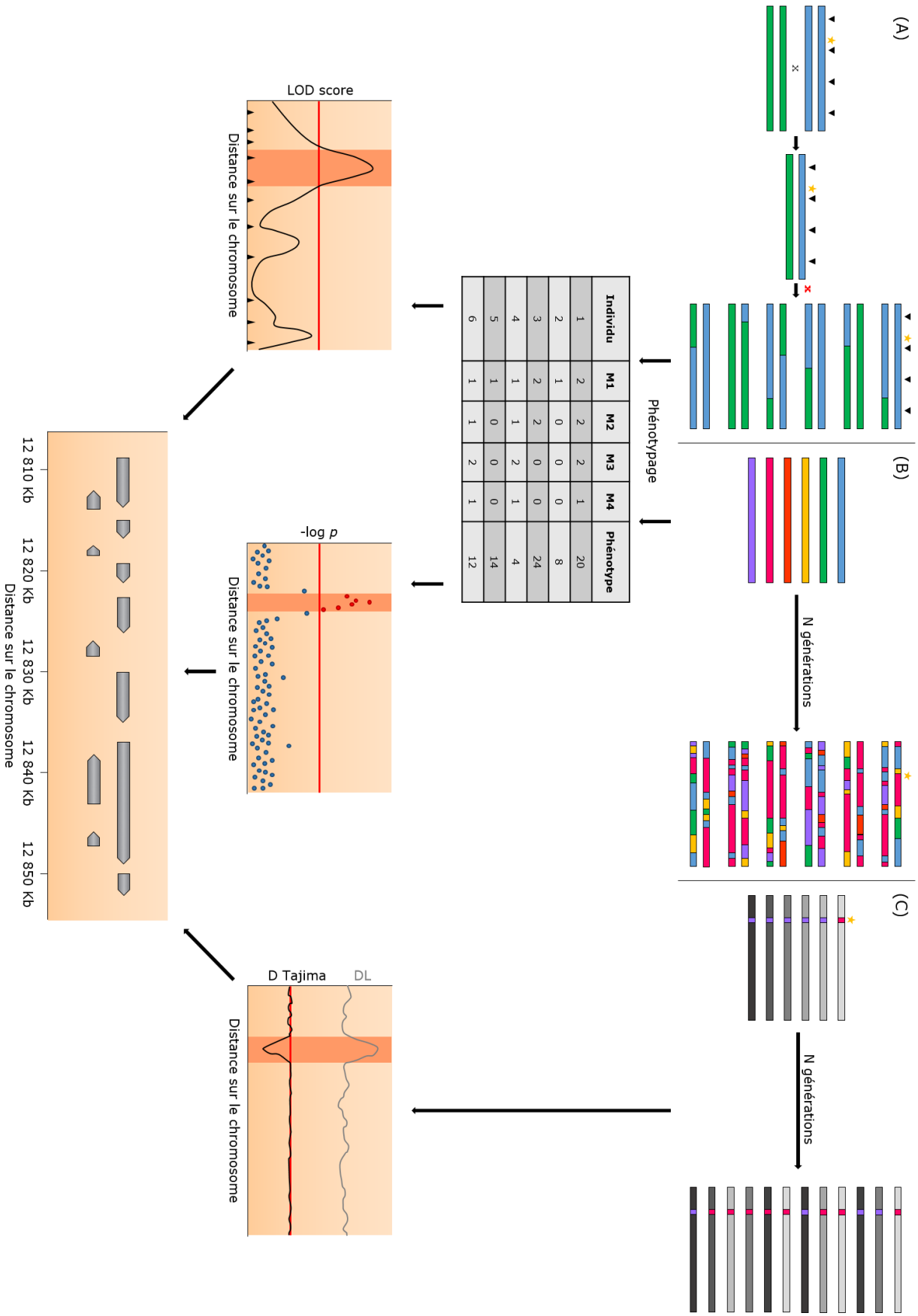


Figure 1 : Principes de base de la cartographie par liaison génétique (A), de la cartographie par génétique d'association (B) et des analyses de génétique des populations sur génome entier (C), adapté de Mackay et al. (2009)

1. La cartographie génétique

Chez les plantes, un grand nombre de caractères agronomiques d'intérêt (rendement, date de floraison, qualité, résistances biotiques ou abiotiques…) sont sous contrôle génétique complexe, reflété par une distribution continue (Mackay, Stone, and Ayroles 2009; Alonso-Blanco and Méndez-Vigo 2014). Avec l'avancée des technologies et des méthodologies, les outils dont nous disposons, et notamment les marqueurs moléculaires, sont de plus en plus adaptés à la recherche des facteurs génétiques responsables de la variation de ces caractères. Le domaine de la biologie dédié à ces études, la génétique quantitative, propose plusieurs alternatives permettant d'identifier et de localiser les zones du génome porteuses des gènes contrôlant la variation des caractères d'intérêt.

1.1. Cartographie « classique » par analyse de liaison

La première approche développée à des fins de localisation des gènes contrôlant la variation des caractères d'intérêt agronomique est la cartographie de loci de caractères quantitatifs (« Quantitative Trait Loci » ou QTLs) par analyse de liaison (Collard et al. 2005). Celle-ci est réalisée sur des individus issus de croisements contrôlés sur une ou plusieurs générations et possédant un degré d'apparentement homogène et connu (familles de pleins frères, descendances F2, descendances back-cross, lignées recombinantes…). Le principe de base consiste à construire une carte (de liaison) génétique à l'aide de marqueurs, en étudiant leur ségrégation au sein des individus de la (des) famille(s) étudiée(s), et à phénotyper en parallèle ces individus (c'est-à-dire de les mesurer ou les évaluer pour le ou les caractères d'intérêt de l'étude), puis de rechercher des corrélations entre marqueurs et phénotypes (Figure 1A). Les principaux marqueurs génétiques utilisés dans ce type d'étude ont longtemps été les RFLP (Restriction Fragment Length Polymorphisms), les RAPD (Random Amplification of Polymorphic DNA), les AFLP (Amplified Fragment Length Polymorphisms), et les microsatellites ou SSR (Single Sequence Repeats ; Phillips and Vasil 2001). Hormis les RFLP qui correspondent à une hybridation de sondes sur de l'ADN digéré, tous ces marqueurs sont des séquences d'ADN amplifiées sur le génome et distinguées grâce à leur taille sur gel ou grâce au séquençage. Plus récemment, les marqueurs SNP (Single Nucleotide Polymorphism) se sont généralisés grâce notamment à leur utilisation possible à très haute densité et de façon largement automatisée.

1.1.1. Cartographie en descendances simples ou connectées

La grande majorité des études de cartographie de QTLs a été réalisée sur des descendances simples telles que les descendances F2 pour les espèces autogames, ou F1 pour les espèces allogames. La puissance et la résolution des analyses de cartographie dépendent essentiellement de la taille des populations étudiées (Collard et al. 2005), la densité de marqueurs sur les cartes n'intervenant qu'en second lieu, dans l'approche de cartographie par intervalle (« Interval Mapping »). Cette approche a cependant permis le clonage de gènes sous-jacents à des QTLs d'intérêt dont la localisation précise a été effectuée en augmentant la taille de la population et la densité locale de marqueurs (Salvi et al. 2007; Chapman et al. 2012).

Un inconvénient majeur de la détection de QTLs en descendances simples tient au fait que l'effet estimé des QTLs peut se révéler spécifique du fond génétique de la population analysée. Un autre inconvénient notable de cette approche est la très faible représentativité de la diversité génétique d'une espèce, puisque seulement deux parents sont concernés dans le croisement initial. Le nombre et la position des QTLs détectés peuvent donc être très variables d'une descendance à l'autre. Des approches de détection en descendances multiples connectées (plans de croisement de type factoriel ou diallèle) ont donc été proposées afin d'élargir la diversité génétique étudiée de manière à mieux explorer l'architecture génétique des caractères (Muranty 1996). Des QTLs ont ainsi été détectés, en nombre plus important qu'en descendances simples, et des tests d'épistasie entre les QTLs et les différents fonds génétiques ont été réalisés (Blanc et al. 2006; Billotte et al. 2010).

1.1.2. Cartographie en pedigree

Afin de mieux tirer parti des populations d'amélioration des programmes de sélection, une approche de cartographie de QTLs a également été développée en utilisant des individus (ou des familles) reliés entre eux par des relations d'apparentement plus complexes au sein d'un pedigree connu (Bink et al. 2002). Cette approche permet d'étudier une diversité plus importante que dans le cas de la cartographie en descendances simples. Elle permet également d'estimer les effets des QTLs dans le fond génétique global de la population étudiée grâce à l'utilisation d'un modèle Bayésien comme celui implémenté dans le logiciel FlexQTL (Bink et al. 2002; Bink et al. 2008). En complément de la grande souplesse d'implémentation de l'approche Bayésienne à des situations de pedigree complexes, cette méthode de détection, quand elle est appliquée à des populations d'amélioration en cours d'exploitation, apporte au sélectionneur l'avantage de pouvoir suivre le long des générations d'amélioration la probabilité de transmission des allèles favorables et défavorables aux QTLs détectés,

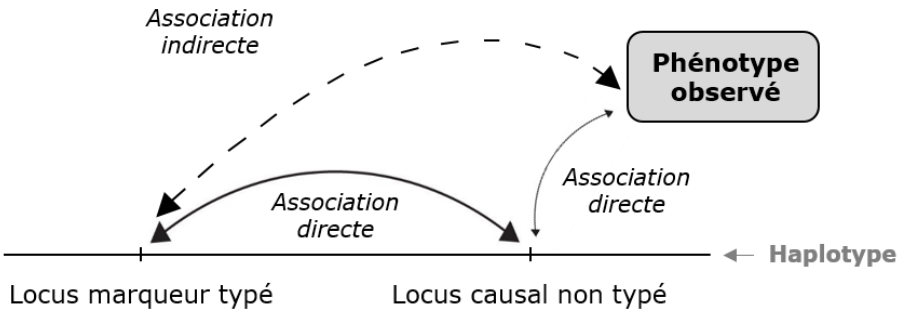


Figure 2 : Schématisation de la relation triangulaire à la base du principe de la génétique d'association, d'après Balding (2006)

et d'estimer les « breeding values » (valeurs reproductives) de l'ensemble des individus (Bink et al. 2014). Cette approche reste cependant limitée à des pedigrees connus et ne permet donc pas d'exploiter la diversité globale de l'espèce étudiée.

1.2. La génétique d'association

Les deux dernières décennies ont vu le développement d'une nouvelle méthode de cartographie des caractères d'intérêt, la génétique d'association ou cartographie par déséquilibre de liaison (Yu and Buckler 2006; Rafalski 2010; Brachi, Morris, and Borevitz 2011; Ingvarsson and Street 2011). L'utilisation de cette méthode a été rendue possible par l'amélioration et la baisse du coût des techniques de génotypage haut débit telles que les puces de génotypage, ou le re-séquençage de génomes entiers. A l'inverse des méthodes de cartographie « classique » décrites ci-dessus, la génétique d'association permet l'étude de populations d'individus non apparentés, ce qui a grandement facilité la cartographie de gènes d'intérêt dans des espèces chez qui les croisements contrôlés et l'étude des descendances étaient trop coûteuses ou trop longues pour pouvoir être envisagées.

1.2.1. Principes de base de la génétique d'association

A la différence des approches de cartographie « classique », la génétique d'association permet l'étude simultanée d'un nombre d'individus et de fonds génétiques différents. La génétique d'association a été développée au début chez l'humain pour rechercher des associations génotype-phénotype dans un cadre où des descendances en ségrégation de grandes tailles sont impossibles à obtenir (Spielman, McGinnis, and Ewens 1993). Cette approche utilise comme échantillon un groupe d'individus non apparentés formant deux cohortes aux phénotypes contrastés, par exemple une cohorte d'individus sains et une cohorte d'individus malades pour une pathologie humaine donnée. Ces individus sont génotypés pour différents marqueurs, puis des corrélations sont recherchées entre le phénotype et les allèles des marqueurs (Transmission/Disequilibrium Test ou TDT). Plus récemment, les analyses de génétique d'association portant sur des caractères continus et complexes se sont développées. Les populations étudiées dans ces analyses sont composées d'individus présentant de grandes variations phénotypiques pour un ou plusieurs caractères (Manolio 2009), et sont souvent basées sur la maximisation de la diversité génétique (« core collections » ; Frankel and Brown 1984) afin d'étudier un groupe représentatif de l'espèce. Le principe de base de la génétique d'association (Figure 1B) repose sur une relation triangulaire (Figure 2) entre le phénotype observé, le locus causal de la variation du phénotype et un locus marqueur en déséquilibre de liaison avec le locus causal (Balding 2006). La

cartographie par génétique d'association s'appuie donc sur une propriété génétique des populations naturelles et domestiques : le déséquilibre de liaison.

1.2.2. Le déséquilibre de liaison

Le déséquilibre de liaison (DL) peut être défini comme l'association non aléatoire entre les allèles de différents loci. A la différence de la liaison génétique (L) qui est définie à l'échelle de l'individu et qui traduit simplement la proximité physique sur le génome de deux loci « liés », le DL est défini à l'échelle d'une population et n'implique pas nécessairement une proximité physique. La liaison physique favorise cependant la mise en place d'un DL. Dans une population sexuée et panmictique, de taille efficace infinie et soumise ni à la sélection, ni à la mutation, ni à la migration, aucun déséquilibre de liaison n'est attendu, les allèles des différents loci étant distribués aléatoirement dans les gamètes. A l'inverse, le DL peut apparaître et croître en fonction de divers processus évolutifs : la dérive, la sélection et la fragmentation des populations. Depuis quelques décennies, un intérêt grandissant pour le DL a vu le jour, principalement dû au fait qu'il permet, au travers des études d'association, de localiser les loci responsables de nombreuses maladies humaines (Risch and Merikangas 1996), mais également de mieux comprendre l'histoire évolutive des espèces (Nordborg and Tavare 2002). Les études d'association détectent en effet des marqueurs qui sont en déséquilibre de liaison avec les loci contrôlant le phénotype.

1.2.2.1. Mode de calcul du déséquilibre de liaison

Le DL se mesure grâce à plusieurs estimateurs, dont les plus communs, D' (Lewontin 1964) et r^2 (Hill and Robertson 1968), reflètent chacun des aspects différents de l'histoire évolutive des loci (Flint-Garcia, Thornsberry, and Buckler 2003). Considérons deux loci ayant chacun deux allèles A/a et B/b. On note les fréquences alléliques de chacun π_A, π_B, π_a et π_b, et les fréquences des haplotypes résultant π_{AB}, π_{Ab}, π_{aB} et π_{ab}. A l'équilibre, les fréquences haplotypiques observées sont égales aux produits des fréquences des allèles ($\pi_{AB} = \pi_A\pi_B$). Le calcul des deux estimateurs repose sur la différence entre les fréquences haplotypiques observées et attendues, qui se traduit en :

$$D_{ab} = \pi_{AB} - \pi_A\pi_B$$

Cependant, D a des propriétés qui ne sont pas toujours satisfaisantes. Il mesure un écart absolu entre les fréquences, dont il dépend directement. Le paramètre D' normalise le déséquilibre de liaison par

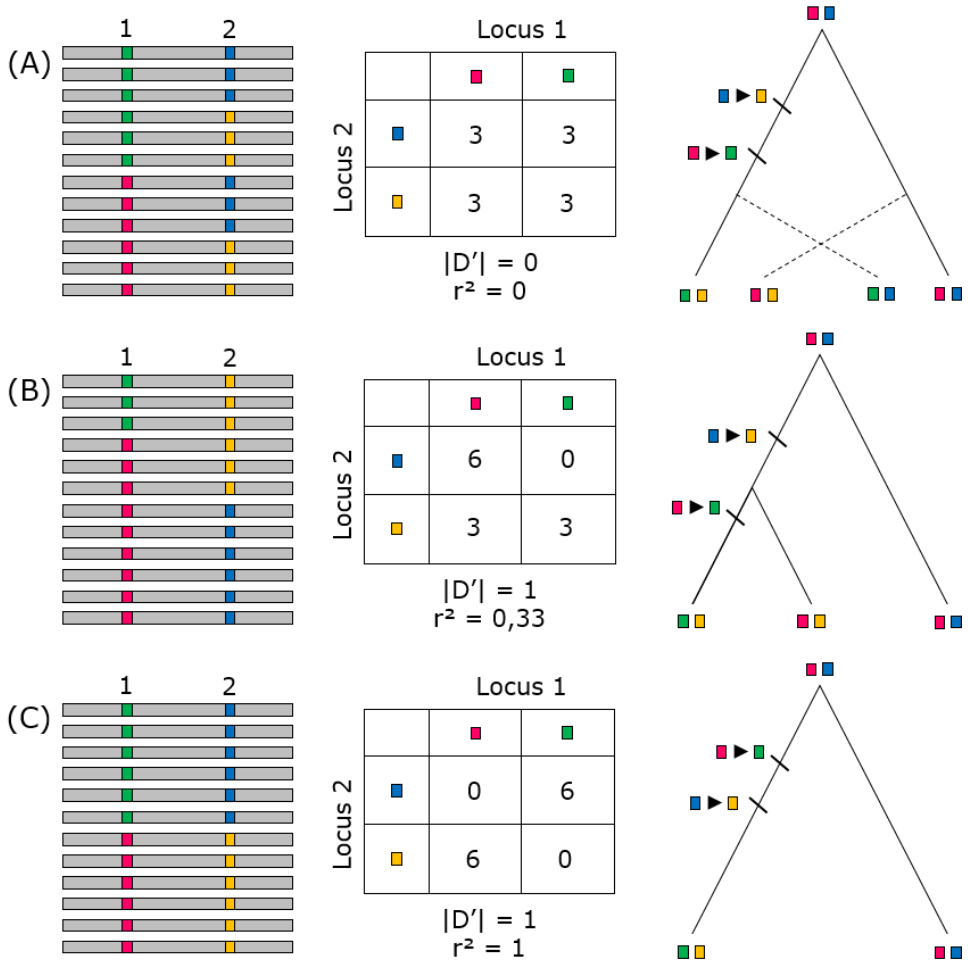


Figure 3 : Exemples de trois scenarii évolutifs et des valeurs prises par D' et r², d'après Flint-Garcia et al. (2003)

(A) les loci sont en équilibre de liaison, il n'y a pas d'association préférentielle entre les allèles aux deux loci ; deux événements de mutation ont eu lieu suivis d'événement de recombinaison engendrant toutes les combinaisons alléliques possibles et ce dans les mêmes fréquences ; (B) les loci sont en déséquilibre de liaison faible ; une des classes alléliques est manquante, une mutation ayant eu lieu sur une seule des trois branches de l'arbre et aucune recombinaison n'ayant eu lieu entre les individus des différentes branches ; (C) les loci sont en déséquilibre de liaison complet ; deux événements de mutation ayant eu lieu sur une seule branche de l'arbre, sans aucune recombinaison entre les individus résultant

rapport à sa valeur maximale en considérant les fréquences alléliques, et se calcule de la manière suivante :

$$D' = \frac{(D_{ab})^2}{\min(\pi_A \pi_{b,} \pi_a \pi_B)} \quad pour\ D_{ab} < 0$$

$$D' = \frac{(D_{ab})^2}{\min(\pi_A \pi_{B,} \pi_a \pi_b)} \quad pour\ D_{ab} > 0$$

D' a l'avantage de varier entre 0 et 1 ce qui permet des comparaisons entre paires de loci ou entre populations. Il est cependant très sensible aux tailles de populations et à la rareté des allèles. Le r^2 est la mesure du DL la plus utilisée actuellement et est similaire à un coefficient de corrélation au carré entre les allèles des deux loci ; il se calcule de la façon suivante et présente également l'avantage de varier entre 0 et 1 :

$$r^2 = \frac{(D_{ab})^2}{\pi_A \pi_B \pi_a \pi_b}$$

Les différents estimateurs du DL ont des comportements et des sensibilités différents aux histoires des loci étudiés. Le r^2 reflète plus particulièrement les histoires de recombinaison et de mutation, contrairement au D' qui reflète uniquement l'histoire de recombinaison. La figure 3 montre plusieurs scenarii évolutifs de deux allèles au sein d'une population ainsi que les valeurs prises par les deux estimateurs en fonction des événements (Flint-Garcia, Thornsberry, and Buckler 2003).

1.2.2.2. Facteurs influençant l'ampleur du déséquilibre de liaison

Plusieurs facteurs sont à l'origine de la mise en place, de la persistance ou de la réduction du DL (Gupta, Rustgi, and Kulwal 2005). Parmi les facteurs entrant en jeu dans la création et le maintien du DL entre allèles se trouvent :

- les mutations, lorsqu'elles se produisent avec une faible probabilité pour un site donné : l'existence du déséquilibre de liaison repose sur l'apparition de mutations responsables de la création de nouveaux allèles qui peuvent former des blocs haplotypiques dans lesquels le DL sera très élevé ; par exemple, si la sélection a rapidement et récemment fait augmenter un allèle en fréquence (Figure 4) de telle sorte que la recombinaison n'a pas eu le temps de « casser » les associations avec les allèles aux sites proches physiquement, l'allèle au site sélectionné est alors lié aux allèles des sites proches du fond génétique dans lequel la mutation est apparue

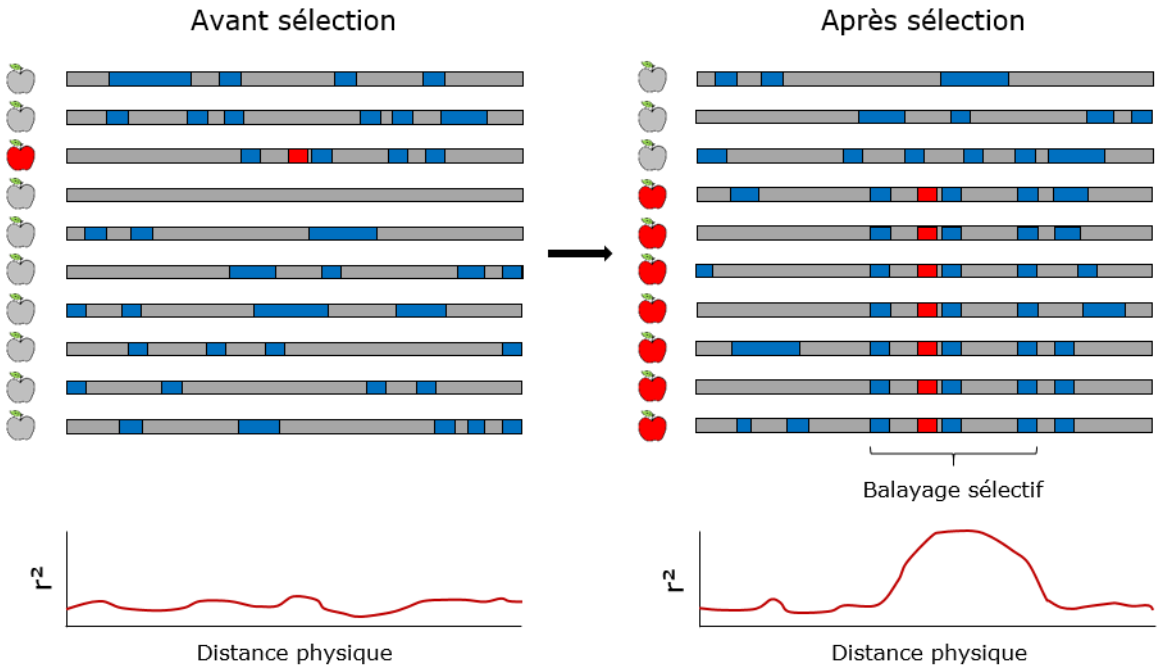


Figure 4 : Illustration d'un balayage sélectif, adaptée de Schaffner et Sabeti (2008)
une mutation favorable (en rouge) apparaît chez un individu d'une population ; au fil des générations cette mutation, sous l'effet de la sélection, va voir sa fréquence augmenter dans la population ; les loci neutres en liaison physique avec cette mutation vont également augmenter en fréquence et créer un bloc haplotypique à l'intérieur duquel le déséquilibre de liaison sera élevé

(Ytournel 2008) ; le rôle des mutations dans la mise en place du DL nécessite une autre force évolutive, la sélection ; lors d'événements de sélection, naturelle ou non, si un groupe d'individus est porteur d'une combinaison d'allèles ou d'un allèle favorables, les différents loci sélectionnés ainsi que les loci neutres à proximité seront en DL (Kim and Nielsen 2004) ;

- la migration des populations : lors de l'hybridation entre deux populations initialement différenciées, les différences de fréquences alléliques dans les populations initiales engendrent un déséquilibre de liaison (Pfaff et al. 2001) qui va rapidement décroître lors des générations suivantes (Figure 5);
- l'autogamie : chez les espèces autogames, le taux de recombinaison efficace est plus faible que chez les espèces allogames, car la recombinaison dans des portions de génome homozygotes n'est pas détectable (Nordborg 2000) ; cela conduit à un maintien des haplotypes sur de longues distances et donc au maintien du DL ;
- la dérive génétique : suite à un événement démographique tel qu'un goulot d'étranglement ou dans des populations de taille efficace limitée, des changements aléatoires de fréquences alléliques vont avoir lieu (Lande 1976), responsables de la disparition de certains allèles ce qui permet le maintien du DL entre les haplotypes qui se trouvent en nombre limité dans cette population.

Au contraire de ces facteurs, d'autres participent à la réduction de l'étendue du DL tels que :

- la recombinaison : lorsqu'un événement de recombinaison a lieu dans une région génomique, le DL qui liait les allèles reste le même de part et d'autre de la recombinaison mais se trouve réduit à néant entre les allèles des deux fragments (Flint-Garcia, Thornsberry, and Buckler 2003); plus le nombre de générations de recombinaison est important depuis un ancêtre commun, moins le DL sera étendu ; les allèles des loci très proches physiquement peuvent ainsi rester en DL pendant de nombreuses générations malgré la recombinaison;
- les mutations (à un taux élevé): lorsqu'elles surviennent de façon récurrente et convergente à l'intérieur de blocs haplotypiques en DL, les mutations vont « casser » l'association préférentielle entre les allèles des loci de ce bloc.

1.2.2.3. Décroissance du déséquilibre de liaison en fonction de la distance physique et conséquences en génétique d'association

Lors de la réalisation d'une étude d'association sans *a priori* (également appelée Genome-Wide Association Study ou GWAS), c'est-à-dire sans cibler de gène candidat et en répartissant des marqueurs

	Population 1			Population 2			Populations 1 + 2		
		A	a		A	a		A	a
	B	81	9	B	2	8	B	83	17
	b	9	1	b	1	89	b	10	90
p_A	0,90			0,03			0,465		
p_a	0,10			0,97			0,535		
p_B	0,90			0,10			0,50		
p_b	0,10			0,90			0,50		
D	0			0			0,183		

Figure 5 : Exemple de création de déséquilibre de liaison lors de la fusion de deux populations en équilibre de liaison, d'après Ytournel (2008)

le long du génome, il est essentiel de déterminer l'étendue de la décroissance du DL dans le génome au sein de la population étudiée, c'est-à-dire la distance physique sur laquelle le DL existe à un taux significatif, et le taux de décroissance du DL en fonction de cette distance physique. Cette mesure va en effet permettre de déterminer le nombre optimal de marqueurs nécessaires pour couvrir la totalité du génome et espérer placer au moins un marqueur par fenêtre de distance (Nordborg and Tavare 2002). Elle va également renseigner sur la précision avec laquelle les facteurs génétiques responsables de la variation phénotypique vont être localisés, c'est-à-dire à quelle distance physique d'un marqueur associé au phénotype on doit chercher le gène responsable. Pour visualiser cette décroissance du DL, un des estimateurs comme le r^2 est calculé entre toutes les paires de marqueurs puis les valeurs sont représentées en fonction de la distance, généralement physique, séparant les marqueurs. La courbe tracée permet ensuite d'identifier la distance à partir de laquelle le r^2 descend en dessous d'une valeur choisie pour l'étude, généralement 0,2, correspondant à la valeur en dessous de laquelle on considère que deux loci ne sont plus significativement en DL. Cette distance, spécifique à chaque population d'étude (Caldwell et al. 2006; Hyten et al. 2007), dépend de nombreux paramètres. Comme indiqué précédemment, le régime de reproduction des individus de la population influe, par exemple, sur l'étendue du DL. Chez des espèces allogames telles que le maïs, le ray-grass ou la vigne, le r^2 décroît en dessous de 0,2 pour des valeurs de 100 pb, 750 pb et 2 cM respectivement (Ponting et al. 2007; Yan et al. 2009; Barnaud et al. 2010). Les espèces autogames telles que la tomate ou l'orge voient quant à elles le DL s'étendre sur des distances égales à plusieurs dizaines de cM (Kraakman et al. 2004; van Berloo et al. 2008). En conséquence, une étude réalisée sur des individus d'une espèce allogame permettra une localisation plus précise d'un facteur génétique qu'une étude réalisée sur les individus d'une espèce autogame, en considérant des tailles de populations identiques et le même nombre de générations depuis l'ancêtre commun le plus récent dans les deux populations. De nombreux biais peuvent également avoir une influence sur l'étendue du DL parmi lesquels on trouve les biais relatifs aux fréquences alléliques. Le r^2 est en effet calculé à partir de ces fréquences aux différents loci et l'utilisation de certaines techniques de génotypage, notamment les puces, peuvent biaiser cette mesure. Cela s'explique par le fait que les SNPs choisis pour ces puces doivent répondre à des critères de sélection stricts qui conduisent la plupart du temps à l'élimination des polymorphismes de trop basse fréquence. On observe ainsi une surestimation de l'étendue du DL calculée à partir de données de puces en comparaison avec celle calculée sur des données de re-séquençage (Lachance and Tishkoff 2013). Les données de re-séquençage peuvent par ailleurs, en fonction de leur qualité et du nombre d'erreurs de génotypage, conduire à une sous-estimation de l'étendue du DL du fait de la présence de faux SNPs. Une surestimation de la distance au-delà de laquelle le DL devient non significatif peut conduire à une localisation moins précise des facteurs d'intérêt en augmentant la taille de la fenêtre étudiée. A l'inverse, une sous-estimation peut conduire à l'exploration d'une fenêtre de distance ne contenant pas le locus causal. D'autres facteurs, tels que la structuration de la population (à savoir un mélange récent

de populations ancestrales ou une structure plus ancienne), l'hétérogénéité de l'apparentement entre les individus de cette population (« structuration cryptique ») ou une épistasie entre gènes, peuvent également générer du DL entre plusieurs loci sans que ceux-ci soient pour autant liés physiquement. Il est alors possible, notamment pour la structuration, de corriger les valeurs du DL afin de minimiser les biais dus à l'échantillonnage de la population étudiée (Mangin et al. 2012; Bouchet et al. 2013). Une limite existe cependant à cette mesure du DL moyen sur le génome entier, à savoir que le DL n'est pas homogène sur l'ensemble du génome. La valeur du r^2 est en effet plus grande dans les régions dans lesquelles peu d'événements de recombinaison ont lieu, telles que les centromères, et cette valeur décroît dans les régions avec de forts taux de recombinaison, telles que les télomères (Drouaud et al. 2006; Comadran et al. 2011).

1.2.3. Les études d'association sur génome entier

Pour mener à bien une étude d'association, les éléments sont donc : une population choisie pour représenter un maximum de diversité génétique et phénotypique telle que les core collections, un nombre de marqueurs génotypés suffisamment grand (soit localement dans le cas d'une approche gène candidat, soit globalement dans le cas d'une approche « genome-wide ») et à évaluer en fonction de l'étendue du DL, et des données phénotypiques pour un ou plusieurs caractères d'intérêt.

1.2.3.1. Modèles statistiques utilisés pour tester l'effet d'un SNP en contrôlant les biais possibles

Un modèle statistique plus ou moins complexe est ensuite appliqué aux données afin de tester l'association entre les marqueurs génotypiques et les valeurs phénotypiques. Le modèle le plus simple est linéaire, suivant l'équation suivante :

$$Y = \mu + bX_{SNP} + \varepsilon$$

dans laquelle Y correspond au vecteur des phénotypes, μ à la moyenne générale de la population, b à la pente de régression représentant l'effet du SNP, X_{SNP} au dosage d'un des allèles au SNP et ε à une résiduelle. Un modèle tel que celui-ci ne prend pas en compte l'impact possible de la structure et de l'apparentement pouvant causer des associations entre le phénotype et des SNPs non liés physiquement aux SNPs causaux (faux positifs).

Des modèles statistiques plus élaborés ont donc été proposés afin d'introduire ces effets dans l'équation. Deux méthodes principales sont utilisées dans le but de déterminer la structuration d'une population. La première consiste en une méthode probabiliste d'assignation des individus en populations ancestrales en minimisant les écarts à l'équilibre de Hardy-Weinberg et le DL entre loci (Pritchard, Stephens, and Donnelly 2000; Alexander, Novembre, and Lange 2009), et la seconde en une Analyse en Composantes Principales ou ACP (Price et al. 2010), qui représente les individus dans un espace en quelques dimensions qui maximisent leur dispersion. Le modèle statistique est donc complété de la façon suivante :

$$Y = \mu + bX_{SNP} + Qv + \varepsilon$$

où Y correspond au vecteur des phénotypes, μ à la moyenne générale de la population, b à l'effet du SNP, X_{SNP} au dosage d'un des allèles au SNP, Q à la matrice d'assignation de chaque individu à chaque sous-population (ou aux valeurs des positions des individus sur les premiers axes de l'ACP), v au vecteur des effets des sous-populations et ε au vecteur des résidus individuels.

Lorsque l'on souhaite prendre en compte, en plus de la structure, l'apparentement pouvant lier les individus entre eux, il est nécessaire de passer à un modèle mixte dans lequel les effets du SNP et de la structure sont des effets fixes, et l'effet de l'apparentement un effet aléatoire. L'apparentement correspond ici au fond génétique de chaque individu non expliqué par le SNP. Le modèle est donc écrit de la manière suivante :

$$Y = \mu + bX_{SNP} + Qv + Zu + \varepsilon$$

où Z correspond à la matrice d'occurrence des individus et u au vecteur des effets du fond génétique de chaque individu. De ce modèle découlent les deux équations suivantes :

$$Var(Y) = Var(u) + V_e \qquad Var(u) = 2K\,V_g$$

où V_g est la variance génétique, V_e la variance environnementale et K la matrice d'apparentement entre individus. L'héritabilité (au sens strict = « chip heritability » ; Speed et al. 2012) du caractère étudié peut alors se calculer comme suit :

$$h^2 = \frac{V_g}{(V_g + V_e)}$$

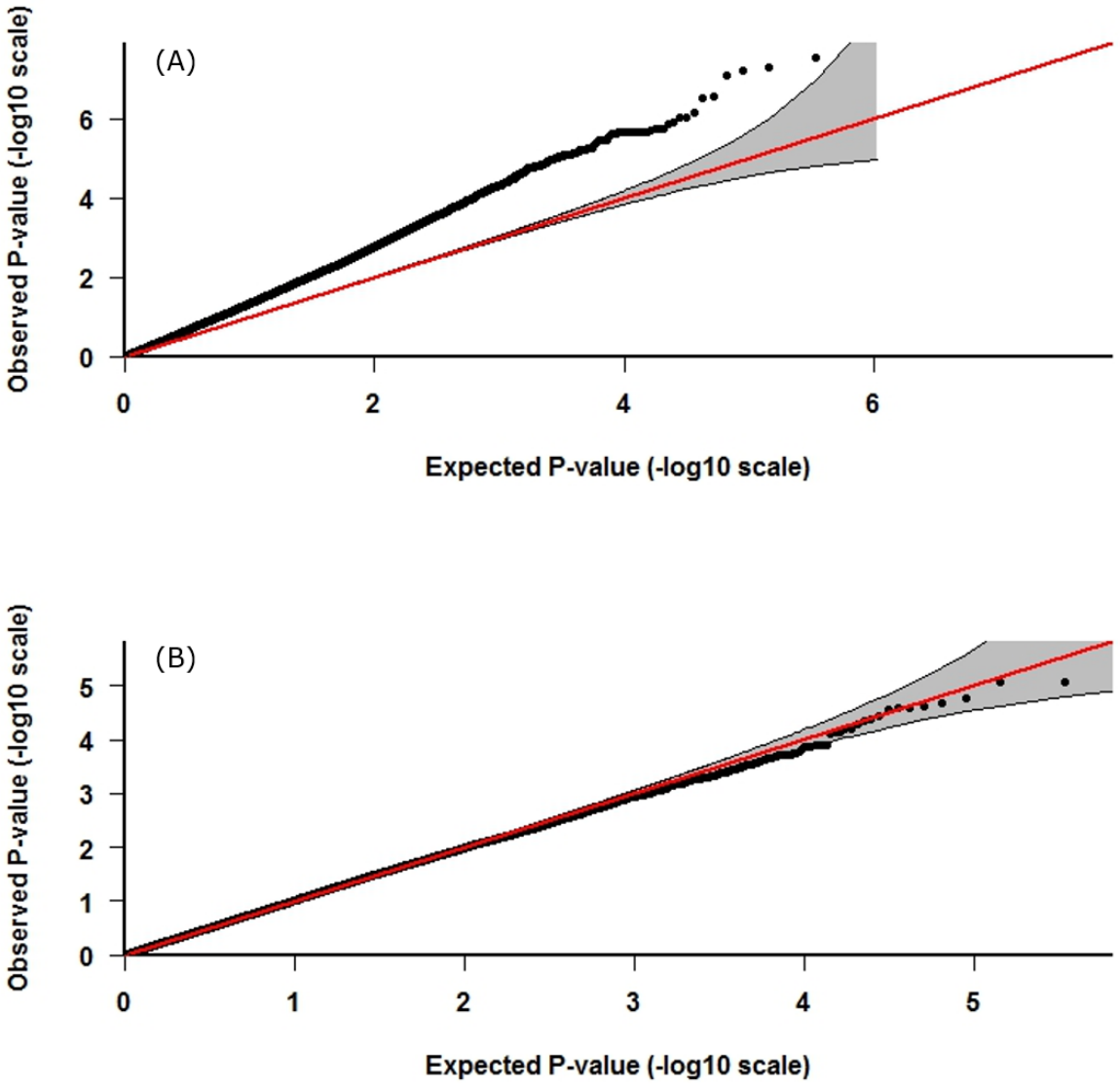


Figure 6 : Exemple de quantile-quantile plots (QQ plot) avant (A) et après correction par l'apparentement et la structure (B)

en abscisse, -log(i / (L + 1)), où L est le nombre de marqueurs, en ordonnée, - log(i^{ème} plus petite p-value) ; la ligne y = x (en rouge) correspond à l'hypothèse nulle ; la zone grisée correspond à un intervalle de confiance, généralement à 95%, sous l'hypothèse nulle ; les points qui s'écartent de la ligne y = x correspondent aux associations significatives

1.2.3.2. Significativité des tests statistiques réalisés

Sous l'hypothèse d'absence d'effet des SNPs (hypothèse nulle), la distribution des valeurs *p* des tests de significativité des effets des SNP (« *p*-value ») est attendue uniforme. Pour vérifier que les résultats ne dévient pas massivement de l'hypothèse nulle, et qu'en particulier les effets de structure et d'apparentement ont bien été pris en compte dans le modèle, les valeurs *p* observées sont projetées contre les valeurs *p* attendues sur un quantile-quantile (QQ) plot (McCarthy et al. 2008) illustré en Figure 6. Cette représentation graphique présente l'avantage de permettre simultanément la visualisation de l'efficacité de la correction et de la présence ou non d'associations significatives.

Une fois le test d'association réalisé, un seuil de significativité doit être choisi afin de contrôler au mieux le nombre de faux positifs détectés (identification d'un SNP pour lequel on déclare une association phénotype-génotype significative alors qu'elle ne l'est pas en réalité). En effet, la réalisation d'un très grand nombre de tests statistiques à un seuil défini de x% entraîne fatalement l'apparition de plusieurs tests significatifs par le seul hasard, si ce seuil ne tient pas compte de la multiplicité des tests. Une première approche consiste à calculer le seuil individuel de significativité en utilisant la correction de Bonferroni selon :

$$S = {}^{\propto}/_{N}$$

où α est le niveau de significativité global souhaité (par exemple 5%) et N le nombre de tests effectués (souvent égal au nombre de SNPs testés). Ce seuil est jugé très conservatif d'une manière générale, et en particulier dans les GWAS, car il suppose que les tests (donc ici les marqueurs) sont indépendants alors qu'un certain nombre de marqueurs sont en DL. Une modification de la correction de Bonferroni a donc été proposée en cherchant à évaluer le nombre de SNPs indépendants dans le jeu de données. Une telle approche est par exemple développée par le logiciel GEC (Li et al. 2012), ce qui permet ensuite de ré-estimer le seuil de significativité selon le nombre de SNPs indépendants. Une deuxième approche consiste à effectuer des tests de permutation mais cela peut se révéler chronophage dès que l'on dispose de grands jeux de données (Gao et al. 2010). Une dernière méthode consiste à calculer le False Discovery Rate (FDR) qui évalue la proportion de faux positifs parmi l'ensemble des tests déclarés positifs (Benjamini and Hochberg 1995; Dolejsi, Bodenstorfer, and Frommlet 2014). Le seuil de significativité de la *p*-value défini selon le FDR est généralement nettement supérieur à celui de la correction de Bonferroni.

Il est également important de réaliser l'analyse d'association sur un jeu de données ne contenant pas de données manquantes, grâce à l'imputation de celles-ci (Bink et al. 2002; Stephens and Donnelly 2003; Scheet and Stephens 2006), ni de marqueurs dont la fréquence serait trop faible, le seuil étant généralement fixé à 0,05. Ces derniers, s'ils sont associés avec le phénotype, ont en effet peu de chance d'être détectés (Myles et al. 2009).

L'ajustement du modèle aux effets de structure et d'apparentement permet de réduire le nombre de faux positifs mais entraîne une réduction de la puissance de détection des vrais positifs. Cette perte de puissance, potentiellement due au fait que les mêmes marqueurs sont utilisés pour tester les associations et estimer l'apparentement, a été estimée plus importante dans les régions en fort DL chez le maïs (Rincent et al. 2014). Deux méthodes alternatives de calcul de la matrice d'apparentement, permettant de pallier cette baisse de puissance, ont donc été proposées : la première consiste à utiliser tous les marqueurs sauf ceux présents sur le même chromosome que le marqueur testé et la seconde à donner un poids à la contribution de chaque marqueur.

1.2.3.3. Logiciels disponibles

De nombreux logiciels, présentant diverses caractéristiques au niveau des temps de calcul et des algorithmes utilisés, ont été développés pour tester les associations à l'échelle du génome. Parmi ces logiciels on peut citer :

- TASSEL (Bradbury et al. 2007) qui implémente une méthode permettant d'éviter l'estimation systématique des composantes de la variance à chaque SNP, en utilisant celles pré-estimées à partir du modèle nul ; le logiciel EMMAX (Kang et al. 2010) utilise la même approximation ;
- GEMMA (Zhou and Stephens 2012, 2014) qui effectue des calculs exacts (sans approximation) sans augmenter le temps d'analyse grâce à la décomposition en valeurs propres de la matrice d'apparentement au départ de l'analyse ; ce logiciel permet aussi de tester les associations phénotype-génotype sur plusieurs caractères traités conjointement de manière à tirer profit des corrélations entre caractères pour augmenter la puissance de détection ;
- Fast-LMM (Lippert et al. 2011) qui utilise une matrice d'apparentement calculée sur un sous-ensemble de marqueurs de manière à diminuer le temps de calcul.

D'autres approches visant à tester conjointement l'effet de plusieurs SNPs ont également été développées, de manière analogue à la recherche de QTLs multiples (« Multiple QTL Mapping ») sur descendances. C'est le cas de l'approche MLMM (« Multi-Loci Mixed Models ») développée par Segura et al. (2012). Des propositions plus élaborées ont été faites pour remplacer la matrice d'apparentement individuel (K) par une matrice d'apparentement entre groupes d'individus, rassemblés à la suite d'une analyse de clustering selon l'approche UPGMA (« Unweighted Pair Group Method with Arithmetic mean »). Une version améliorée de cette démarche a été implémentée dans GAPIT (Lipka et al. 2012). Un nouveau logiciel apparu très récemment semble à nouveau montrer un gain en matière de rapidité de calcul et de puissance de détection d'associations significatives : le logiciel BOLT-LMM (Loh et al.

2015) modélise l'architecture génétique des caractères, non pas selon le modèle infinitésimal avec distribution normale des effets génétiques (comme implicitement considéré dans les logiciels précédents), mais selon un mélange *a priori* de distributions d'effets *via* une approche Bayésienne. Enfin, l'utilisation de plusieurs SNPs en DL formant des blocs haplotypiques a également été proposée afin de mieux tenir compte de la variabilité allélique au niveau du locus causal de la variation phénotypique (Lorenz, Hamblin, and Jannink 2010; Barendse 2011).

1.2.3.4. Mise en œuvre de la génétique d'association

La génétique d'association a permis de grandes avancées en recherche, notamment en santé humaine, avec l'identification de très nombreux polymorphismes causaux de pathologies (Imamura and Maeda 2011; Visscher et al. 2012; Fachal and Dunning 2015), mais également en biologies animale et végétale avec l'identification de nombreux loci présentant une association avec des facteurs d'intérêt agronomique (Bolormaa et al. 2011; Wang et al. 2012; Kang et al. 2015). Cependant, l'approche de génétique d'association ne conduit pas souvent à l'identification du locus causal mais plutôt des marqueurs à son voisinage (Korte and Farlow 2013), ce qui permet de dresser une liste de gènes candidats responsables de la variation phénotypique observée. Le développement de nouveaux modèles statistiques devrait permettre, dans un futur proche, d'améliorer les résultats des études de génétique d'association et la localisation des facteurs génétiques recherchés (Huang and Han 2014).

Les nombreuses analyses GWAS ont pu être effectuées grâce à la synthèse de nombreuses puces de génotypage comprenant d'abord de faibles densités de marqueurs SNP (Bachlava et al. 2012; Verde et al. 2012), puis de moyennes et hautes densités (Kranis et al. 2013; Bianco et al. 2014; Lee et al. 2014; Unterseer et al. 2014; Wang, Wong, et al. 2014; Bassil et al. 2015). De nombreuses questions restent cependant d'actualité malgré les progrès rapides permis par la génétique d'association, telles que le déterminisme génétique des caractères complexes régulés par de nombreux loci d'un génome, ou la détection de polymorphismes causaux présentant des fréquences alléliques trop faibles pour être détectés par une telle approche (Myles et al. 2009; Korte and Farlow 2013). Sous-jacente à cette absence de signal trouvée dans de nombreuses études, se trouve l'héritabilité manquante ou « missing heritability » (Maher 2008), définie par le fait que la totalité de la variation génétique ne peut être expliquée par les données génotypiques, qui est en phase d'être élucidée. Deux approches ont été développées, à savoir une première similaire à de la prédiction génomique (Goddard and Hayes 2007), qui utilise un modèle où l'ensemble des SNPs sont impliqués sous forme d'effets aléatoires tel qu'implémenté dans le logiciel GCTA (Yang et al. 2011), et une seconde visant à réaliser des Epigenome-Wide Association Studies (EWAS), qui prennent en compte les modifications épigénétiques

de l'ADN des individus étudiés, et dont les résultats peuvent être mis en commun avec les résultats de GWAS (Verma 2012).

1.3. La génétique des populations au service de la cartographie

1.3.1. Principes de la génétique des populations

L'incorporation des principes de Mendel à la théorie de l'évolution a donné naissance au début du XXème siècle à une nouvelle discipline, la génétique des populations, qui peut également être utilisée à des fins de cartographie du génome. Le but de cette discipline est la compréhension des différents processus évolutifs que sont la migration, la sélection, la dérive génétique et la mutation (Nielsen 2005). La génétique des populations peut permettre de repérer des gènes ou régions du génome impliqués dans les caractères liés à l'adaptation, notamment à travers la modification de leur polymorphisme. Des gènes ou régions génomiques impliqués dans la domestication ou dans la sélection variétale (empirique ou orientée) peuvent ainsi être identifiés au niveau du génome. Un des principes importants de génétique des populations est la théorie neutraliste de l'évolution qui stipule que, au niveau moléculaire, la plupart des changements et des variations intra- et inter-espèces ne sont pas dus à la sélection naturelle mais à la dérive génétique (Kimura 1968). Quand on veut inférer l'existence de sélection, on doit donc tester si les données génétiques de type séquences (AFLP, RFLP, SSR) ou de type génome entier, SNPs (Helyar et al. 2011), ne peuvent pas être expliquées par la théorie neutraliste, sous différents scenarii démographiques. Lorsque l'hypothèse neutraliste est rejetée, on considère que le locus étudié a été soumis à sélection. Différents estimateurs sont ainsi utilisés à différentes fins pour déterminer (i) s'il y a eu sélection au locus étudié, (ii) le type de sélection qui a eu lieu, et (iii) si des événements démographiques susceptibles de changer les spectres de fréquences ont eu lieu. Il existe plusieurs types de sélection parmi lesquels se trouvent :

- la sélection directionnelle positive : favorise l'allèle bénéfique jusqu'à fixation dans la population ;
- la sélection directionnelle négative : désavantage l'allèle défavorable jusqu'à fixation de l'autre allèle dans la population ;
- la sélection équilibrante : favorise le maintien de plusieurs allèles au même locus ; se produit lorsque la valeur des hétérozygotes est supérieure à celle des homozygotes, lorsque l'allèle rare a un avantage, ou lorsqu'il existe une sélection hétérogène dans le temps ou dans l'espace.

Lors des événements de sélection positive, on observe aux alentours du locus sélectionné une empreinte particulière au niveau moléculaire, appelée balayage sélectif, car toute la diversité génétique est éliminée. En effet, quand une nouvelle mutation est fortement sélectionnée, elle augmente rapidement en fréquence, entraînant avec elle les allèles des sites proches physiquement, phénomène appelé l'autostop génétique (Smith and Haigh 2007). On observe donc autour de ce locus sous sélection une diminution de la diversité nucléotidique ainsi qu'une augmentation du déséquilibre de liaison. L'étendue de la zone autour du locus sous sélection dépend de l'intensité de la sélection et diminue lorsque le nombre de générations augmente depuis la fixation, conjointement avec le nombre de recombinaisons (Sabeti et al. 2002). Inversement, la sélection équilibrante maintient des allèles sur de grandes périodes de temps, plus longtemps qu'attendu sous l'hypothèse neutraliste, ce qui laisse le temps pour l'accumulation de nombreuses substitutions entre les allèles. La sélection équilibrante augmente donc le polymorphisme autour du locus sous sélection.

Au-delà des différents types de sélection pouvant avoir lieu dans une population, divers événements démographiques influent également sur le spectre de fréquences alléliques. Parmi ceux-là on peut citer les goulots d'étranglement qui correspondent à une diminution drastique de la taille de la population. Un tel événement entraîne une augmentation des taux de consanguinité, une diminution de la variation génétique et une augmentation du risque de fixer des allèles délétères dans la population (Cornuet and Luikart 1996). A l'opposé des goulots d'étranglement, on trouve l'expansion de population qui correspond à une augmentation rapide de la taille de la population. Au niveau génétique, un tel événement a pour conséquences un excès de variants à basse fréquence et un déficit en variants de fréquence intermédiaire (Excoffier, Foll, and Petit 2009). Les événements de sélection autant que les événements démographiques ont donc un impact sur le spectre de fréquences alléliques. Cependant, la démographie a un effet sur le génome entier tandis que la sélection a un effet sur une partie du génome. Ainsi, il est nécessaire de réaliser des tests pour déterminer si le spectre de fréquences alléliques est explicable par la démographie seule ou si la sélection a également joué un rôle.

1.3.2. Estimateurs utilisés en génétique des populations

Les principaux estimateurs utilisés pour détecter les signatures de sélection au niveau moléculaire sont :

- le D de Tajima (Tajima 1989) : il compare le nombre de sites polymorphes dans une séquence et la diversité moyenne par site ; en l'absence de sélection ces deux quantités sont censées être égales et donner un D de Tajima non significativement différent de 0 (attendu neutre) ; dans le cas d'une sélection équilibrante, on observe un excès d'allèles à fréquences intermédiaires, ce qui résulte en un D de Tajima

négatif ; dans le cas d'une sélection positive, on observe un excès de variants rares, conduisant à un D de Tajima positif ; certains événements démographiques peuvent également engendrer des écarts à la neutralité du D de Tajima tels que les goulots d'étranglement récents (D > 0) ou les expansions de populations (D < 0) ;

- le F de Fu et Li (Fu and Li 1993) : il est similaire au D de Tajima du fait qu'il permette la détection des biais dans le spectre de fréquences alléliques, et à la différence près qu'il fait la distinction entre les mutations anciennes et récentes ; il étend le principe du D de Tajima en incluant une autre espèce permettant de polariser les mutations en allèles ancestraux et dérivés ; lors d'une expansion de la population ou après un autostop génétique, on observe un excès de polymorphismes à basse fréquence qui résulte en une valeur négative de F ; dans le cas d'une sélection équilibrante ou d'un goulot d'étranglement récent, on observe un déficit en polymorphismes de fréquence intermédiaire qui se traduit par une valeur positive de F ; le F de Fu et Li est moins sensible aux phénomènes d'expansion de population et d'autostop génétique et permet donc une meilleure interprétation du D de Tajima ;
- le H de Fay et Wu (Fay and Wu 2000) : il détecte un excès d'allèles dérivés à haute fréquence, témoins d'une sélection positive, et prend alors une valeur négative ; il n'est pas sensible au phénomène d'expansion de population contrairement au D de Tajima ; il requiert des données d'une autre espèce.

Ces estimateurs permettent d'obtenir une idée plus précise de l'histoire évolutive d'un locus, tant au niveau de son histoire sélective que démographique.

1.3.3. Applications de la génétique des populations en cartographie

L'une des applications possible à la génétique des populations en cartographie est l'identification et la localisation des facteurs génétiques ayant subi des pressions de sélection, naturelle ou liée à l'activité humaine (Figure 1C). La plupart des études de génétique des populations partent d'une observation phénotypique et progressent jusqu'au(x) facteur(s) génétique(s) responsable(s) de cette observation (Wright and Gaut 2005). Ces études ciblent en général des gènes candidats, censés avoir participé à l'adaptation des populations à leur environnement ou avoir été sélectionnés par l'Homme, et depuis quelques années, grâce à l'avancée du génotypage haut-débit, certaines sont réalisées sur des génomes entiers sans *a priori*. Parmi ces études, celles réalisées sur des populations humaines ont très vite bénéficié du génotypage haut débit, et ont réussi à détecter des régions génomiques sous sélection ainsi que les gènes sous-jacents, permettant d'expliquer plus précisément l'histoire évolutive

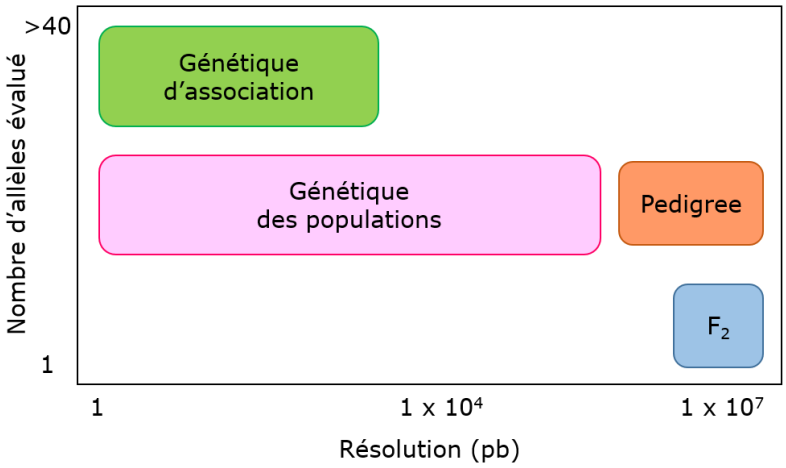


Figure 7 : Comparaison des trois approches pour la cartographie de gènes d'intérêt, d'après Flint-Garcia et al., (2003)

de l'Homme (Sabeti et al. 2007). Chez les animaux, des études ont été menées sur des espèces d'importance économique majeure tels que le bœuf (Larkin et al. 2012), le porc (Amaral et al. 2011) ou le chien domestique (Axelsson et al. 2013). D'autres espèces utiles à l'Homme comme les champignons utilisés dans l'agro-alimentaire (Cheeseman et al. 2014) ont également été étudiées par l'approche de génétique des populations en cartographie. Chez les plantes, les différentes études ont été réalisées sur des espèces d'intérêt agronomique comme le maïs (Zhang et al. 2002), mais également sur des espèces pour lesquelles la cartographie « classique » sur descendances était difficile à mettre en place. Parmi ces études on peut citer celles ayant permis la confirmation de gènes impliqués dans la vernalisation et la photopériode chez l'épicéa commun *Picea abies* (Heuertz et al. 2006), ou l'identification de gènes impliqués dans divers processus physiologiques chez le pin à torches *Pinus taeda* L. (Eckert et al. 2010).

1.4. Comparaison des différentes méthodes

Chacune des méthodes de localisation de gènes ou régions génomiques contrôlant la variation de caractères agronomiques ou adaptatifs présentées ci-dessus possède des avantages et des inconvénients, listés ci-dessous et résumés dans la Figure 7.

1.4.1. Choix des caractères étudiés

Les cartographies par liaison génétique ou génétique d'association sont appliquées à des caractères mesurés (sur les populations étudiées) et donc choisis de manière ciblée. Chaque caractère peut ainsi être traité séparément pour la détection de QTLs. Au contraire, les analyses de génétique des populations peuvent permettre d'identifier des régions génomiques sous sélection sans que l'on connaisse d'avance précisément le(s) caractère(s) adaptatif(s) qui a (ont) été le(s) plus influent(s). Dans ce second cas, il faut alors prendre en compte d'autres caractéristiques ou informations sur la population étudiée (par exemple le gradient écologique ou pédo-climatique) pour émettre des hypothèses sur les caractères potentiellement responsables de la sélection, pour lesquels les traces de sélection indiqueront des régions génomiques porteuses de gènes candidats. Par ailleurs, les traces de sélection peuvent correspondre à l'effet de plusieurs caractères ayant subi simultanément une sélection qui aurait induit des modifications de fréquences alléliques à différents endroits du génome. L'attribution des signatures de sélection identifiées aux différents caractères potentiellement impliqués n'est donc pas simple.

1.4.2. Choix de la population d'étude

La cartographie par liaison génétique implique de pouvoir produire des individus apparentés issus de croisements contrôlés entre des parents présentant des phénotypes contrastés (pour les lignées), ou fortement hétérozygotes pour un même caractère. La cartographie par génétique d'association peut être réalisée plus simplement sur des cohortes contrastées de type case-control, ou sur des collections rassemblant une diversité génétique et phénotypique suffisantes. La constitution de core collections maximisant cette diversité implique une bonne connaissance préalable de la diversité génétique inhérente à l'espèce. Les analyses de génétique des populations sont en général réalisées sur des individus choisis au hasard au sein d'une ou plusieurs populations, la plupart du temps sauvages, la bonne représentativité de l'échantillonnage vis-à-vis du (des) caractère(s)s adaptatif(s) ciblé(s) ou supposé(s) pouvant être un point crucial pour cette dernière approche.

1.4.3. Ampleur de la diversité étudiée

La cartographie par liaison génétique permet d'étudier au maximum quatre allèles par croisement, sur les loci ségrégeant, dans le cas d'individus diploïdes hétérozygotes (descendance F1). Au contraire, la cartographie par génétique d'association, réalisée aussi bien sur des cohortes case-control que sur des collections, permet potentiellement d'identifier un nombre de loci et d'allèles par loci beaucoup plus grand. Cependant, plus le nombre de loci ségrégeant est important, plus la complexité de l'architecture génétique du caractère va être difficile à explorer, en particulier si des relations d'épistasie existent entre certains loci. Par ailleurs, un nombre d'allèles trop important peut rendre difficile la lecture d'un signal à l'aide de SNPs bialléliques, le passage à la détection de signaux par bloc haplotypique pouvant partiellement résoudre cette difficulté. Les analyses de génétique des populations étant réalisées dans le but d'identifier des signatures de sélection, caractérisées entre autres par une diminution locale de la diversité, les gènes identifiés présentent la plupart du temps un faible nombre d'allèles. Comme indiqué précédemment à propos du choix des caractères étudiés, cette dernière approche peut également être restreinte en matière d'exploration de la diversité génétique d'un caractère donné en fonction de son implication réelle dans l'adaptation de la population considérée.

1.4.4. Résolution de la cartographie

D'une manière générale, la précision de la localisation des gènes ou régions génomiques impliqués dans la variation des caractères est directement dépendante du nombre de recombinaisons ayant eu lieu entre les individus et leur plus proche ancêtre commun. En ce qui concerne la cartographie de QTLs par liaison génétique, la résolution est donc limitée puisqu'il n'y a qu'une génération de recombinaison considérée. La cartographie par génétique d'association permet en revanche une résolution bien supérieure surtout si le plus proche ancêtre commun « moyen » est relativement éloigné, et il peut arriver que cette approche aille jusqu'à la localisation de SNPs causaux. Les analyses de génétique des populations, si elles sont menées sur des gènes candidats, permettent d'identifier les haplotypes responsables d'un caractère phénotypique donné. Les analyses sans *a priori* réalisées sur génome entier permettent également d'identifier des gènes sous sélection, à condition que le marquage moléculaire et le DL soient adaptés, mais ces gènes ont de grandes chances d'appartenir à plusieurs voies métaboliques, ce qui peut compliquer l'interprétation.

2. Le modèle d'étude : le pommier cultivé

2.1. *Malus* x *domestica*

2.1.1. Généralités

Le pommier cultivé, *Malus* x *domestica*, est un arbre de la famille des *Rosaceae*, de la sous-famille des *Maloïdeae*, de la tribu des *Pyrae* et du genre *Malus*, qui comprend une trentaine d'espèces. La sous-famille des *Maloïdeae* est caractérisée par des fruits possédant de deux à cinq carpelles. La pomme, qui en possède cinq, est considérée comme un faux-fruit puisqu'elle résulte du développement, non pas de l'ovaire comme les fruits classiques, mais de l'endocarpe, tissu qui entoure l'ovaire de la fleur. La reproduction des pommiers est principalement allogame du fait d'une auto-incompatibilité gamétophytique (Frankel and Galun 1977), l'autofécondation restant possible chez certaines variétés mais avec un taux de réussite très faible. De ce fait, les vergers de production nécessitent d'être plantés, en plus des arbres qui porteront les fruits, de variétés pollinisatrices. La majorité des variétés de pommiers sont diploïdes (2n = 34), bien qu'un nombre important de variétés anciennes soient triploïdes, ce qui reflète une sélection empirique efficace en faveur de cette caractéristique qui est souvent associée à un calibre de fruit plus important (Lassois et al. 2015). Une première version de la

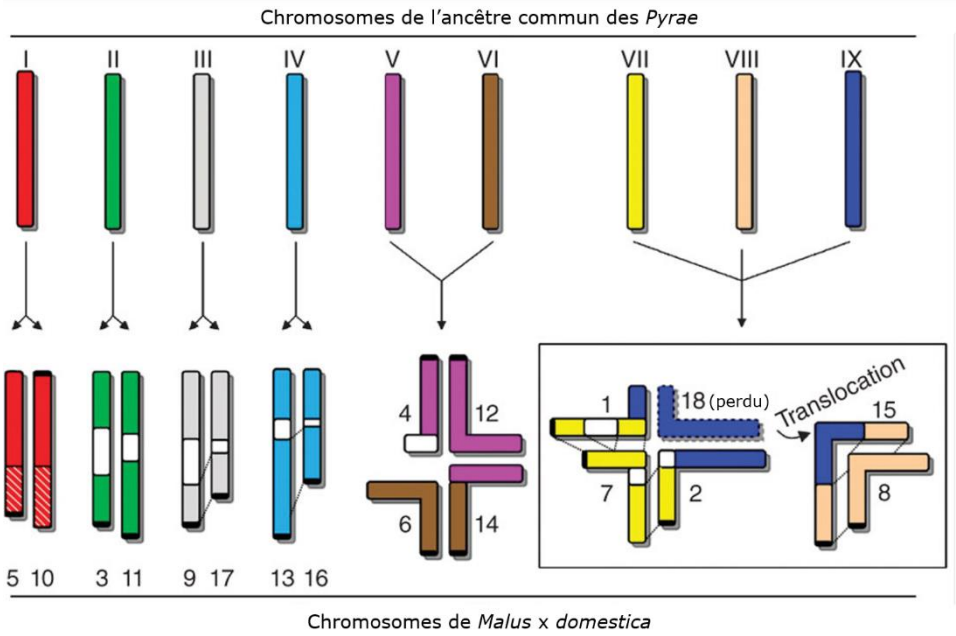


Figure 8 : Réarrangements chromosomiques ayant donné, depuis l'ancêtre commun à 9 chromosomes, le génome actuel des *Pyrae* à 17 chromosomes, d'après Velasco et al. (2010)

séquence génomique de la variété « Golden Delicious » a été publiée, facilitant ainsi les études de type moléculaire (Velasco et al. 2010).

2.1.2. Le pommier dans le monde

La pomme est le fruit le plus répandu dans les régions tempérées, arrivant en seconde place de la production mondiale derrière la banane, avec plus de 76 millions de tonnes produites en 2012 (FAO). Le pommier peut aussi bien pousser dans des régions situées à de hautes latitudes, et où les températures descendent largement en dessous de 0°C, que dans des régions tropicales (Forsline et al. 2003). Cultivées depuis l'Antiquité par les civilisations grecque et romaine, les pommes ont gagné leur popularité grâce à leur longue durée de conservation (jusqu'à un an après la récolte en conditions contrôlées de type « Ultra Low Oxygen »), leur facilité d'utilisation et grâce aux nombreuses façons possibles de les consommer (fraîches, cuites, sous forme de compote, de jus, de cidre…).

2.1.3. Origine du pommier cultivé

Le genre *Malus* serait originaire des provinces du Sud de la Chine où sont retrouvées une grande partie des espèces de *Malus* existant. L'apparition du genre, datée à la période de l'Eocène (55,5 à 33,7 millions d'années), aurait pris place en même temps que celles d'autres espèces de la famille des *Rosaceae* (Juniper and Mabberley 2006). Le génome d'un ancêtre commun au pommier et au poirier possédant neuf chromosomes aurait subi une auto-polyploïdisation et aurait donné, après certains réarrangements chromosomiques (soit la perte d'un chromosome, soit la fusion de deux chromosomes ; Figure 8), le génome à 17 chromosomes caractéristique des *Pyreae* (Velasco et al. 2010). Le pommier cultivé serait issu d'une hybridation interspécifique mettant en jeu les quatre espèces de pommiers sauvages présentes depuis l'Asie jusqu'à l'Europe, *Malus sieversii*, *Malus orientalis*, *Malus sylvestris* et *Malus baccata*. Une des hypothèses concernant l'apparition du pommier cultivé stipule qu'il serait apparu à partir de *M. sieversii* en Asie centrale il y a environ 8 000 ans, qu'il aurait ensuite été transporté le long de la route de la Soie, où des hybridations avec *M. orientalis* et *M. baccata* auraient eu lieu, et qu'il aurait finalement été hybridé avec *M. sylvestris* après son introduction en Europe, il y a 3 000 ans (Forsline et al. 2003). Les implications de *M. sieversii* et *M. sylvestris* ont récemment été confirmées grâce à des données moléculaires (Cornille et al. 2012). Certains cultivars présentant des caractères d'intérêt pour l'Homme et probablement issus d'hybridations aléatoires, ont été maintenus grâce à la propagation végétative mise en place depuis l'Antiquité chez les ligneux par le greffage (Forsline et al. 2003). De ces anciens cultivars sont nées peu à peu les variétés de pommiers telles que nous les

connaissons aujourd'hui, adaptées aux utilisations diverses pour lesquelles l'Homme les a sélectionnées.

2.2. Les deux types variétaux de pommiers cultivés

Parmi les nombreuses variétés de pommiers domestiques répertoriées et destinées à la production de fruits, deux types variétaux se distinguent en fonction de l'utilisation qu'en fait l'Homme. Le premier type, représentant une grande majorité des variétés, est constitué par les pommes dites « à couteau » et le second type est constitué par les pommes dites « à cidre ». Les principales caractéristiques permettant de classer ces deux types variétaux de pommes sont présentées ci-dessous.

2.2.1. Le pommier à couteau

Les pommes à couteau représentent la grande majorité des pommiers cultivés, avec plus de 20 000 variétés réparties dans le monde entier (Juniper and Mabberley 2006). Egalement appelées pommes de table, les pommes à couteau sont destinées à la consommation telles quelles ou cuites, ainsi qu'à la transformation en jus. Ce sont des fruits de grande taille à la chair sucrée et douce, en comparaison avec les fruits des pommiers sauvages, petits et amers. Une très grande variabilité phénotypique existe entre les différentes variétés de pommes à couteau. Ces différences s'observent pour une multitude de caractères tels que la date de floraison, de maturité (variétés précoces récoltées à la fin de l'été ou variété tardives récoltées jusqu'à fin novembre), la taille et la forme des fruits ainsi que de nombreuses caractéristiques organoleptiques telles que la teneur en sucre, l'acidité ou le contenu en minéraux (Mratinić and Fotirić-Akšić 2011).

2.2.2. Le pommier à cidre

Contrairement aux pommes à couteau qui peuvent être utilisées à plusieurs fins, les pommes à cidre n'ont été sélectionnées que dans le but de produire du cidre. Pour cette raison, elles possèdent des caractéristiques bien particulières, notamment une teneur en polyphénols plus importante (Sanoner et al. 1999), un fruit de plus petite taille et une alternance de production en moyenne plus prononcée que celle des pommiers à couteau (Dapena, Minarro, and Blazquez 2005). Il existe quatre catégories de pommes à cidre en fonction de leur degré d'astringence, correspondant à une teneur en polyphénols

Tableau 1 : Tableau récapitulatif de la teneur en tanins et en acidité des quatre types de pommes à cidre, d'après www.cider.org.uk

Catégories	Tanins (%)	Acidité (%)
Pommes douces-amères	> 0,2	< 0,45
Pommes aigres-amères	> 0,2	> 0,45
Pommes aigres	< 0,2	> 0,45
Pommes douces	< 0,2	< 0,45

plus ou moins élevée, et de leur degré d'acidité, en grande partie dû à la teneur en acide malique (Pereira-Lorenzo, Ramos-Cabrer, and Fischer 2009). On retrouve les pommes dites « douces-amères » qui contiennent plus de 0,2% de tanins et moins de 0,45% d'acidité, les « aigres-amères » contenant plus de 0,2% de tanins et plus de 0,45% d'acidité, les « aigres » contenant moins de 0,2% de tanins et plus de 0,45% d'acidité, et les « douces » qui contiennent moins de 0,2% de tanins et moins de 0,45% d'acidité (Tableau 1). De ces caractéristiques va dépendre la qualité du cidre produit, en sachant qu'un cidre résulte souvent du mélange des jus de plusieurs catégories de pommes à cidre, et même du jus de pommes à couteau, ajouté le plus souvent pour augmenter les volumes de production.

2.3. Caractères ciblés dans les programmes de sélection

Un très grand nombre de variétés de pommiers cultivés ont vu le jour depuis l'événement de domestication il y a 4 000 ans de cela (Cornille et al. 2012). Malgré l'impressionnant nombre de variétés disponibles, une poignée seulement domine le marché. Ces variétés ont été sélectionnées notamment pour leur forte productivité, leur capacité de conservation, et leurs excellentes qualités gustatives, allant des pommes sucrées telles que la Pink Lady® aux pommes acidulées comme la Ariane Les Naturianes®, en passant par des variétés plus juteuses telles que la « Elstar ». Le principal inconvénient de ces variétés réside dans le fait qu'elles sont pour la plupart sensibles aux nombreuses maladies qui touchent les pommiers en vergers (Gessler et al. 2006). Selon les conditions environnementales, certains vergers peuvent ainsi nécessiter plus d'une vingtaine de traitements par an pour fournir aux consommateurs des fruits esthétiquement parfaits. Nous décrirons ici les deux principales maladies du pommier, la tavelure et le feu bactérien, ainsi que les principaux caractères de qualité du fruit ciblés dans les programmes de sélection visant à créer de nouvelles variétés qui possèderont à la fois des niveaux de résistance aux maladies supérieurs et des qualités gustatives exceptionnelles.

2.3.1. La tavelure

La tavelure est une maladie fongique affectant les différents membres de la sous-famille des *Maloïdeae* et causée, chez le pommier, par le champignon ascomycète *Venturia inaequalis*. Cette maladie est responsable de l'apparition de lésions brunes sur les différents organes de la plante, principalement sur les feuilles et les fruits, et représente donc un problème majeur du secteur de production des pommes.

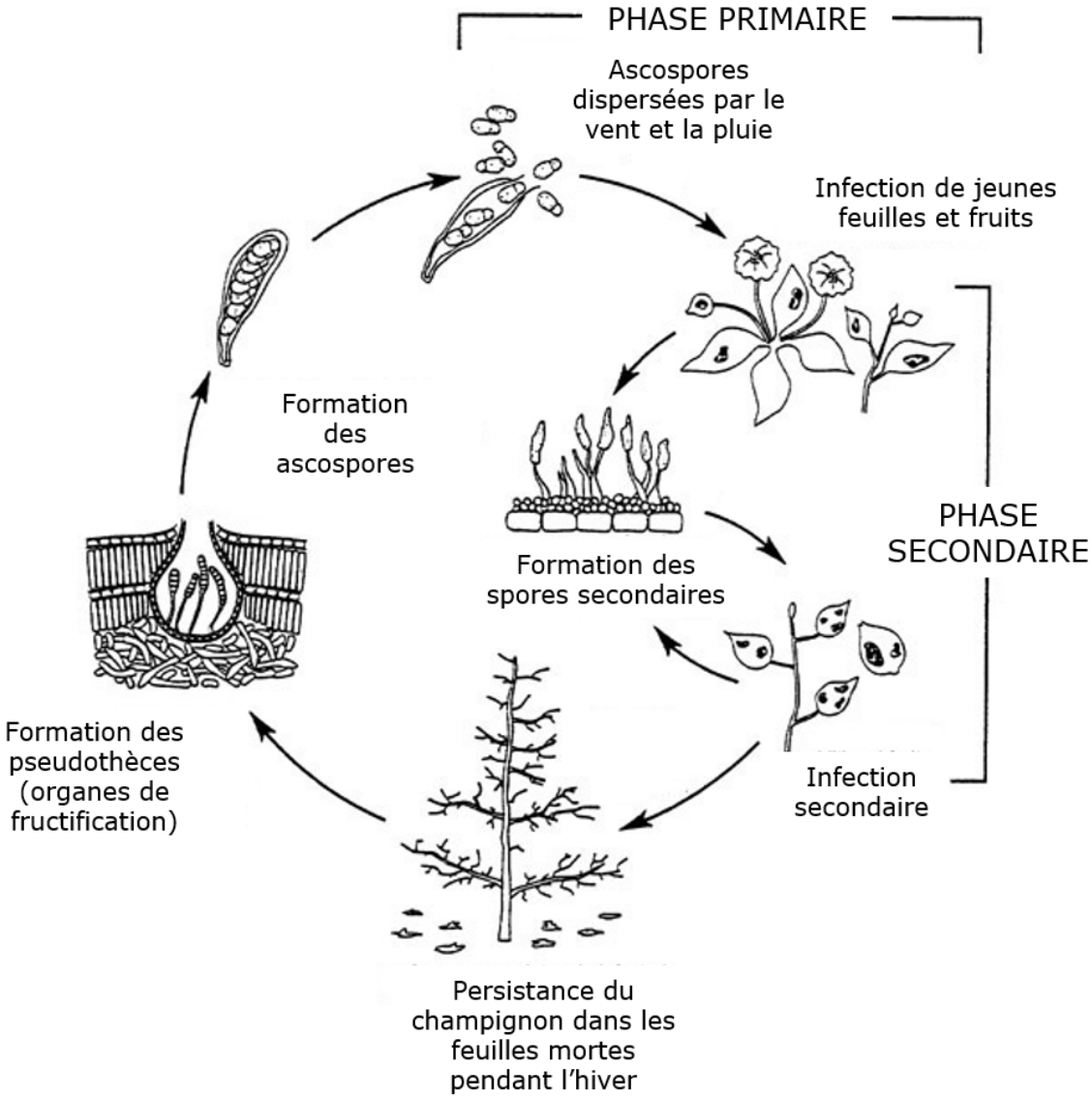


Figure 9 : Cycle de vie du champignon phytopathogène *Venturia inaequalis*, agent de la tavelure du pommier, d'après Vaillancourt et Hartman (2000)

2.3.1.1. Cycle de vie et méthodes de lutte

Le cycle de vie du champignon *Venturia inaequalis* voit sa première étape annuelle se dérouler en hiver alors que le champignon persiste sous forme de pseudothèces, issus de la reproduction sexuée, dans les feuilles tombées au sol (Figure 9). Il subsiste ainsi dans la litière en décomposition pendant la phase saprophyte de son cycle (Bowen et al. 2011). Au printemps, lors des pluies, les ascospores sont libérées des pseudothèces, généralement au moment où les bourgeons éclosent et où les feuilles se développent (MacHardy and Gadoury 1986) : elles constituent l'inoculum primaire. Le risque d'infection est maximal lorsque les organes atteints par les spores, feuilles ou fruits, sont jeunes (Xu and Robinson 2005), et que les conditions climatiques sont humides. Une fois la spore à la surface d'un des organes de la plante, et à condition que la spore soit recouverte d'une gouttelette d'eau libre pendant un temps suffisant, un tube germinatif se forme et perce la cuticule à l'aide d'un appressorium (Smereka, MacHardy, and Kausch 1987) pour former des structures parenchymateuses localisées dans les espaces intercellulaires d'où le champignon tire les nutriments dont il a besoin. Le développement de ces structures et des conidiophores, puis des conidies qu'elles vont produire, conduit à l'apparition des tâches caractéristiques de l'infection par *Ventruria*. Les conidies, obtenues par reproduction asexuée, vont à leur tour être dispersées par le vent et la pluie et constituent l'inoculum secondaire, responsable de l'augmentation de la prévalence de la maladie. Plusieurs cycles de ré-inoculation, infestation, sporulation sont ainsi possibles au printemps, mais plus rarement en été où le temps d'humectation nécessaire à la germination des spores n'est pas atteint, et où les températures trop élevées sont défavorables à la croissance du champignon. Les infestations peuvent reprendre à l'automne sur feuilles plus âgées. Le champignon continue alors de se développer sur les feuilles sénescentes et celles tombées à terre dans le cadre de la phase saprophytique de son cycle. C'est durant cette phase qu'a lieu la reproduction sexuée entre souches compatibles (Keitt and Palmiter 1937) qui donnera naissance aux pseudothèces, qui à leur tour pourront démarrer un nouveau cycle le printemps suivant.

La plupart des variétés de pommiers cultivées sont sensibles à la tavelure (Gessler et al. 2006) ce qui conduit à des dépenses conséquentes pour le contrôle des infections dans les vergers commerciaux (Manktelow et al. 1995). Des traitements fongicides sont par exemple appliqués plusieurs fois par saison de production et ce dans la plupart des régions (Holb 2006), et ont eu pour conséquence l'émergence de souches de *Venturia inaequalis* résistantes à de nombreuses classes de molécules (Köller 1994; Köller et al. 2004). Des stratégies ont été mises en place pour retarder le développement de ces résistances telles que des applications de fongicides raisonnées dans le temps (Brent and Hollomon 1995). Des méthodes de lutte prophylactique ont également été développées et consistent en des pratiques culturales visant à réduire la quantité d'inoculum primaire (élimination des feuilles mortes en hiver ; Holb 2006), et la possibilité de propagation de l'inoculum secondaire (espacement

des arbres dans les vergers, tailles d'aération). La dernière stratégie de lutte repose sur la résistance de certaines variétés à la tavelure à des niveaux plus ou moins élevés.

2.3.1.2. Interaction et résistance

Lors de l'infection d'une variété de pommier sensible par *Venturia inaequalis*, le développement du champignon se fait vraisemblablement grâce à l'action de ses effecteurs qui agissent, tout du moins pour certains d'entre eux, comme des suppresseurs des mécanismes de défense de l'arbre (Bowen et al. 2011). Cette hypothèse est soutenue par la capacité du champignon à pénétrer la cuticule et à se différencier en larges structures pseudo-parenchymateuses (« stroma ») dans l'espace sous-cuticulaire, sans initiation d'une réponse de défense efficace de la plante. Cependant, le développement du champignon va être affecté par plusieurs facteurs, et notamment l'âge de l'organe cible. Toujours chez une variété de pommier sensible, un organe jeune sera plus sensible et verra le développement du champignon former d'importantes lésions à sa surface, tandis qu'un organe plus âgé verra des lésions de taille moindre se développer, puis l'arrêt de la croissance du champignon (MacHardy 1996). Ce phénomène, attribué à la résistance ontogénique, n'est cependant pas suffisant pour protéger efficacement les arbres d'un verger. La résistance ontogénique semble en effet diminuer lors des étapes de sénescence des feuilles et de nouvelles lésions dues au champignon se développent en fin d'été (Kollar 1996). Si le champignon a pu, avant que la résistance ontogénique ne soit complètement mise en place ou après qu'elle a diminué, se développer suffisamment dans les feuilles, l'événement de reproduction sexué prenant place à l'automne ne sera pas compromis (MacHardy, Gadoury, and Gessler 2001). Des sources alternatives de résistance sont donc nécessaires pour qu'un individu présente une résistance accrue à la tavelure.

De nombreux QTLs et gènes de résistance à la tavelure ont été identifiés grâce à des analyses de cartographie de liaison (voir Chapitre 1, 3.1). Les QTLs confèrent généralement une résistance partielle contre une large gamme de souches de pathogènes. Chez la pomme, la plupart des QTLs identifiés semblent cependant être spécifiques de certaines souches de *Venturia* et avoir des efficacités variables (Bowen et al. 2011). Trois classes de gènes de résistance ont été identifiées chez le pommier, regroupés en fonction des symptômes observés sur feuilles :

- la croissance du champignon est arrêtée rapidement après la pénétration dans la plante ; les symptômes sont de type « pin-point » correspondant à des petites dépressions visibles au niveau de l'épiderme et similaires à des piqûres d'épingles ; on parle de réaction d'hyper-sensibilité (HR) ;
- le champignon se développe de façon limitée au niveau sous-cuticulaire et induit des symptômes de type nécrose en étoile ;

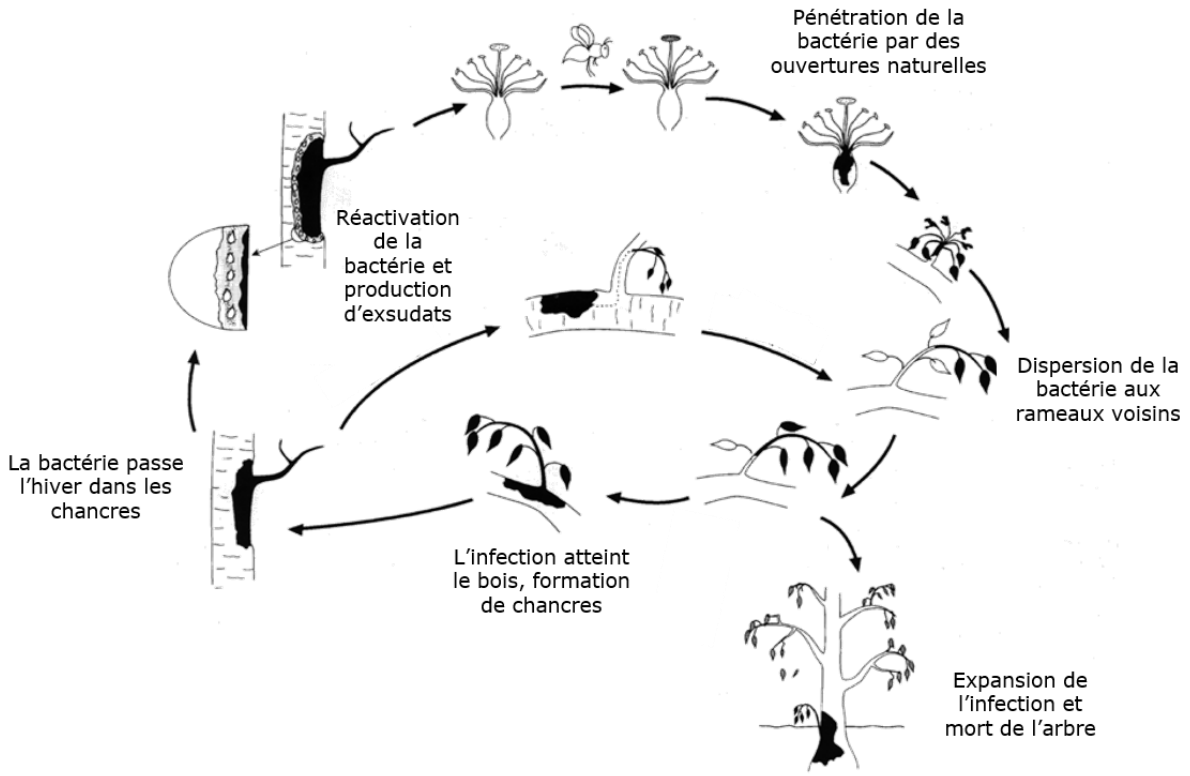


Figure 10 : Cycle de vie de la bactérie phytopathagène *Erwinia amylovora*, agent du feu bactérien chez les *Maloideae*, d'après Wilcox (1994)

- le champignon se développe en causant des symptômes de chlorose et produit une sporulation de faible intensité.

Au niveau moléculaire, le modèle le plus probable est celui où certains effecteurs du champignon agissent comme des protéines d'avirulence reconnues par les produits des gènes de résistance du pommier avec lesquels ils entrent en contact. Lors d'une interaction entre une protéine de résistance et la protéine d'avirulence correspondante, la plante déclenche divers mécanismes de défense (Hammond-Kosack and Jones 1996) : production d'espèces réactives de l'oxygène (ROS), épaississement de la paroi cellulaire, activation de voies de biosynthèse de métabolites secondaires… Les différences d'expression phénotypiques de la résistance présentées précédemment reflètent probablement des différences dans la cascade de signalisation et dans la mise en place des défenses basales (Bowen et al. 2011).

2.3.2. Le feu bactérien

Le feu bactérien est une maladie qui atteint un grand nombre d'espèces de la famille des *Rosaceae*, et en particulier de la sous-famille des *Maloïdeae* (pommier, poirier, cognassier, aubépine…). L'agent pathogène de cette maladie, la bactérie nécrogène *Erwinia amylovora*, est une bactérie Gram-négative de la famille des *Enterobacteriaceae*. Les symptômes de cette maladie sont des nécroses qui touchent aussi bien les branches que les fruits de l'arbre, et qui sont responsables de lourdes pertes économiques (Thomson 2000).

2.3.2.1. Cycle de vie et méthodes de lutte

Le cycle d'*Erwinia amylovora* commence par la pénétration de la bactérie dans la plante (Figure 10). Cette étape se déroule au printemps, en particulier par temps chaud et humide, lorsque la bactérie pénètre dans les tissus par les nectaires des fleurs ou par des blessures sur les feuilles ou les tiges en croissance. L'âge de la fleur, de même que l'âge des organes pour l'interaction du pommier avec *Venturia inaequalis*, influe sur l'efficacité de l'infection. Les fleurs les plus jeunes sont ainsi plus sensibles que les fleurs âgées (Malnoy et al. 2012). La bactérie passerait l'hiver dans des chancres de l'année précédente et contaminerait les fleurs au printemps grâce aux pluies et aux insectes qui dispersent les cellules bactériennes (Miller 1929). La bactérie se multiplie ensuite dans les espaces intercellulaires en formant des nécroses, et contamine peu à peu, d'abord l'inflorescence, puis la branche entière en se déplaçant dans les vaisseaux du xylème (Thomson 2000). Lorsque la concentration en bactéries devient importante, des gouttelettes d'exsudat, composé de bactéries et

d'exopolysaccharides et attractif pour les insectes (Malnoy et al. 2012), se forment à la surface des tissus infectés et constituent une source d'inoculum secondaire.

Les méthodes de lutte contre le feu bactérien sont principalement prophylactiques et consistent en la destruction des parties de la plante infectées, voire à la destruction du verger si un trop grand nombre d'individus sont atteints. L'élimination des chancres à l'intérieur desquels la bactérie serait susceptible de passer l'hiver constitue un élément essentiel dans le contrôle de la dispersion d'*Erwinia amylovora* (Brooks 1926). Les possibilités de lutte chimique sont limitées, la molécule la plus efficace étant la streptomycine, un antibiotique dont l'usage est cependant interdit ou réglementé de manière très stricte selon les pays. L'utilisation de stimulateurs de défense des plantes pourrait offrir une solution alternative puisque certains éliciteurs, tels que l'acibenzolar-S-méthyl, analogue de l'acide salicylique, ou bien le prohexadione-calcium, un réducteur de croissance, montrent expérimentalement des potentialités de protection, partielle mais significative, contre le feu bactérien (Brisset et al. 2000). L'utilisation de bactéries antagonistes a également été proposée (Paternoster et al. 2011). La création de variétés résistantes par sélection ou éventuellement par transgénèse représente donc une option de lutte non négligeable.

2.3.2.2. Interaction et résistance

De la même manière que pour l'interaction pommier-tavelure, l'infection de pommiers sensibles par la bactérie *Erwinia amylovora* est médiée par l'intervention d'effecteurs qui ont été caractérisés, en particulier à travers l'étude de souches mutées. Le pouvoir pathogène de la bactérie (revu par Oh and Beer 2005) repose essentiellement sur sa capacité à injecter ses effecteurs dans les cellules végétales ou à les sécréter dans l'apoplaste *via* un système de sécrétion de type III (Mudgett 2005). Des génotypes résistants de pommiers, chez lesquels la mort cellulaire se limite au point d'infection, ont été identifiés (Durel, Denance, and Brisset 2009; Vogt et al. 2013). Les mécanismes sous-jacents à la résistance de la variété « Evereste » ont été explorés par des approches métaboliques et transcriptionnelles (Dugé De Bernonville et al. 2012; Gaucher 2012; Gaucher et al. 2013). Au niveau métabolique, les recherches se sont focalisées sur les dihydrochalcones, classe de flavonoïdes majoritaire dans les feuilles, et en particulier sur les mécanismes de transformation de ces composés pendant l'infection (oxydation et déglucosylation). Une capacité de transformation différentielle entre les génotypes « Evereste » et « MM106 » (sensible) suggère un rôle possible de ces mécanismes dans le devenir de l'interaction (résistance *versus* sensibilité). Au niveau transcriptionnel, l'analyse des principales voies de signalisation suggère l'implication de la voie de l'acide jasmonique dans la résistance.

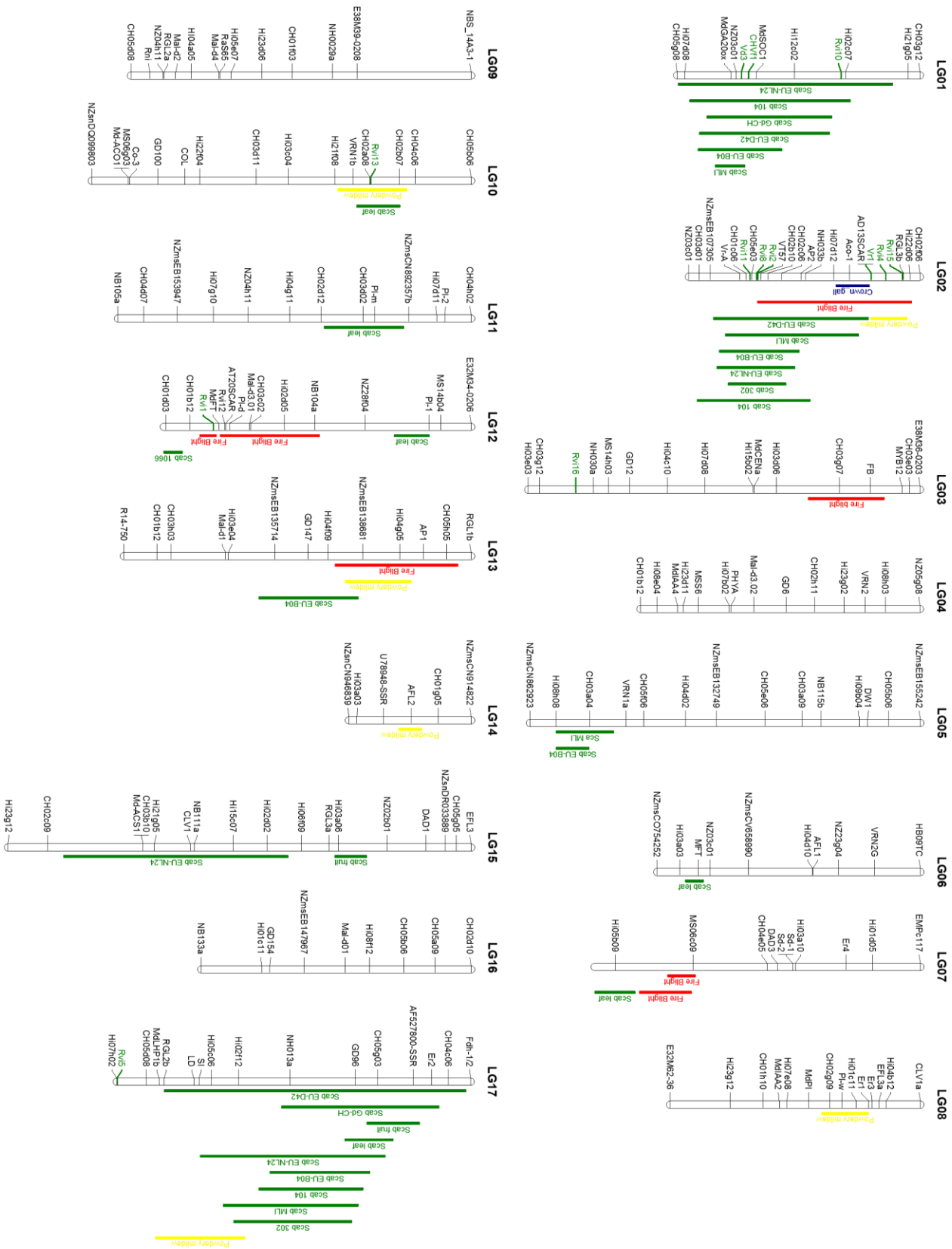


Figure 11 : Carte synthétique de QTLs de résistance aux maladies identifiés chez le pommier par une approche de cartographie en descendance F1

tavelure (vert), feu bactérien (rouge), mildiou (jaune) et galle du collet (bleu)

2.3.3. Caractères de qualité du fruit

Des études sensorielles ont montré que, du point de vue du consommateur, la texture des fruits est l'un des caractères les plus importants (Jaeger et al. 1998). Chez les pommes, les différentes composantes de la texture du fruit sont le croquant, la fermeté, le fondant et la jutosité pour les caractères recherchés par le consommateur, et la farinosité pour les caractères négatifs (Oraguzie et al. 2009). Tous ces caractères résultent de propriétés physiques et biochimiques de la paroi cellulaire, influencés par des facteurs génétiques et environnementaux (Oraguzie et al. 2004). Les caractères ayant attrait au goût sont également importants pour le consommateur et notamment la perception de l'acidité, du sucre et des arômes.

3. Etat de l'art des études de cartographie chez le pommier

L'importance économique du pommier en fait une espèce qui a été et reste la cible de nombreux programmes de sélection. Ci-dessous sont décrits les principaux QTLs et gènes majeurs identifiés lors d'études de cartographie.

3.1. Etudes sur la résistance du pommier aux maladies

Comme évoqué précédemment, les vergers de pommiers de production peuvent subir annuellement plus d'une vingtaine de traitements fongicides et insecticides du fait de la grande sensibilité des cultivars modernes face à un grand nombre de maladies. Nous décrirons ci-dessous les principales études de cartographie ayant permis l'identification et la localisation des facteurs génétiques contrôlant la résistance aux principales maladies affectant le pommier cultivé. Les principaux pathogènes ou ravageurs s'attaquant au pommier sont les champignons *Venturia inaequalis* et *Podosphaera leucotricha*, agents de la tavelure et de l'oïdium respectivement, la bactérie *Erwinia amylovora,* agent du feu bactérien, et le puceron cendré *Dysaphis plantaginea* (Lespinasse, Rousselle-Bourgeois, and Rousselle 1992).

Un certain nombre d'études ont été réalisées sur des descendances F1 de pommiers et ont permis l'identification de QTLs de résistance à ces maladies (Figure 11). Parmi les QTLs identifiés pour la résistance à la tavelure, certains se sont révélés être spécifiques à une souche de *Venturia* (exemple des QTLs localisés sur les groupes de liaison (LG) 13 et LG15 par Calenge et al. 2004), alors que d'autres présentent une résistance vis-à-vis plusieurs souches comme ceux localisés sur les LG01, LG02

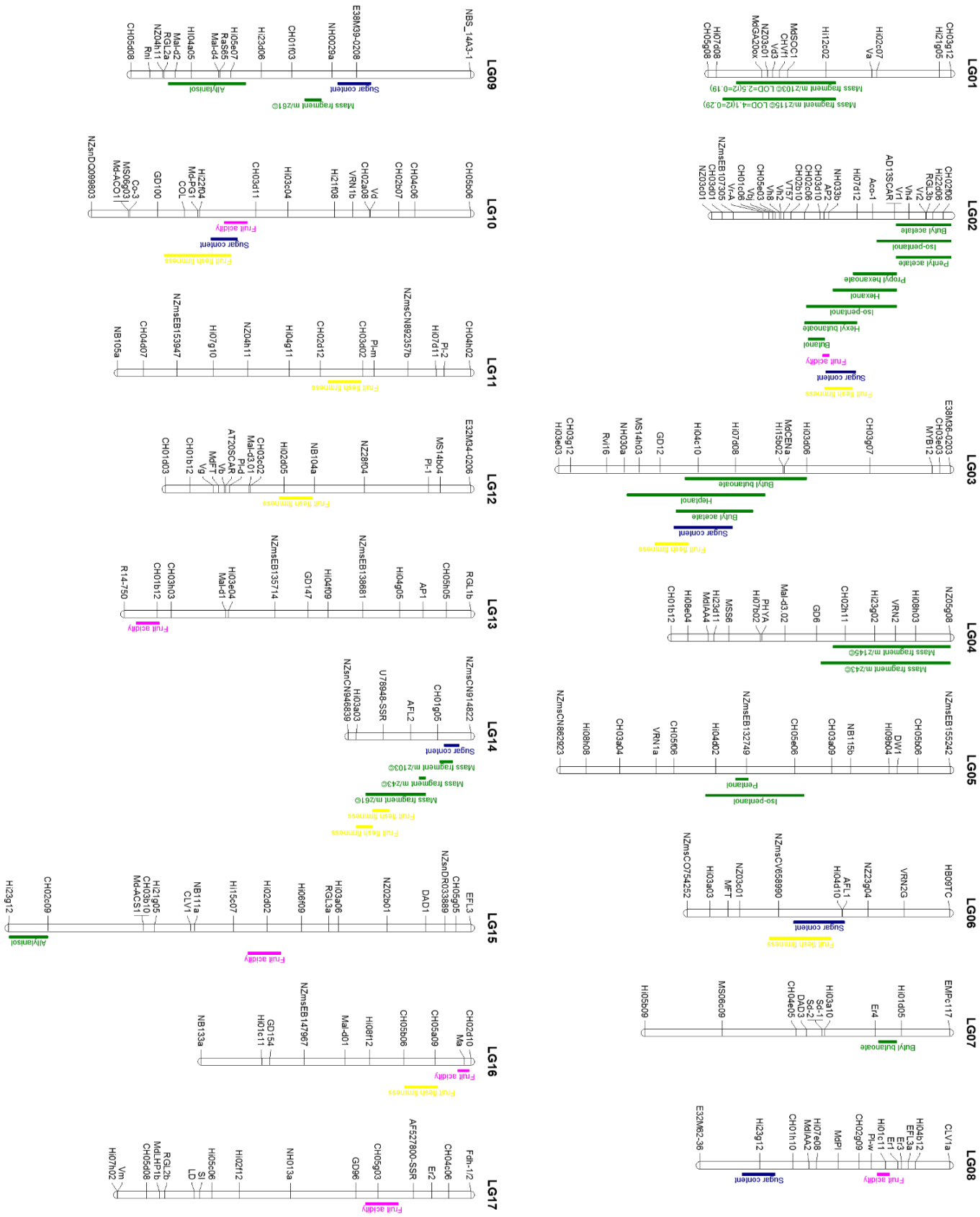


Figure 12 : Carte synthétique de QTLs de caractères de qualité du fruit identifiés chez le pommier par une approche de cartographie en descendance F1

goût (vert), acidité (rose), sucre (bleu foncé), fermeté (jaune)

et LG17 (Calenge et al. 2004; Lê Van et al. 2013). Des gènes majeurs codant vraisemblablement pour des protéines impliquées dans des interactions gène pour gène avec des effecteurs de *Venturia inaequalis* ont également été identifiés, chacun responsable d'un phénotype de résistance particulier (Bus et al. 2011). Concernant la résistance au feu bactérien, cinq groupes de liaison ont été identifiés comme étant porteurs de QTLs de résistance (Calenge et al. 2005; Khan et al. 2006; Khan et al. 2007; Durel, Denance, and Brisset 2009). A ce jour, une seule étude a permis l'identification d'un gène codant pour une protéine ayant une relation gène pour gène avec une protéine d'*Erwinia amylovora* (Vogt et al. 2013).

3.2. Etudes sur la qualité du fruit

3.2.1. Qualité des pommes à couteau

De nombreuses études de cartographie ont été menées sur la qualité des pommes à couteau. Ces études portent la plupart du temps sur les caractères de texture ou de goût, qui sont les caractères reconnus par les consommateurs et ayant donc une grande importance économique. La figure 12 répertorie un certain nombre des QTLs identifiés jusque-là concernant les caractères de texture et de goût (Liebhard et al. 2003; Zini et al. 2005; Kenis, Keulemans, and Davey 2008; Dunemann et al. 2009; Costa et al. 2010).

3.2.2. Qualité des pommes à cidre

A ce jour, une seule étude de cartographie par liaison génétique a été réalisée sur des individus ayant dans leur pedigree des pommiers à cidre (Verdu et al. 2014). Cette étude porte sur la localisation de QTLs responsables des teneurs en polyphénols dans la chair des pommes et a permis l'identification de régions génomiques portant des QTLs spécifiques d'une classe de polyphénols, ou au contraire d'un ensemble de classes. D'autres études ont été réalisées sur les QTLs de teneur en polyphénol sur des descendances de pommiers à couteau (Chagne, Krieger, et al. 2012; Khan, Chibon, et al. 2012). Ces deux études ont notamment permis l'identification d'une zone portant de multiples QTLs de teneur en polyphénols localisée sur le haut du LG16 (Figure 13). Une étude ultérieure (Khan, Schaart, et al. 2012) a montré que le gène responsable de ces QTLs serait le gène *LAR* codant pour une leucoanthocyanidine réductase et localisé en amont de la voie des procyanidines, responsables notamment de l'amertume du cidre.

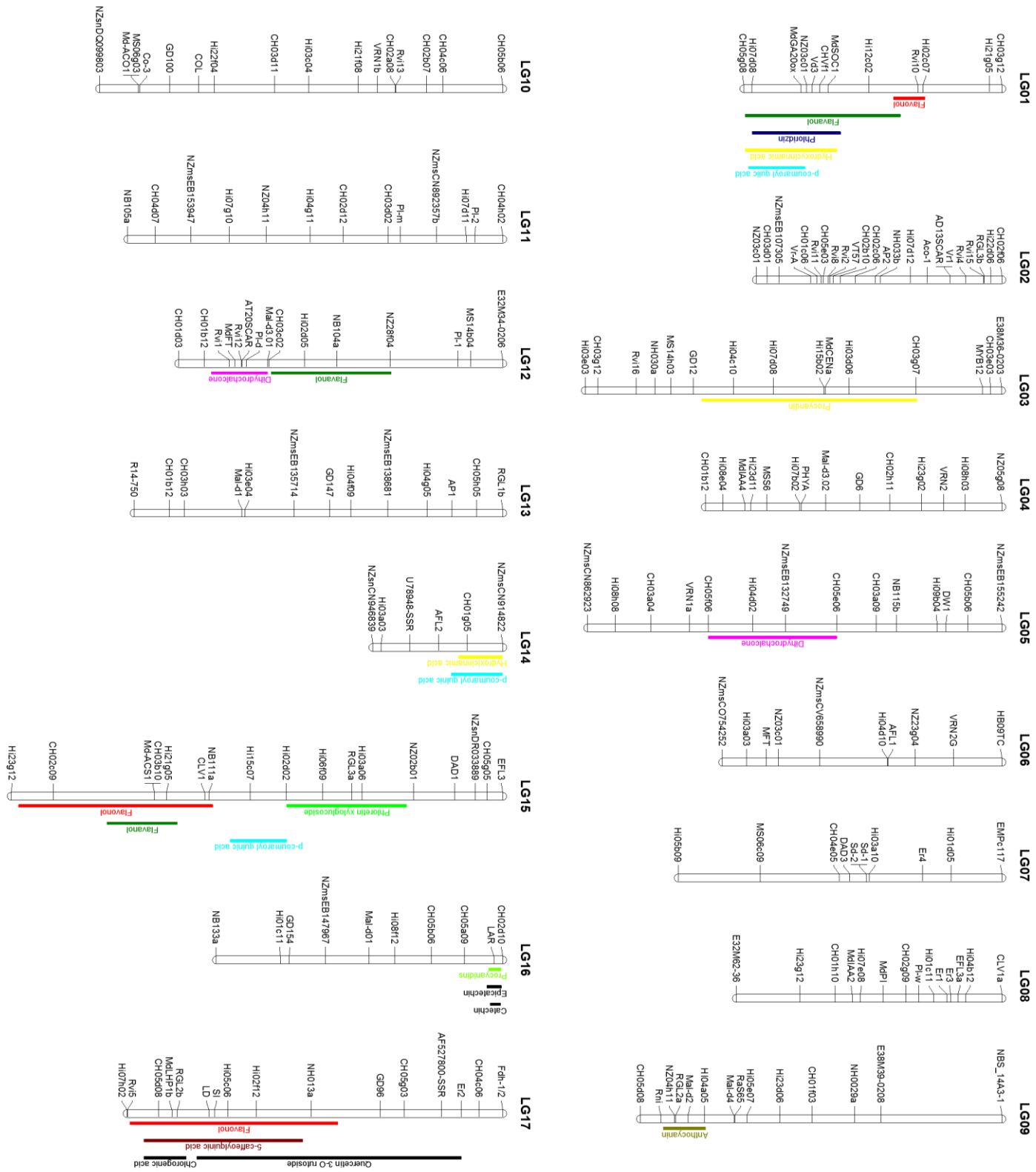


Figure 13 : Carte synthétique de QTLs de teneur en polyphénols identifiés chez le pommier par une approche de cartographie en descendance F1

flavonols (rouge), flavanol (vert foncé), dihydrochalcones (rose), phloridzine (bleu), acide *p*-coumaroyl quinique (cyan), anthocyanines (kaki), procyanidines (vert clair), acide hydroxycinnamique (jaune), divers (noir)

4. Objectifs de thèse

Ce premier chapitre bibliographique permet de mettre en avant les limites rencontrées lors de la recherche de facteurs génétiques responsables de la variation de caractères quantitatifs dans des descendances F1 ou dans des populations de pedigree. Du fait de la taille limitée des populations de cartographie généralement étudiées et du faible nombre de recombinaisons considéré, la localisation de ces facteurs est souvent imprécise. Par ailleurs, l'analyse d'une population biparentale hétérozygote ne permet d'identifier qu'au maximum quatre allèles à un locus donné. Enfin, le nombre de QTLs détectés dans une descendance reste généralement limité. Face à ces limites, la recherche de facteurs génétiques contrôlant les caractères par génétique d'association dans des collections regroupant une diversité génétique plus large paraît une démarche qui devrait permettre de localiser davantage de régions génomiques impliquées, et de les localiser plus finement du fait du plus grand nombre d'événements de recombinaison considérés.

L'objectif de cette thèse consistait donc à mettre en œuvre une première étude de génétique d'association chez le pommier en analysant une core collection regroupant une large gamme de variétés anciennes de pommiers à couteau. Une étude de cette ampleur n'a pour l'instant jamais été publiée chez le pommier. Une étude du déséquilibre de liaison préalable était prévue dès le départ de la thèse en élargissant l'analyse à une petite collection de pommiers à cidre. Cela nous a donné l'opportunité d'explorer plus avant les éventuelles traces de différenciation et de signature de sélection entre ces deux pools génétiques. Les données de génotypage à haute densité n'ayant finalement été disponibles que tardivement au cours de la thèse, les analyses de génétique d'association ont été réalisées en fin de thèse sans pouvoir approfondir autant que nécessaire les résultats obtenus.

Chapitre 2

Matériel végétal et méthodes générales

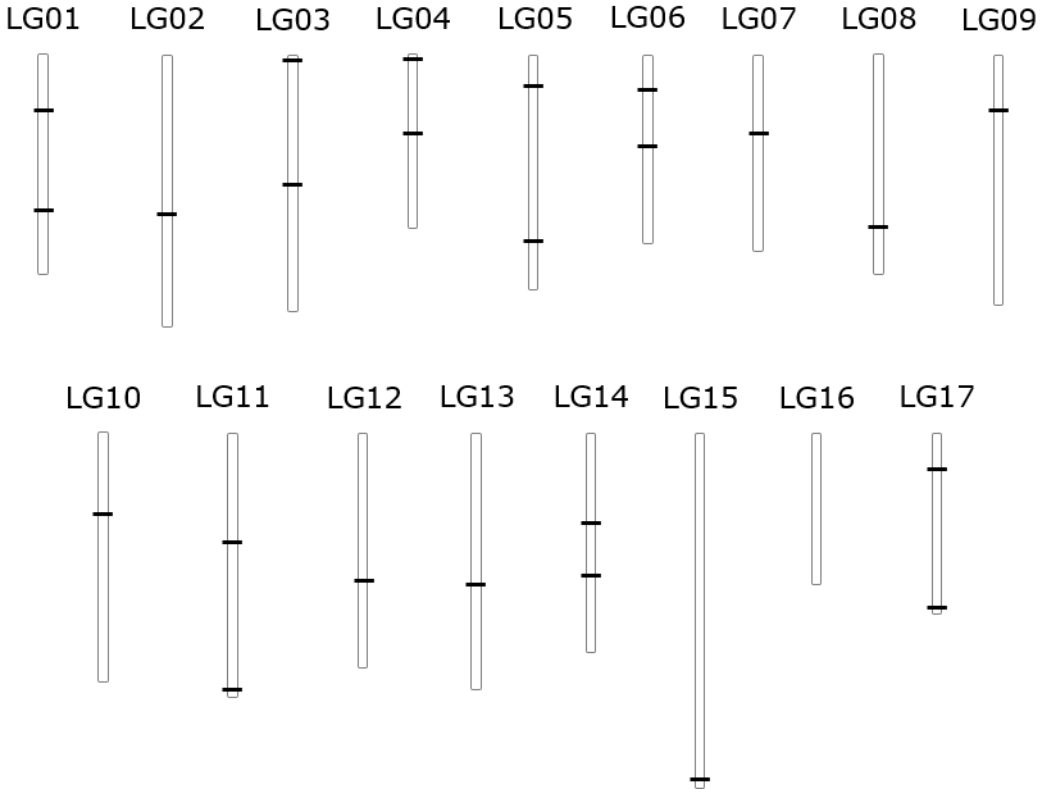


Figure 14 : Répartition des marqueurs microsatellites utilisés pour l'étude de diversité et la construction des core collections sur les 17 groupes de liaison du génome du pommier dans l'étude de Lassois et al. (2015)

1. Matériel végétal

Trois core collections de variétés anciennes de pommiers ont été étudiées au cours de cette thèse. Les variétés constituant ces core collections proviennent des vergers conservatoires de ressources génétiques de l'INRA d'Angers.

1.1. Présentation des collections INRA

L'INRA d'Angers possède une importante collection de variétés de pommiers parmi lesquelles on trouve des variétés anciennes et modernes, françaises et internationales, à cidre et à couteau, d'ornement et de consommation. Ces variétés, plus d'un millier, sont rassemblées au sein d'un Centre de Ressources Biologiques. L'INRA d'Angers est par ailleurs animateur du réseau national des conservatoires des fruits à pépins qui rassemble de nombreuses associations d'amateurs (Croqueurs de Pommes, Mordus de la Pomme, Verger Conservatoire de Pétré…), des jardins botaniques, et des collections régionales (Centre Régional de Ressources Génétiques du Nord-Pas de Calais, Centre Végétal Régional d'Aquitaine) et nationales (Jardin du Luxembourg), représentant une très grande diversité à l'échelle du territoire français. Une étude récente (projet CorePom) financée par la Fondation pour la Recherche sur la Biodiversité (FRB) a permis de caractériser génétiquement un ensemble de 2 163 accessions (~1 060 accessions de la collection INRA et ~1 100 accessions des autres partenaires) correspondant à un échantillonnage des variétés les plus diverses possibles du point de vue de la pomologie, des variétés locales et anciennes, des variétés synonymes ainsi que des variétés décrites et référencées qui ont servi de témoins de vérification pour les analyses de diversité décrites ci-dessous.

1.2. Définition des core collections

Préalablement à ce travail de thèse, des analyses de diversité ont été réalisées par l'équipe ResPom sur les individus rassemblés dans le cadre du projet CorePom. Chaque individu a été génotypé à l'aide de 24 marqueurs microsatellites (SSR) répartis de façon régulière sur les 17 groupes de liaison du pommier (Figure 14). La première étape a été d'identifier les individus « doublons », c'est-à-dire les individus présents sous des noms différents dans la collection mais présentant un profil génétique similaire, et de choisir un seul exemplaire à conserver dans les analyses suivantes. Un des buts du projet CorePom étant la constitution de core collections maximisant la diversité génétique en minimisant le nombre d'individus, un individu possédant les mêmes allèles qu'un autre devient de ce fait inintéressant. Sur les 2 163 accessions génotypées, 1 422 ont été considérées comme étant uniques

et ont été incluses dans les analyses de diversité. La deuxième étape de ce travail a été d'identifier les individus triploïdes qui représentent une proportion significative du total (18%), et dont la prise en charge par les différents logiciels est rare. Les individus pour lesquels au moins trois marqueurs sur 24 possédaient trois allèles distincts ont donc été écartés des analyses. La liste sur laquelle les analyses de diversité ont été réalisées, après retrait des individus ayant plus de 30% de données manquantes, était constituée de 188 variétés anciennes de pommes à cidre, 737 variétés anciennes de pommes à couteau et 159 variétés modernes de pommes à couteau. Cette thèse portant principalement sur la recherche de facteurs d'intérêt dans les variétés anciennes de pommes, les variétés modernes n'ont pas été prises en compte dans les analyses. Trois core collections ont été constituées à partir de ces données de génotypage (Lassois et al. 2015), et utilisées au cours de cette thèse:

- une core collection de variétés anciennes de pommiers à cidre contenant les 48 individus représentant le mieux la diversité génétique de la collection INRA de pommiers à cidre et nommée ci-après CC48 Cidre,
- une core collection de variétés anciennes de pommiers à couteau contenant les 48 individus représentant le mieux la diversité génétique de la collection INRA de pommiers à couteau et nommée ci-après CC48 Couteau,
- une core collection de variétés anciennes de pommiers à couteau contenant les 278 individus représentant le mieux la diversité génétique de la collection INRA de pommiers à couteau (et donc contenant les individus de la CC48 Couteau), et nommée ci-après CC278 ; par ailleurs, une dizaine d'individus supplémentaires ont été choisis en tant qu'individus « bonus » dans le cas où certaines variétés de la CC278 n'avaient pas assez de données de phénotypage ou de génotypage.

Lors de cette thèse, les deux CC48 ont été utilisées afin d'identifier les bases génétiques de la différenciation entre pommiers à cidre et pommiers à couteau. La CC278 a quant à elle été utilisée pour une étude de génétique d'association visant à identifier de nouveaux facteurs d'intérêt chez les pommiers à couteau.

2. Phénotypage du matériel végétal

Dans le but de réaliser une étude de génétique d'association sur des variétés anciennes de pommiers à couteau, une série de phénotypages a été réalisée sur plusieurs caractères d'intérêt agronomique dans les programmes de sélection actuels. Les données de résistance ont été acquises dans le cadre de ma thèse, les données de qualité du fruit ont été acquises par ailleurs.

Tableau 2 : Liste des variétés de pommiers utilisées comme témoins de sensibilité ou de résistance pour les tests de résistance à la tavelure et au feu bactérien

Tests tavelure	Témoins communs	Test feu bactérien
« Ariane » « Florina » « Gala » « TN10-8 »	« Discovery » « Golden Delicious » *Malus floribunda* « Red Pippin »	« Enterprise » « Evereste » « Idared » « Prima »

2.1. Acquisition des données de résistance

2.1.1. Tests de résistance à la tavelure

Au total, trois tests de résistance à la tavelure ont été réalisés au printemps 2012, au printemps 2013 et à l'automne 2013. Pour les tests réalisés au printemps, des baguettes de bois de l'année ont été prélevées sur les individus de la CC278 dans les parcelles de l'INRA et greffées sur des porte-greffes « MM106 » en huit exemplaires par variété. Pour le test réalisé à l'automne, les plants utilisés pour le test du printemps ont été traités avec des fongicides à la fin de l'essai, rabattus, puis laissés en serre durant l'été. Une dizaine de variétés, témoins de sensibilité ou de résistance, ont été rajoutées à chaque test (Tableau 2), ainsi que plusieurs variétés notoirement ou présumées résistantes. Les greffes ont été réalisées dans le courant du mois de février de chaque année et les plantes ont ensuite été placées en serre. Les greffons provenant de nombreuses variétés anciennes, la pousse des plantes n'a pas été aussi homogène que pour un test sur descendance. De nombreuses plantes ont débourré mais ont rapidement arrêté leur croissance, les rendant impropres à l'inoculation. Les plantes présentant des caractéristiques visibles de croissance ont été choisies et placées de façon aléatoire sur les tablettes, puis inoculées. Pour chaque test, plusieurs inoculations ont été réalisées en fonction du stade de développement des plantes. Les inocula ont été préparés par P. Expert de l'équipe EcoFun à partir de feuilles de pommiers sur lesquelles ont été multipliées les différentes souches :

- l'inoculum du printemps 2012 correspondait à la souche monoconidiale EU-B04 (race 1,10 selon la classification de Bus et al. (2011) ; voir également Caffier et al. (2014) ;
- l'inoculum du printemps 2013 contenait un mélange équiprobable des cinq souches 104/163/EU-B04/EU-NL24/EU-D42, correspondant respectivement aux races (1), (5), (1,10), (1,3,6,7) et (1,6,10) ;
- l'inoculum de l'automne 2013 contenait la souche monoconidiale 104 (race 1).

Lors de l'inoculation, l'inoculum est pulvérisé sur les plantes de façon à ce que des gouttelettes se forment à la surface des feuilles, mais en évitant d'atteindre le stade du ruissellement de manière à ne pas enlever l'inoculum des feuilles. Les plantes sont ensuite placées sous une bâche plastique pendant 48 heures, ce qui assure le maintien d'une hygrométrie maximale. La température est ensuite maintenue à 17°C, le niveau d'hygrométrie entre 75 et 85% d'humidité et la luminosité réduite à l'aide d'ombrières.

Une fois l'inoculation effectuée, les plantes ont été notées à 7, 14, 21 et 28 jours post-inoculation. Pour cela, deux échelles de notations ont été utilisées :

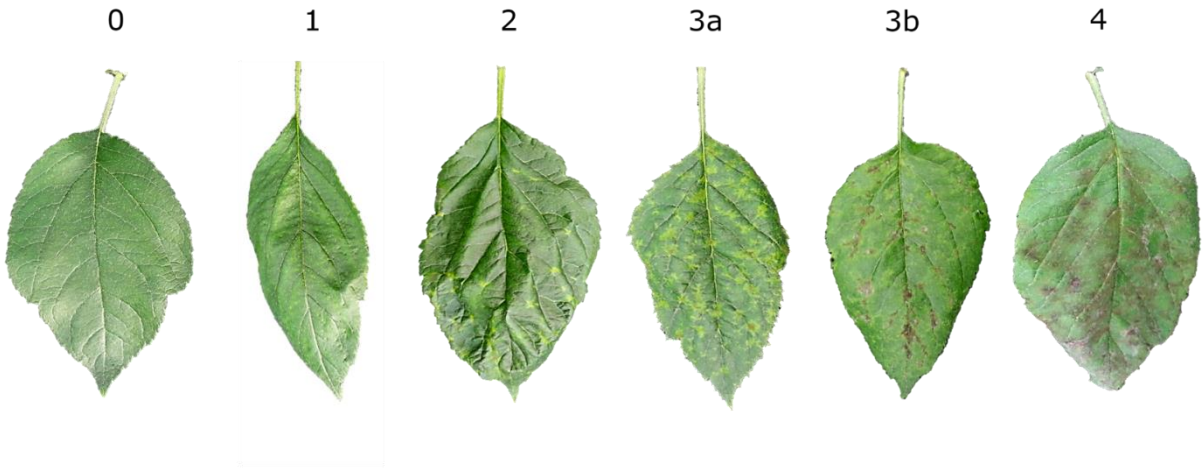


Figure 15 : Classes phénotypiques de pommiers infectés par le champignon responsable de la tavelure *Venturia inaequalis* selon Chevalier et al. (1991)

classe 0 : aucun symptôme visible ; classe 1 : symptômes caractéristiques de « pin-point » ; classe 2 : symptômes de résistance (chlorose, nécrose, crispation) sans sporulation ; classe 3a : symptômes de résistance avec quelques tâches de sporulation peu abondante ; classe 3b : symptômes de résistance avec tâches de sporulation abondante ; classe 4 : pas de symptôme de résistance et sporulation abondante

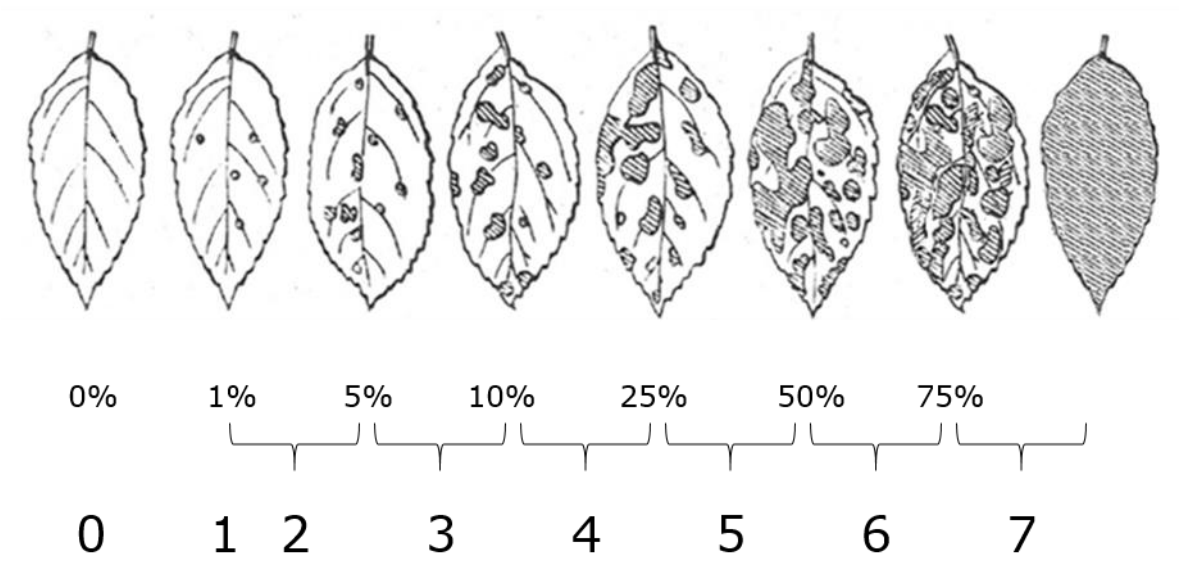


Figure 16 : Adaptation de l'échelle de sévérité de sporulation de Croxall et al. (1952)

0 : aucune sporulation visible ; 1% de surface foliaire couvert de sporulation : 1 ; de 1 à 5% : 2 ; de 5 à 10% : 3 ; de 10 à 25% : 4 ; de 25 à 50% : 5 ; de 50 à 75% : 6 ; de 75 à 100% : 7

- la première note évalue la classe de résistance à laquelle un individu appartient (Chevalier, Lespinasse, and Renaudin 1991) ; cette échelle est qualitative et classe les individus en six catégories (Figure 15) ;
- la seconde note, adaptée de (Croxall, Gwynne, and Jenkins 1952), est quantitative et reflète le pourcentage de surface foliaire couvert par la sporulation du champignon (Figure 16).

La présence de symptômes de résistance, à savoir la nécrose, la chlorose et la crispation des feuilles (Figure 17), a également été notée et chaque symptôme a été quantifié individuellement selon une échelle allant de 0 à 3 en fonction de l'intensité du symptôme. La notation a été effectuée sur une seule feuille par plante en prenant soin de choisir la feuille présentant le plus de symptômes de la maladie. Quelques traitements ponctuels anti-oïdium et acaricides inefficaces contre la tavelure ont eu lieu pendant les essais.

2.1.2. Tests de résistance au feu bactérien

Tout comme pour les tests de résistance à la tavelure, des greffons des variétés de la CC278 ont été prélevés en champ et greffés sur porte-greffes « MM106 » en huit exemplaires, ainsi que quelques variétés témoin (Tableau 2). Plusieurs inoculations ont également eu lieu pour le test de résistance au feu bactérien, en fonction de l'état de croissance des plantes. L'inoculation avec *Erwinia amylovora* nécessite en effet une bonne croissance des plants et une taille minimale de la pousse inoculée de 15 cm. Dans le cas contraire, si une plante non poussante est inoculée et qu'elle reste apparemment saine, on ne peut pas savoir si elle est réellement résistante ou si elle a échappé à la maladie du fait de l'absence de croissance. L'inoculum, gardé à 4°C avant utilisation, est une solution de bactéries de la souche CFBP1430 (souche de référence de la Collection Française de Bactéries associées aux Plantes ; https://www6.inra.fr/cirm/CFBP-Bacteries-associees-aux-Plantes), à concentration 10^7 unités par millilitre, préparée par R. Chartier de l'équipe ResPom. Les plantes sont inoculées de la façon suivante : avec des ciseaux trempés dans l'inoculum on coupe la première feuille entièrement déroulée à l'extrémité de la pousse, puis on répète l'opération sur la 2ème feuille la plus jeune. La température de la serre est ensuite réglée à 22°C le jour et 18°C la nuit, l'hygrométrie à 80% d'humidité et les ombrières se ferment lorsque 600W par mètre carré sont dépassés.

Une fois l'inoculation effectuée, les plantes ont été notées à 7, 14 et 21 jours post-inoculation. Plusieurs mesures ont été effectuées sur chaque plante :

- une première mesure de la longueur totale de la pousse inoculée ;
- une mesure de la longueur de nécrose totale.

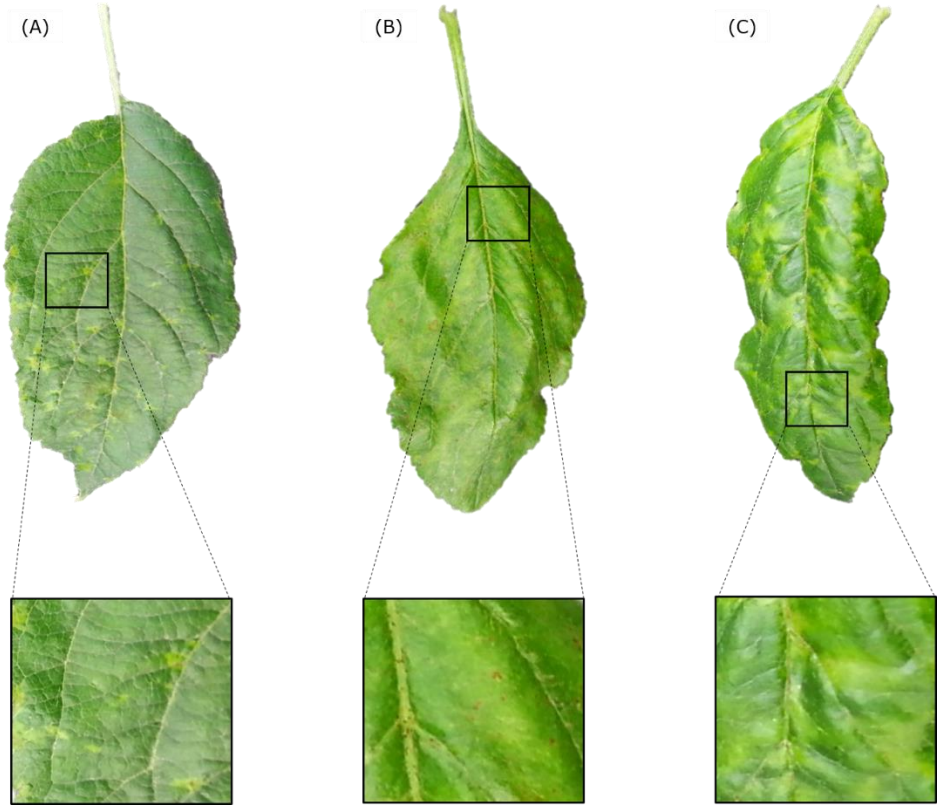


Figure 17 : Symptômes de résistance observés lors d'une infection de feuilles de pommier par *Venturia inaequalis*, l'agent de la tavelure

(A) chlorose ; (B) nécrose ; (C) crispation

Tableau 3 : Caractères évalués par des dégustations de fruits dans la core collection de pommiers à couteau

Caractère	Définition
Acidité	Perception de l'intensité de la sensation acide
Ratio Sucre/Acidité	Perception du rapport entre sucre et acidité
Croquant	Quantité de bruit lors de la première morsure
Teneur en fibres	Perception physique de la fibre de la chair
Fermeté	Force requise pour mordre dans la pomme
Granulosité	Perception physique des grains de la chair
Jutosité	Quantité de jus libérée de la chair
Farinosité	Perception de la chair sèche et granuleuse
Fondant	Force requise pour écraser la chair entre la langue et le palais
Russeting	Estimation visuelle de l'intensité des marques liégeuses se développant sur l'épiderme
Sucre	Perception de l'intensité de la sensation sucrée
Goût	Perception de l'intensité aromatique du fruit

Les plantes présentant une nécrose uniquement au niveau des nervures des feuilles inoculées ont été notées NN, pour « nécrose nervure », et celles présentant en plus des nécroses au niveau du pétiole de ces feuilles NNP, pour « nécrose nervure et pétiole ».

2.2. Acquisition des données de qualité du fruit

Une série de dégustations effectuées par un panel d'experts des équipes VaDiPom et FruitQual de l'IRHS a été réalisée sur les fruits des individus de la CC278 récoltés à maturité physiologique, pour les années 2012, 2013 et 2014 et sur les 12 caractères présentés dans le Tableau 3. Les notes attribuées pour chaque caractère vont de 1 à 9. Selon les années, plusieurs modalités de dégustation ont été mises en place :

- en 2012, plusieurs notateurs faisant partie de plusieurs binômes ont dégusté un même fruit sans la peau et attribué une note pour chacun des caractères ; une note consensus a ensuite été attribuée par fruit et par binôme (dans le cas d'une trop grande différence entre les notes de notateurs d'un même binôme, un deuxième fruit a pu être dégusté),
- en 2013 et 2014, deux notateurs ont dégusté indépendamment un même fruit sans la peau et ont ensuite ajusté les notes entre eux (dans le cas d'une trop grande différence entre les notes des deux notateurs, un deuxième fruit a pu être dégusté).

3. Extractions d'ADN

Les feuilles des individus des core collections ont été récoltées en plaques 96 puits et en tubes Eppendorf 2 mL, soit en champ, soit en serre, et placées à -80°C. Les extractions ont été réalisées en utilisant le kit NucleoSpin® Plant II (Macherey-Nagel GmbH and Co KG, Düren, Germany) en suivant les indications du protocole fourni, et les ADN ont été dosés en utilisant le Nanodrop 1 000. La plupart des extractions n'ayant pas permis l'obtention d'ADN en quantité ou en qualité nécessaires pour les étapes de génotypage, les ADN des individus concernés ont été ré-extraits en utilisant les feuilles stockées dans les tubes Eppendorf. Une fois la quantité d'ADN nécessaire atteinte, les ADN ont été précipités à l'éthanol (voir protocole du Chapitre 3, 2.1) afin de garantir une qualité optimale notamment pour l'hybridation des ADN sur les différentes puces de génotypage.

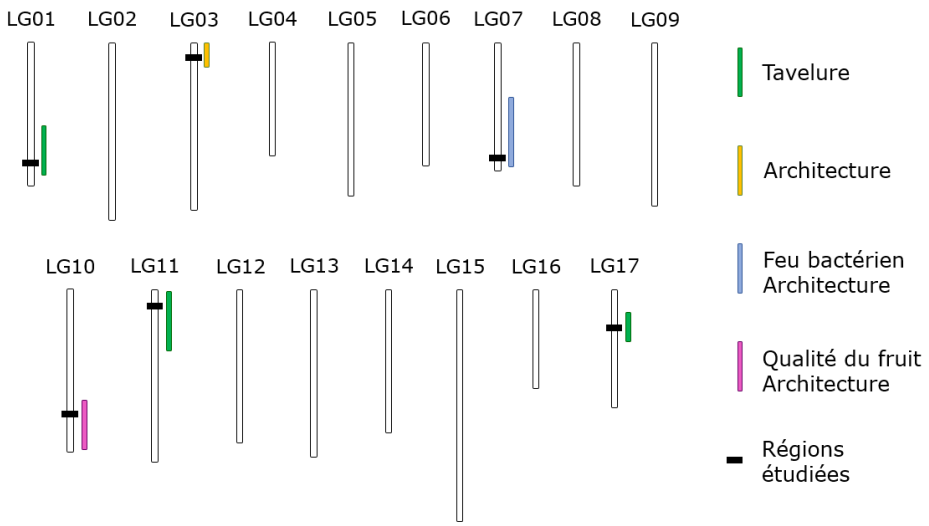


Figure 18 : Localisation des régions génomiques portant les gènes re-séquencés et nature des QTLs associés

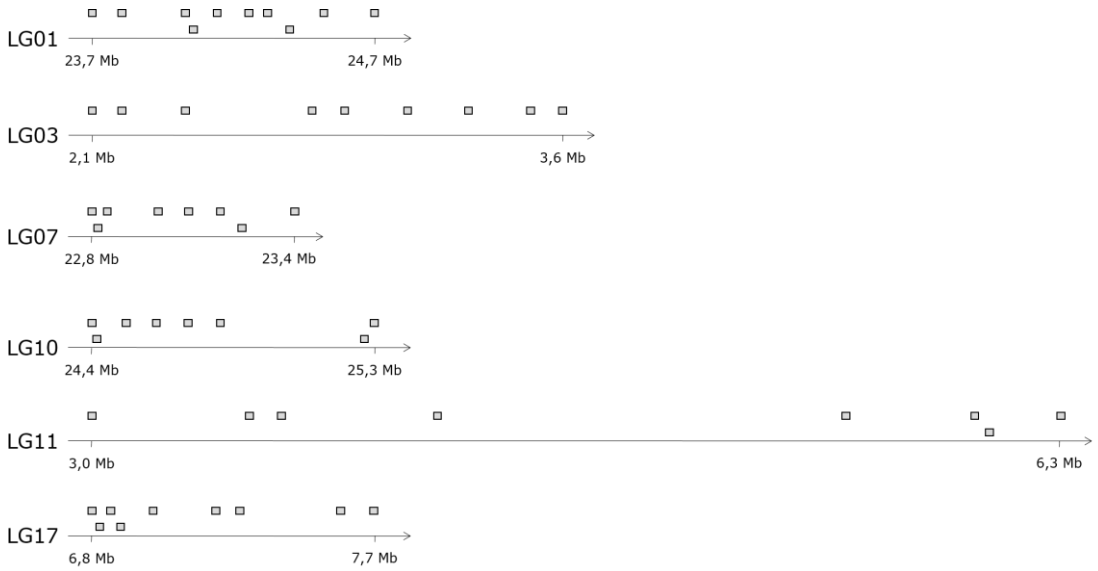


Figure 19 : Représentation schématique de la répartition des 52 gènes candidats positionnels dans les 6 régions du génome du pommier contenant des QTLs d'intérêt

4. Génotypages

Lors de cette thèse, deux techniques de génotypage ont été utilisées sur les ADN des individus des différentes core collections. La première technique de génotypage s'appuie sur deux puces SNP de différentes densités ; l'International RosBREED SNP Consortium (IRSC) apple 8k SNP array v1 (Chagne, Crowhurst, et al. 2012) et l'Axiom-Apple-480k SNPs (Bianco, Durel, Troggio, et al., in prep.) ont été utilisées pour réaliser le génotypage des individus des deux CC48 et des individus de la CC278 respectivement. Les données de la puce 8k ont été utilisées pour étudier la différenciation génétique entre les pommiers à cidre et les pommiers à couteau, et les données de la puce 480k pour réaliser l'étude de génétique d'association. Le premier génotypage a été réalisé sur la plateforme « SNP, Transcriptomique et Epigénomique » (STE) du CHU d'Angers et le second sur la plateforme « Gentyane » de l'INRA de Clermont-Ferrand (équipe de Charles Poncet), dans le cadre du projet européen FruitBreedomics. A la suite de ce génotypage, 4 234 marqueurs SNP ont été retenus pour la puce 8k (criblage visuel) et 240 129 ont été classés comme faisant partie de la classe des Poly High Resolution par le workflow d'Affymetrix® (plus de 99% de reproductibilité ; travail réalisé par C. Denancé), et ont donc été utilisés dans les analyses pour la puce 480k SNPs.

La seconde technique de génotypage, correspondant à du re-séquençage de gènes, a été réalisée uniquement sur les individus de la CC278 grâce à un financement d'une AIP Bio-ressources. Ce génotypage a été réalisé en collaboration avec l'équipe « Etude du Polymorphisme des Génomes Végétaux » (EPGV) d'Evry. Afin d'étudier le déséquilibre de liaison à fine échelle, six régions génomiques co-localisant avec des QTLs d'intérêt (Figure 18) ont été choisies dans le génome du pommier. A l'intérieur de ces six régions, des gènes candidats positionnels, avec des distances croissantes entre paire de gènes, ont été choisis (Figure 19). Un support bio-informatique d'Evry a permis le design de couples d'amorces en théorie spécifiques de ces gènes, à raison d'une ou deux régions amplifiées par gène, en fonction de sa structure. L'ancêtre du pommier ayant subi une auto-polyploïdisation de son génome, certains couples d'amorces sont susceptibles d'amplifier deux fragments sur des groupes de liaison différents mais de séquence homologue. Les amorces ont principalement été choisies dans des régions exoniques, qui sont plus conservées que les régions introniques, afin de maximiser les chances d'amplification sur de si nombreuses variétés. Un total de 96 couples d'amorces répartis dans 52 gènes ont ainsi été designés. Les PCR pré-séquençage ont été réalisées en utilisant la technique Integrated Fluidic Circuits (IFC) de Fluidigm® qui permet l'amplification simultanée de 48 couples d'amorces sur 48 individus (Figure 20). Chaque amorce est dotée d'une séquence « code barre » qui va permettre, à la fin de la PCR, de réunir les 48 amplicons d'un même individu dans une seule solution et de pouvoir différencier ces différents amplicons lors de l'étape de séquençage. Une fois les PCR effectuées, les solutions contenant les pools d'amplicons ont été passées sur un séquenceur MiSeq d'Illumina® qui génère des reads de 250 pb en forward et en reverse, soit des reads de 500 pb.

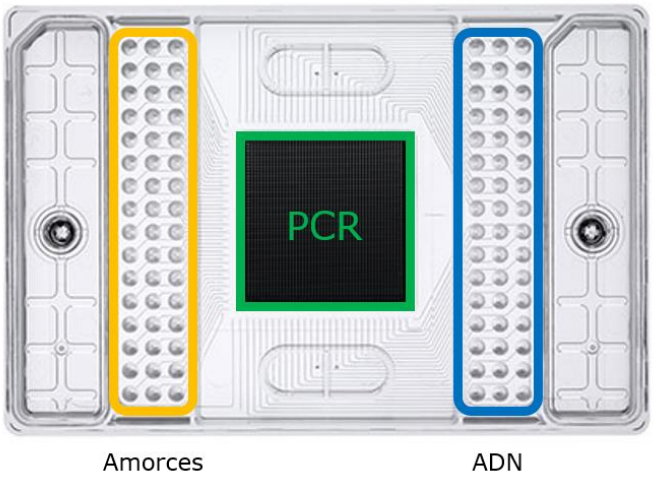


Figure 20 : Schéma d'une plaque Integrated Fluidic Circuits (IFC) de Fluidigm®
les 48 puits en jaune reçoivent les 48 couples d'amorces et le mélange réactionnel, les 48 puits en bleu reçoivent les ADN de 48 individus, les différentes solutions sont ensuite distribuées à l'aide d'un système de capillaires dans la cellule en vert contenant 48 x 48 puits, où ont lieu les PCR simultanées

5. Analyses statistiques et bio-informatiques

5.1. Bases génétiques de la différenciation entre pommes à cidre et pommes à couteau

5.1.1. Analyses de différenciation génétique

Dans le but d'identifier et de localiser les régions génomiques responsables des différences phénotypiques entre pommes à cidre et pommes à couteau, des analyses de structure et de différenciation génétique (F_{ST}) ont été réalisées (voir Chapitre 3, 2.1).

5.1.2. Analyses de génétique d'association

Une étude de génétique d'association prenant en compte soit le type variétal de pommes, à cidre ou à couteau, soit le caractère amer de certaines pommes à cidre, a été réalisée en corrigeant par la structure et l'apparentement (voir Chapitre 3, 2.1). Les notes 0 et 1 ont été attribuées aux individus de type couteau et cidre respectivement, puis aux individus doux et amers, afin de réaliser une analyse d'association de type case-control.

5.1.3. Analyses de génétique des populations

A partir des données de génotypage de la puce 8k, les hétérozygoties observée (H_o) et attendue (H_e) ont été calculées avec le logiciel GENETIX 4.05 (Belkhir et al. 1996-2004). Plusieurs estimateurs de génétique des populations ont été calculés, en collaboration avec A. Branca et T. Giraud du laboratoire « Ecologie, Systématique et Evolution » (ESE) de l'université d'Orsay, avec une approche de fenêtres glissantes, pour une taille de fenêtre de 20 SNPs et un pas de 5 SNPs, en utilisant la librairie libsequence (Thornton 2003). L'écart à la neutralité des sites a été étudié en calculant le D de Tajima, le F de Fu et Li et le H de Fay et Wu. L'état ancestral des allèles a été obtenu en alignant le génome de la pomme à celui de la pêche (*Prunus persica*) en utilisant le logiciel YASS (Noé and Kucherov 2005). Seuls les blocs orthologues de plus de 1 000 pb ont été gardés pour déterminer l'état ancestral de l'allèle. Le génome de la pêche (Verde et al. 2013) a été utilisé plutôt que celui de la poire, pourtant plus apparenté, parce que, tout comme le génome de la pomme, celui de la poire a subi une duplication

du génome assez récente (Velasco et al. 2010), ce qui aurait rendu la détermination de l'état ancestral moins fiable à cause des potentiels paralogues. Les valeurs extrêmes des différents estimateurs, D, F et H, ont été choisies au-dessus du 95ème percentile et les fenêtres contenant potentiellement des signatures de sélection ont été choisies quand plusieurs estimateurs montraient des valeurs extrêmes.

5.2. Variations du déséquilibre de liaison

5.2.1. Obtention des fichiers de génotypage par analyses bioinformatiques

Les données de re-séquençage de gènes en sortie du séquenceur MiSeq ont été traitées avec le logiciel CLC Genomics Workbench (http://www.clcbio.com). Le logiciel CLC Genomics permet de charger les données brutes issues du séquenceur MiSeq et d'en extraire les données de génotypage. Des vérifications concernant la qualité des séquences ont été effectuées. Chacun des reads ayant passé les contrôles qualité a été aligné avec la séquence de référence issue de la première version du génome du pommier (Velasco et al. 2010), et un fichier d'alignement (.sam) a été généré pour chaque fragment et chaque individu. L'outil Variant Detection du package Resequencing Analysis de CLC Genomics a été utilisé sur ces alignements afin de détecter, pour un individu et un fragment donnés, les positions auxquelles plus de 30% des reads étaient différents de la séquence de référence. Les variants trouvés chez tous les individus ont ensuite été mis en commun et une liste de SNPs par fragment établie (Annexe 1). Cette étape a été réalisée afin de faciliter les analyses réalisées sur les données de re-séquençage en faisant en sorte que tous les individus possèdent des haplotypes de même longueur pour chacun des fragments.

Dans un deuxième temps, un script perl permettant d'extraire les données de génotypage, écrit par P. Barre de l'INRA de Lusignan avec quelques ajouts de M. Leforestier (Annexe 2), a été utilisé. Il se décompose en six parties :

- la première fonction du script est de lire le code CIGAR présent dans les fichiers d'alignement obtenus en sortie de CLC Genomics ;
- à partir de ce code, la séquence de chacune des paires de reads est extraite et retranscrite dans un fichier texte ; les reads forward sont placés sur les lignes impaires et les read reverse sur les lignes paires ;
- pour chaque paire de reads, si le read forward et le read reverse sont présents dans le fichier, le script permet le phasage des données en fusionnant les deux lignes concernées ; dans le cas où le nombre de bases lues par le séquenceur pour les

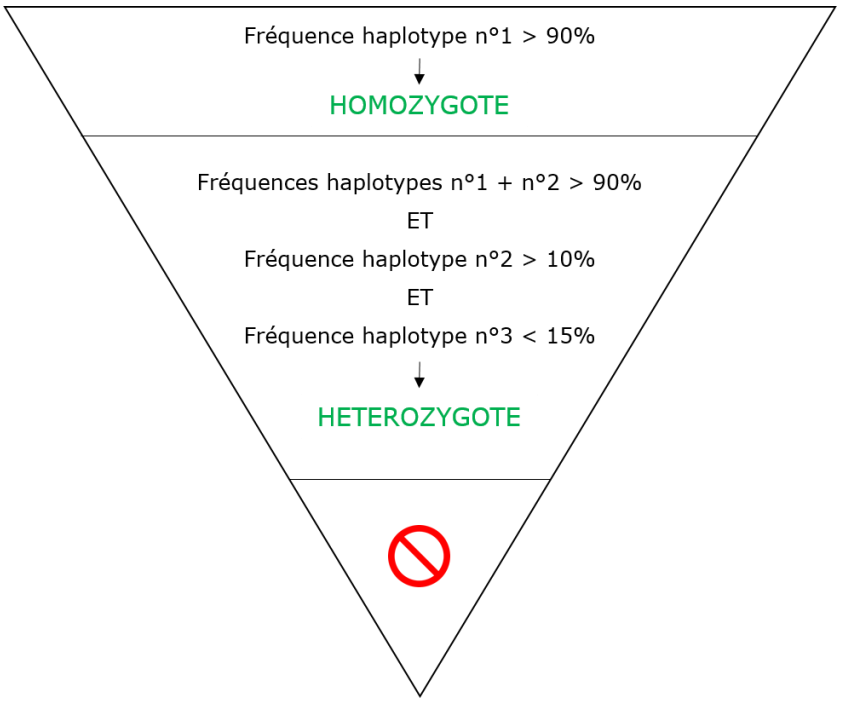


Figure 21 : Tri appliqué aux haplotypes contenus dans les fichiers de re-séquençage identifiés à l'aide d'un script Perl permettant de différencier les haplotypes homozygotes et hétérozygotes en fonction de leur fréquence

les fragments pour lesquels aucune décision n'a pu être prise ont été supprimés des analyses

reads forward et reverse est supérieur au nombre de bases de la séquence, c'est-à-dire s'il y a chevauchement des reads sur la partie médiane de la séquence, seules les bases provenant du read forward sont conservées ;

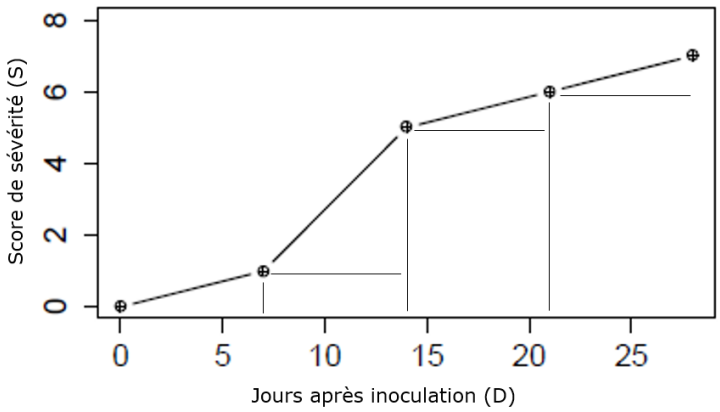
- en utilisant le fichier où sont stockées les positions des SNPs par fragment, le script extrait l'information de séquence uniquement à ces positions pour tous les reads générant ainsi des haplotypes ;
- les haplotypes identiques sont mis en commun et comptabilisés ;
- le script calcule la fréquence des six haplotypes les plus représentés et dont la fréquence est supérieure à 0,1 et génère un fichier par individu et par fragment de gène contenant cette information, ainsi que les haplotypes en question.

Plusieurs filtres ont ensuite été appliqués aux fichiers de sortie du script perl afin de différencier les individus homozygotes des individus hétérozygotes à un locus donné, et d'éliminer, le cas échéant, les fichiers pour lesquels aucune décision ne pouvait être prise (Figure 21). En effet, l'étape de PCR par Fluidigm® et le re-séquençage MiSeq peuvent avoir entrainé des erreurs de génotypage générant par la suite des faux haplotypes au sein d'un individu diploïde, pour lequel on attend pourtant seulement deux haplotypes en situation hétérozygote.

Dix génotypes ont été re-séquencé en double afin de permettre une estimation du taux d'erreurs de séquençage. Les mêmes filtres que précédemment ont été appliqués aux fichiers et les différences entre les deux fichiers d'un même fragment et d'un même individu ont été comptabilisées puis comparées avec le nombre total de comparaisons. Les positions auxquelles la base lue était un N n'ont pas été prises en compte dans ces calculs.

5.2.2. Analyses statistiques

Les données de génotypage de la puce 480k ont été utilisées pour étudier l'étendue du DL au niveau du génome entier. Les données manquantes ont d'abord été imputées grâce au logiciel Beagle 4.0 (Browning and Browning 2007). Les marqueurs, répartis tous les 3 kb en moyenne le long du génome, ne comprennent pas les SNPs localisés sur le LG0 de la version 3 du génome du pommier (le LG0 a été constitué avec tous les contigs qui n'ont pas pu être placés sur les autres groupes de liaison, faute d'alignement de séquence), ni les SNPs attribués aux groupes de liaison mais sans position attribuée. Un programme en fortran, écrit par C. Carillier et A. Ricard de l'INRA de Toulouse, a ensuite été utilisé pour calculer le r^2 entre chaque paire de marqueurs, sur un même groupe de liaison et pour une distance entre marqueurs ne dépassant pas 500 kb. Les valeurs obtenues par groupe de liaison ont ensuite été compilées et le r^2 moyen calculé pour des incréments de 10 kb. Les données des puces 8k et 480k ont été analysées avec le logiciel Haploview 4.2 (Barrett et al. 2005). Ce logiciel permet le



$$AUDPC = \sum_{i=7}^{D=28} [\,(\frac{(D_{i+7} - D_i)(S_{i+7} - S_i)}{2}) + (D_{i+7} - D_i)(S_i - S_0)] + \frac{(D_7 - D_0)(S_7 - S_0)}{2}$$

Figure 22 : Graphique représentant l'aire sous la courbe de progression de la maladie (AUDPC) et formule utilisée pour la calculer, pour une notation hebdomadaire

calcul des différents estimateurs du DL (r^2 et D'), la visualisation de ces estimateurs entre chaque paire de marqueurs et la constitution de blocs haplotypiques qui indiquent les marqueurs en fort DL (voir Chapitre 3, 2.1).

Les données de re-séquençage de gènes ont par ailleurs été utilisées dans le but d'étudier le déséquilibre de liaison à très fine échelle, en situation intra- et inter-génique. Pour cela, le logiciel Haploview 4.2 a été utilisé. Les analyses ont été réalisées sur les SNPs des 6 régions étudiées séparément, en prenant comme seuils pour la fréquence de l'allèle mineur (MAF), pour l'équilibre de Hardy-Weinberg et pour le pourcentage de données manquantes autorisé 0,01, 0,001 et 0,4 respectivement. La décroissance moyenne du déséquilibre de liaison a été calculée par incréments de 1,25 kb et représentée sous forme de graphique dans un environnement R. Il a été choisi, au vu des résultats obtenus lors de l'utilisation de la puce 8k SNPs, de ne pas corriger les valeurs de r^2 obtenues, ni par la structure, ni par l'apparentement.

5.3. Recherche de régions génomiques contrôlant la variation de caractères quantitatifs par génétique d'association

5.3.1. Analyses statistiques des données de phénotypage

Avant d'analyser les différents jeux de données de résistance, les notes recueillies pour chaque test ont subi une transformation. Pour les données de résistance à la tavelure des différentes variétés, l'aire sous la courbe de progression de la maladie (AUDPC) a été calculée (Figure 22). Une note de 0 a été attribuée à tous les individus pour la date de l'inoculation. Les données de résistance au feu bactérien ont quant à elles été utilisées pour calculer le pourcentage de la tige atteint par la nécrose à 21 jours. Une longueur de tige nécrosée de 0,5 cm a été attribuée aux individus notés NN, une longueur de 1 cm aux individus notés NNP, et un centimètre a été rajouté aux longueurs mesurées sur les individus pour lesquels la nécrose avait atteint la tige. La note de chaque individu a ensuite été calculée en divisant la longueur de nécrose par la longueur totale de la pousse. Par ailleurs, il a été décidé ici de retirer des analyses les génotypes pour lesquels une ou deux copies seulement ont été notées. En effet, bien que les plantes aient poussé en serre que l'on pourrait considérer comme un environnement contrôlé, certains facteurs dont la variation au sein des serres est mal contrôlée, jouent sur la croissance des pathogènes sans que l'on puisse exactement quantifier cet effet. C'est le cas par exemple du système de chauffage de la serre qui est localisé à une extrémité des tablettes sur lesquelles poussent les plantes, et qui réduit considérablement le développement de *Venturia*. Les pousses présentant des

notes aberrantes, c'est-à-dire les pousses NN ou NNP appartenant à un génotype pour lequel la majorité des pousses présentait une nécrose importante, ont également été éliminées de l'analyse des données de résistance au feu bactérien.

Des analyses de variance ont été menées à l'aide du logiciel R en utilisant les données de résistance. Le modèle d'analyse de variance appliqué est le suivant :

$$Y_{ijk} = \mu + B_i + G_j + (B \times G)_{ij} + \varepsilon_{ijk}$$

dans lequel Y_{ijk} est le phénotype de la répétition k du génotype j dans le bloc i, μ la moyenne générale de la population, B l'effet du bloc i, G l'effet du génotype j, (B x G) l'interaction entre le génotype et le bloc et ε_{ijk} l'effet résiduel. Dans le cas des tests de résistance à la tavelure, B correspond aux effets combinés de la localisation de la copie phénotypée dans la serre (par bloc) et de la date d'inoculation. Dans le cas des tests de résistance au feu bactérien, B correspond à l'effet de la date d'inoculation. Aucun effet notateur n'a été pris en compte pour les tests tavelure puisque toutes les notations ont été effectuées par le même notateur, ni pour le test feu bactérien puisque la note attribuée est une mesure physique qui ne dépend pas de la perception du notateur.

Les données des différents caractères de qualité du fruit ont été analysées en utilisant le modèle de variance suivant :

$$Y_{ijk} = \mu + B_i + \mathrm{N(B)}_{ik} + G_j + (B \times G)_{ij} + \varepsilon_{ijk}$$

dans lequel Y_{ijk} est le phénotype de la répétition k du génotype j dans le bloc i, μ la moyenne générale de la population, B l'effet de l'année i, N(B) l'effet du notateur k emboîté dans l'année i, G l'effet du génotype j, (B x G) l'interaction entre le génotype et l'année et ε_{ijk} l'effet résiduel.

Le modèle d'analyse de variance le plus adapté aux données (interactif ou additif) a été choisi en calculant le Bayesian Information Criterion (BIC) pour chaque test et en choisissant le modèle pour lequel le BIC était le plus faible. D'autre part, les valeurs génotypiques des variétés ont été estimées en utilisant le Best Linear Unbiased Predictors (BLUP), correspondant à une régression géno-phénotypique (Pinheiro and Bates 2000). Pour les données de qualité du fruit, les BLUP correspondant aux valeurs génotypiques moyennes des individus sur l'ensemble des années (BLUP_G), ceux correspondant à la valeur génotypique spécifique des individus pour une année donnée (BLUP_G2012, BLUP_G2013 et BLUP_G2014), ainsi que les BLUP des interactions génotype-année (BLUP_I2012, BLUP_I2013, BLUP_I2014) ont été calculés. Les héritabilités au sens large ont été calculées à partir des résultats des analyses de variance suivant l'une ou l'autre des formules suivantes, selon que le modèle de variance était additif ou interactif :

$$H^2 = \frac{\sigma_G^2}{[\sigma_G^2 + \left(\sigma_\varepsilon^2/n\right)]} \quad \text{ou} \quad H^2 = \frac{\sigma_G^2}{[\sigma_G^2 + \left(\sigma_{B\times G}^2/z\right) + \left(\sigma_\varepsilon^2/nz\right)]}$$

dans lequel H^2 est l'héritabilité au sens large, σ^2_G la variance génotypique, σ^2_ε la variance de la résiduelle, σ^2_{BxG} la variance de l'interaction entre le génotype et l'année de dégustation, n le nombre moyen de copies par génotype (pour une année ou un test donnés) et z le nombre d'années ou de blocs. La première formule s'applique aux modèles de variance additifs et la seconde aux modèles de variance interactifs.

Des corrélations (Pearson et Spearman) ont ensuite été calculées entre tous les caractères de qualité du fruit d'une part, sur la base des BLUP_G, et pour les tests de résistance à la tavelure d'autre part.

5.3.2. Recherche d'associations phénotype-génotype

5.3.2.1. Estimation de la structure et de l'apparentement

Comme indiqué dans le Chapitre 1, 2.3.2, il est important, lors de la réalisation d'une étude de génétique d'association, de se prévaloir de tous les effets qui pourraient augmenter le nombre de faux positifs. Deux effets qui génèrent ce type d'erreurs sont la structure et l'apparentement. Le package factoMiner de R (Lê, Josse, and Husson 2008) a été utilisé pour réaliser une ACP. Lors de ces analyses, seul un sous-ensemble des marqueurs de la puce, contentant environ 55k SNPs, a été utilisé. Ces marqueurs ont été choisis pour ne pas être en fort DL ($r^2 < 0,1$) grâce au logiciel plink (Purcell et al. 2007). La matrice centrée d'apparentement K a été calculée en utilisant le logiciel GEMMA sur le jeu de données total (Zhou and Stephens 2012).

5.3.2.2. Etude d'association

Comme indiqué en 5.2.2, les données manquantes présentes dans le jeu de données de la puce de génotypage 480k ont été imputées en utilisant le logiciel Beagle 4.0 (Browning and Browning 2007).

Avant de réaliser l'analyse d'association à proprement parler, le modèle à appliquer a été choisi en calculant le BIC selon la formule indiquée par Courtois et al. (2013) :

$$BIC = -2\ln(L) + k\ln(n)$$

dans laquelle L est égal à la valeur de la vraisemblance dans le modèle testé, k au nombre de paramètres estimés dans le modèle et n à la taille de l'échantillon. Deux modèles différents ont été testés : un modèle prenant en compte l'apparentement estimé avec le logiciel GEMMA (Zhou and Stephens 2012) et un prenant simultanément en compte la structure, estimée en réalisant une ACP, et l'apparentement en utilisant le logiciel GEMMA. Le modèle ayant la valeur de BIC la plus faible a été choisi comme modèle pour les analyses d'association.

Le modèle linéaire mixte univarié implémenté dans GEMMA a été utilisé selon le modèle suivant afin de détecter des associations entre les données génotypiques complètes et les BLUP pour les différents caractères étudiés séparément :

$$y = W\alpha + x\beta + u + \varepsilon$$

dans lequel y représente le vecteur des données phénotypiques, W une matrice de covariables (moyennes phénotypiques et matrice de structure), α l'effet des covariables, x le vecteur des génotypes, β l'effet du SNP (sous forme de dose allélique), u un vecteur d'effets aléatoires (dans lequel est prise en compte la matrice d'apparentement K) et ε le vecteur des résiduelles. Pour les analyses des données de qualité du fruit, les valeurs génotypiques des individus par année (BLUP annuels), les BLUP des individus sur les trois années (BLUP_G), et les BLUP des interactions génotype-année (BLUP_I) ont été utilisés. Compte-tenu du nombre de marqueurs utilisés pour les analyses (~ 239k), un seuil de significativité a été calculé selon la correction de Bonferroni pour un seuil de significativité global de α = 0,05. Une autre valeur seuil a été estimée en utilisant les résultats du logiciel GEC (Li et al. 2012) qui estime le nombre de tests, et donc de SNPs, indépendants lors des analyses (~ 207k). Au vu des résultats d'association, les deux seuils (6,7 et 6,6) ont été jugés trop stringents puisque très peu d'analyses montraient des SNPs avec des valeurs significatives. Un seuil de *p*-value = 1.10^{-5} a donc été choisi pour déclarer les associations significatives, malgré le risque de détecter davantage de faux positifs. D'autres SNPs, pour lesquels la *p*-value est supérieure à 10^{-5}, présentaient également des profils intéressants. Certaines régions génomiques abritaient en effet des « pics » formés par plusieurs SNPs présentant des *p*-values décroissantes puis croissantes (avec un SNP central ayant une valeur minimale de *p*-value) en fonction de la distance qui les sépare du SNP le plus significatif. Certaines de ces régions génomiques ont donc été étudiées. L'héritabilité au sens strict (h^2) a été calculée en utilisant la formule présentée dans le Chapitre 1, 1.2.3.1.

5.3.2.3. Effets des SNPs ayant une *p*-value significative

Pour certains des SNPs présentant des *p*-values significatives pour un caractère donné, une analyse de variance a été réalisée sous R selon le modèle suivant :

$$Y = \mu + M_1 + M_{...} + M_n + \varepsilon$$

dans lequel Y représente le vecteur des données phénotypiques de ce caractère, μ la moyenne générale de la population, M_n le vecteur des génotypes (AA, AB, BB) au marqueur SNP retenu sur la base de sa *p*-value et ε à l'effet résiduel. Le R^2, correspondant à la part de variation phénotypique expliquée, a été calculé pour tous les SNPs significatifs d'un test d'association, soit de façon globale soit de façon individuelle. Il est à noter que ces calculs de R^2 ont été réalisés avec un modèle statistique ne prenant en compte ni les effets de structure ni d'apparentement, à la différence du modèle linéaire mixte utilisé précédemment. Ils ne coïncident donc pas exactement avec les valeurs de *p*-value sur la base desquelles les SNPs ont été retenus. Par ailleurs, le modèle utilisé ici considère l'effet de chaque SNP selon les 3 classes génotypiques possibles (AA, AB, BB) et non selon le nombre de doses alléliques (0, 1, 2). Un test de comparaison de moyennes de Newman-Keuls a donc été réalisé pour mieux décrire les différences entre les moyennes par classe génotypique (AA, AB et BB) pour ces SNPs et pour évaluer d'éventuels effets de dominance en comparant la moyenne AB à la moyenne AA + BB. Les valeurs de ces effets ont été obtenues en considérant que la différence entre les moyennes des deux homozygotes est égale à 2a et que la valeur de la moyenne de l'hétérozygote est égale à a + d.

5.3.2.4. Recherche de gènes candidats

Lorsque le contig sur lequel un SNP détecté lors de l'analyse d'association avait une longueur inférieure à 20 kb, des fenêtres d'environ 20 kb (10 kb de part et d'autre du marqueur) autour du marqueur ont été examinées et les gènes présents dans ces fenêtres étudiés. Pour les cas où le contig était plus long que 20 kb, l'ensemble du contig a été considéré comme étant la fenêtre à examiner. Les séquences des gènes ont été obtenues à partir du site http://www.rosaceae.org/data/download, et des megablast ont été réalisés sur la banque de séquences nucléotidiques non redondante en gardant les paramètres par défaut. Les prédictions de gènes réalisées sur la version 1 du génome du pommier datent en effet de 2010, et des gènes identifiés comme n'ayant pas de fonction prédite peuvent à ce jour être alignés avec des gènes identifiés récemment et présentant une fonction putative.

5.3.2.5. Version du génome du pommier et carte physique disponible

La première version du génome du pommier publiée en 2010 par Velasco et al. (Fondation Edmund Mach – FEM, Italie) a été réalisée à partir de données de séquençage de la variété « Golden Delicious », très hétérozygote. L'assemblage du génome a donc été fastidieux et a créé des regroupements erronés de contigs et scaffolds. Des estimations réalisées par FEM après la publication de la version 1 du génome ont révélé que la proportion de contigs mal placés pouvait atteindre 25% sur certains groupes de liaison. Après une version 2 du génome très légèrement remaniée et nécessaire à la publication d'une puce de génotypage de densité intermédiaire (puce 20k SNPs, Bianco et al., 2014), une version 3 a été produite fin 2014 par FEM dans l'optique d'améliorer le placement des contigs. Les données d'une carte génétique dense réalisée grâce aux données de génotypage issus de cette puce 20k ont en particulier été utilisées. Par ailleurs, certains contigs ont été regroupés ou fusionnés sur la base de leur similarité de séquences et d'autres contigs (ou scaffolds) ont été dissociés.

Dans le cadre du projet européen FruitBreedomics, FEM a transmis à l'équipe ResPom un ensemble d'informations sur la position des SNPs sur les contigs, la position et l'orientation des contigs dans les scaffolds, la position des scaffolds sur les groupes de liaison (mais sans les orientations dans certains cas du fait du manque de marqueurs génétiques permettant de les orienter). Par ailleurs, un nombre conséquent de contigs ou de scaffolds n'avaient pas pu être ancrés sur le génome (pas de données de marquage génétique disponibles). A partir de l'ensemble de ces informations, l'équipe ResPom (H. Muranty) a généré un fichier donnant la position physique des SNPs sur les groupes de liaison pour tous les SNPs appartenant à des contigs ancrés. Une position hypothétique sur un groupe de liaison « fantôme » (baptisé « groupe de liaison 0 ») a aussi été donnée pour les SNPs appartenant à des contigs non ancrés (mis bout à bout) de manière à disposer d'une position physique (même erronée). Cependant dans le cadre de cette thèse, ces derniers SNPs n'ont pas été considérés.

La puce 8k SNPs (Chagne, Crowhurst, et al. 2012) a été designée à partir de la version 1 du génome et la puce 480k SNP à partir de la version 3. Les positions physiques de la plupart des contigs ayant été modifiées lors du passage à la version 3, la position des contigs étudiés suite aux études d'association (Chapitre 5) a été comparée à celle de la version 1, notamment dans le cas de co-localisations avec des QTLs.

Chapitre 3

Bases génétiques de la différenciation entre pommes à cidre et à couteau

1. Introduction

L'étude des processus évolutifs ayant eu lieu pendant la diversification des variétés présente un intérêt fondamental pour la compréhension de l'évolution des espèces et pour diverses applications en matière de sélection chez des espèces cultivées. Cette étude peut notamment permettre de développer les connaissances sur les bases génétiques de caractères d'intérêt agronomique et sur les moyens de préserver les ressources génétiques. Une étude récente sur l'origine du pommier cultivé a montré que, contrairement aux attentes, les pommiers à cidre ne présentaient pas plus d'introgression de séquences génomiques depuis le pommier sauvage *Malus sylvestris* que les pommiers à couteau (Cornille et al. 2012). Cette hypothèse reposait sur le fait que les pommes à cidre ressemblent fortement aux pommes sauvages, étant généralement de plus petite taille et amères au goût. Le scénario d'une participation plus importante du pommier sauvage européen dans l'émergence du pommier à cidre n'étant pas retenu, les deux sortes de pommes dérivent probablement du même pommier ancestral et ont donc dû subir des pressions de sélection différentes. Au cours de cette thèse, il a donc été proposé d'étudier deux core collections de pommes à couteau et pommes à cidre dans le but d'identifier les régions génomiques qui participent aux différences phénotypiques entre les deux sortes de pommes, et de rechercher les signatures de sélection dans les deux génomes. Ce travail, mené grâce à l'appui du laboratoire « Ecologie, Systématique et Evolution » (ESE) de l'université d'Orsay (A. Branca et T. Giraud) a fait l'objet d'un article récemment accepté dans la revue *Evolutionary Applications*. L'article est intégralement présenté ci-dessous, puis suppléé par des analyses complémentaires basées sur le spectre de fréquences de sites dont les résultats sont ensuite discutés.

2. Genomic basis of the differences between cider and dessert apple varieties

Diane Leforestier[a], Elisa Ravon[b], Hélène Muranty[b], Amandine Cornille[c,d], Christophe Lemaire[a], Tatiana Giraud[c,d*], Charles-Eric Durel[b*], Antoine Branca[c,d*]

[a] Université d'Angers, UMR 1345 Institut de Recherche en Horticulture et Semences, F49045 Angers, France.

[b] INRA, UMR 1345 Institut de Recherche en Horticulture et Semences, SFR 4207 QUASAV, 49071 Beaucouzé, France.

[c] ESE, Université Paris-Sud, 91405 Orsay France

[d] ESE, CNRS, 91405 Orsay Cedex

[*] co-senior authors

Abstract

Unravelling the genomic processes at play during variety diversification is of fundamental interest for understanding evolution, but also of applied interest in crop science. It can indeed provide knowledge on the genetic bases of traits for crop improvement and germplasm diversity management. Apple is one of the most important fruit crops in temperate regions, having both great economic and cultural values. Sweet dessert apples are used for direct consumption while bitter cider apples are used to produce cider. Several important traits are known to differentiate the two variety types, in particular fruit size, biennial *versus* annual fruit bearing and bitterness, caused by a higher content in polyphenols. Here, we used an Illumina 8K SNP chip on two core collections, of 48 dessert and 48 cider apples, respectively, for identifying genomic regions responsible for the differences between cider and dessert apples. The genome-wide level of genetic differentiation between cider and dessert apples was low, although 17 candidate regions showed signatures of divergent selection, displaying either outlier F_{ST} values or significant association with phenotypic traits (bitter *versus* sweet fruits). These candidate regions encompassed 420 genes involved in a variety of functions and metabolic pathways, including several colocalizations with QTLs for polyphenol compounds.

Keywords

Malus domestica, BayeScan, outlier, F_{ST}, linkage disequilibrium, genome-wide association

2.1. Introduction

Domestication and variety diversification have been models for studying the mechanisms underlying adaptation since Darwin (1856), being the result of a strong and recent selection by humans for desired traits in organisms used as food (Meyer, DuVal, and Jensen 2012; Larson and Burger 2013; McTavish et al. 2013), ornaments (Yuan et al. 2014), pets (Axelsson et al. 2013) or for their metabolic abilities (Douglas and Klaenhammer 2010). Dissecting the genomic changes occurring during domestication and variety diversification has thus a fundamental importance for our understanding of evolutionary processes, in addition to applied interests for improving the desired traits in domesticated organisms and managing the germplasm diversity. Studying the footprints of adaptation in genomes may indeed allow to identify the important traits or metabolic pathways that were under selection during domestication and variety diversification, as well as the genetic bases of these traits (Wang et al. 1999; Whitt et al. 2002; Palaisa et al. 2003; Gallavotti et al. 2004; Wang et al. 2005; Yamasaki et al. 2005; Walsh 2008). Identifying the genomic regions involved may accelerate further improvement of traits controlling agricultural productivity and performance, such as yield, organoleptic or nutritional quality, and resistance to biotic and abiotic stresses, using marker-assisted selection (Soller 1994;

Collard and Mackill 2008; Prada 2009). It may also help conservation management programs aiming at maintaining important functional biodiversity in core collections as well as in wild relatives of crop species.

The cultivated apple tree (*Malus domestica* Borkh.) is one of the most important fruit crops in temperate regions, with great economic and cultural values (Juniper and Mabberley 2006). Dessert apples are popular because of their taste, nutritional properties, storability and convenience of use. The fruits of the specific varieties used to produce cider are smaller and bitter, as are those from crabapples, i.e., the fruits of the wild apple species. The bitterness is due to a high content in polyphenols (Sanoner et al. 1999). Not all cider cultivars are however extremely bitter (Pereira-Lorenzo, Ramos-Cabrer, and Fischer 2009). Cider apples are also known for their fibrous structure, which allows longer storage (Lea and Piggott 2003; Campo del et al. 2005). In addition, cider apples more often display biennial bearing (Dapena, Minarro, and Blazquez 2005), *i.e.*, with crop occurring only every two years. Finally, cider apples are more susceptible than dessert apples to fire blight, a disease caused by the bacteria *Erwinia amylovora* (Paulin et al. 1988; Lespinasse and Paulin 1990). Thousands of apple cultivars have been documented (Morgan, Richards, and Dowle 2002), although only a few now dominate the market. Surprisingly, the history of apple domestication has just begun to be unraveled (Cornille et al. 2014). Genetic analyses have revealed a Central Asian origin of cultivated apple, with an initial divergence from the wild species *Malus sieversii*, together with an unexpectedly large secondary contribution through introgression from the European wild species *Malus sylvestris* (Velasco et al. 2010; Cornille et al. 2012). In contrast to expectations, cider cultivars did not appear the most introgressed by wild species based on microsatellites (Cornille et al. 2012). This suggests either a recent selection in the cider varieties for traits favorable for apple-based beverages from the standing genetic variation in the domesticated gene pool, or the introgression of only few genes from crabapples into the cider varieties. However, cider beverage has been produced for centuries in Western Europe especially by the Celts using native crabapples even before the invasion of the Romans who brought the domesticated apples. Much effort has been devoted since the 17th century in Europe to generate cider apple cultivars with high contents in sugar and polyphenols for producing high quality cider (Morgan, Richards, and Dowle 2002).

Although some *M. sieversii* individuals produce large apples, the variability in fruit size and color is wide. The selection by humans in cultivated apples targeted many phenotypic traits, including among others the number of fruits, their size, color, shape, flavor, taste, texture, storage capacity, harvesting ease, juvenile phase length and disease resistance (Janick 2005). QTL mapping has been used to dissect the genetic architecture of several desired traits, through crosses between cultivars (Calenge et al. 2004; Segura, Cilas, and Costes 2008; Celton et al. 2011; Guitton et al. 2012; Longhi et al. 2012; Verdu et al. 2014). However, the footprints of selection have been little studied so far in apples compared to annual crops (Yamasaki et al. 2005; Camus-Kulandaivelu et al. 2008). The recently

released 'Golden Delicious' genome sequence (Velasco et al. 2010) and the availability of medium-density genotyping tools (Chagne, Crowhurst, et al. 2012) have made it possible to generate population-scale data for investigating genome-wide patterns of selection.

In the present study, we set out to identify genomic regions under divergent selection between cider and dessert apples using two core collections, one of each variety type (N=48 each), and 3,704 SNP markers. First, we analyzed the population genetic structure in our sample to assess the differentiation between dessert and cider apple varieties using a much higher number of markers than in a previous study (Cornille et al. 2012). We also investigated the extent of linkage disequilibrium (LD) as a function of genomic distance within the genome in order to infer the expected maximal distance between the causal variation and the markers displaying association with the phenotype. We then looked at F_{ST} statistics for identifying outlier loci that would differentiate cider and dessert varieties significantly more than the average genomic background. Finally, a genome-wide association analysis was performed, taking into account genetic structure and kinship, contrasting the 1) cider *versus* dessert variety types or 2) high *versus* low bitterness cultivars. Altogether, these analyses aimed at localizing the genomic regions that have been under divergent selection and responsible for the phenotypic differences between cider and dessert apples. We then examined in these regions the putative functions of genes to find candidates that have potentially undergone differential changes during the divergence between cider and dessert apples. Recent selection programs on cider apples aim at improving yield, regularity of production, resistance to pests and pathogens, while maintaining their specific technologic characteristics (*e.g.* high content in polyphenols). The identification of the genomic regions responsible for the differences between dessert and cider variety types could therefore be of great use for instance in a marker-assisted selection approach trying to select new cider varieties combining a higher content in polyphenols with the agronomic performances of dessert apples such as regular annual bearing, higher yield and fruit size.

2.2. Material and methods

2.2.1. Plant material

The two apple core collections used in this study had been previously constituted by choosing the individuals that maximized the genetic diversity based on a set of 24 microsatellite markers in the INRA Angers germplasm collection of dessert and cider apple cultivars. Shortly, the core collections were built by retaining individuals from larger sets of apple accessions (737 and 188 for dessert and cider apples, respectively) using the 'Maximum Length Subtree' option of the DARwin software (Perrier,

Flori, and Bonnot 2003; Perrier and Jacquemoud-Collet 2006). The two core collections included 48 dessert and 48 cider apple cultivars respectively (Annexe 3). Reflecting the content of the INRA germplasm collection, both core collections mainly include old (generated before the 1950's) French apple cultivars, some of them being clones of cultivars grown in other European countries under different names. Because Western Europe has been the main place where the selection of dessert and cider apples has taken place (Morgan, Richards, and Dowle 2002), the core collections we studied should be quite representative of the selection history of dessert and cider apples.

2.2.2. SNP arrays

Genomic DNA was extracted from leaves of the 96 individuals using the NucleoSpin® Plant II kit (Macherey-Nagel GmbH and Co KG, Düren, Germany). Because apple leaves are full of polysaccharides and phenols that contaminate the extracted DNA and may prevent hybridization on the array, DNA samples were purified as follows: 0.1 volume of sodium acetate (final concentration 0.3 M), 2.5 volumes of cold 100% ethanol and 1 µL of glycogen were added, the tubes were centrifuged at 13,000 g for 30 minutes, the supernatant was discarded, 200 µL of 70% ethanol were added, the tubes were centrifuged at 13,000 g for 10 minutes, the supernatant was discarded, the tubes were air-dried overnight and the DNA was resuspended in the appropriate volume of water. DNA samples were then checked for quality using Nanodrop 1000, quantified using PicoGreen® and processed onto the International RosBREED SNP Consortium (IRSC) apple 8K SNP array v1 (Chagne, Crowhurst, et al. 2012) following the Illumina® protocol.

2.2.3. SNP filtering

SNPs were filtered using the Genotyping Module (version 1.8.4) of the Illumina® GenomeStudio software (Illumina Inc.). A visual inspection of each SNP was performed and SNPs exhibiting a good genotypic clustering in distinct spots were kept. Paralogous SNPs were removed by performing BLAST onto the apple genome and removing probes having two equally good best hits onto the reference genome. This step was necessary for avoiding potential paralogy, due to the whole genome duplication having occurred in the apple evolutionary history (Velasco et al. 2010). There were a few missing data in the dataset obtained from GenomeStudio, we therefore used fastPHASE 1.2 (Scheet and Stephens 2006) with the default parameters, and we indicated whether an individual belonged to the cider subgroup or to the dessert one to impute the missing data and to phase the SNPs belonging to a given linkage group (LG). Because the core collections were designed to maximize the genetic diversity and

because SNPs for the 8K array were chosen among the most polymorphic markers in 27 dessert apple genomes (Chagne, Crowhurst, et al. 2012), the allelic frequencies obtained may be biased compared to the full genetic pools of dessert and cider apples. Therefore, we excluded analyses based on the site frequency spectrum and focused only on analyses less sensitive to such biases.

2.2.4. Estimation of Linkage Disequilibrium

The levels of linkage disequilibrium were estimated using the r^2 parameter between all pairwise comparisons using the Haploview 4.2 software (Barrett et al. 2005) and a minor allele frequency (MAF) cutoff of 0.01. A first analysis was run without taking into account the structure and kinship in the collections; the levels of linkage disequilibrium were then corrected for population structure (see below) and kinship using the R package LDcorSV (Mangin et al. 2012). The kinship matrix, reflecting the degree of genetic covariance among individuals, was calculated with the Cocoa 1.1 software (Maenhout, De Baets, and Haesaert 2009).

2.2.5. Analysis of population structure

The ADMIXTURE 1.23 software (Alexander, Novembre, and Lange 2009) was used to investigate the genetic population structure in the dataset. The number of genetic clusters K was assessed using values ranging from 1 to 10 and we chose the number of clusters for which the cross-validation error was the lowest. The cross-validation procedure masks one fifth of the genotypes (five runs altogether) and calculates estimates for these genotypes. Each genotype is then predicted and the software calculates a prediction error across all masked genotypes. The Q matrix, *i.e.* the posterior probabilities for each individual to belong to a given cluster, outputted by ADMIXTURE 1.23 was used for the genotype-phenotype association analysis.

2.2.6. Differentiation between cider and dessert apples - Detection of outlier loci

Pairwise single locus F_{ST} between the two core collections was calculated using either GENETIX 4.05 (Belkhir et al. 1996-2004) or BayeScan 2.1 (Foll and Gaggiotti 2008). The Bayesian method implemented in the latter (Beaumont and Balding 2004) was run to detect outlier loci using the following

parameters: after 20 pilot runs of 50,000 iterations and an additional burn-in of 500,000 iterations, we used 3,000,000 iterations (thinning interval of 50 and sample size of 50,000).

2.2.7. Phenotype-genotype association

A genome-wide association study (GWAS) was run using the univariate linear mixed model (LMM) implemented in GEMMA (Zhou and Stephens 2012), taking into account the centered kinship matrix (K) calculated in GEMMA and the Q matrix from ADMIXTURE. A first analysis was performed on the two core collections by giving cider cultivars a score of 1 and dessert cultivars a score of 0. However, because not all cider cultivars are bitter, a second analysis was performed, this time not considering the cider *versus* dessert cultivars classification, but instead the bitterness of the cider apple cultivars, as recorded in the literature (Boré and Fleckinger 1997): bitter cider cultivars were given a score of 1 while sweet cider cultivars and dessert cultivars were given a score of 0. Both binary situations were treated as quantitative traits, as the linear mixed model is recognized as a robust approximation of a generalized linear model (Zhou, Carbonetto, and Stephens 2013). Markers were considered significantly associated with the phenotype for p-value $\leq 10^{-3}$. P-values obtained from GEMMA were used in R environment using the qqman package to generate a Manhattan plot (Turner 2014).

2.2.8. Identification of candidate genes

The online apple genome browser hosted on http://www.rosaceae.org/, containing the gene model predictions made on the apple genome sequence, was used to investigate the putative functions of genes present in the genomic regions detected in the tests above. The Blast2Go 3.0 software (Conesa et al. 2005) was used to perform BLASTX on these sequences with a maximum Blast ExpectValue of 10^{-3}. After gene ontology (GO) functional annotation, the KEGG tools were used to visualize the corresponding metabolic pathways. A BLASTN was run and its results were used as inputs in Blast2GO 3.0 to retrieve GO annotations for the entire gene set of the apple genome. The regions of interest were then tested for enrichment of particular gene functions.

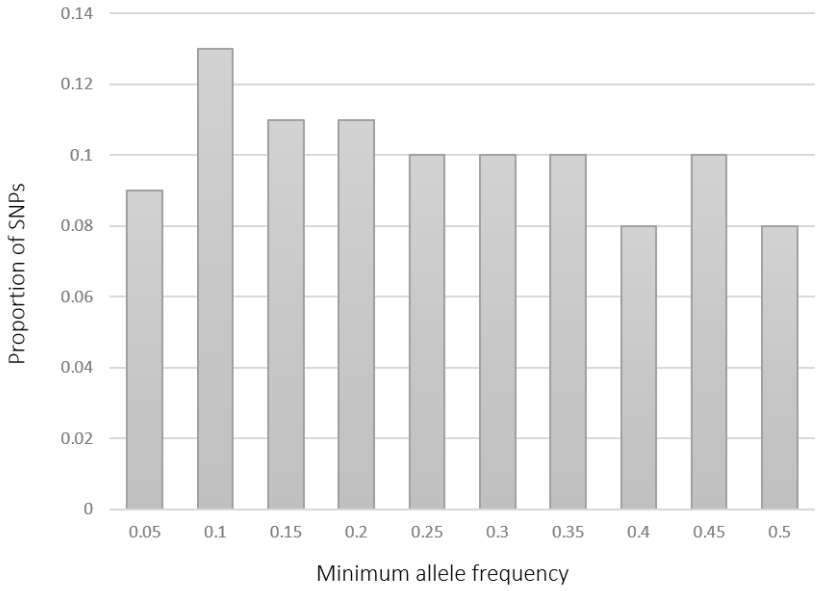


Figure 23 : Allele frequency spectrum of the 3,704 SNP markers of the array

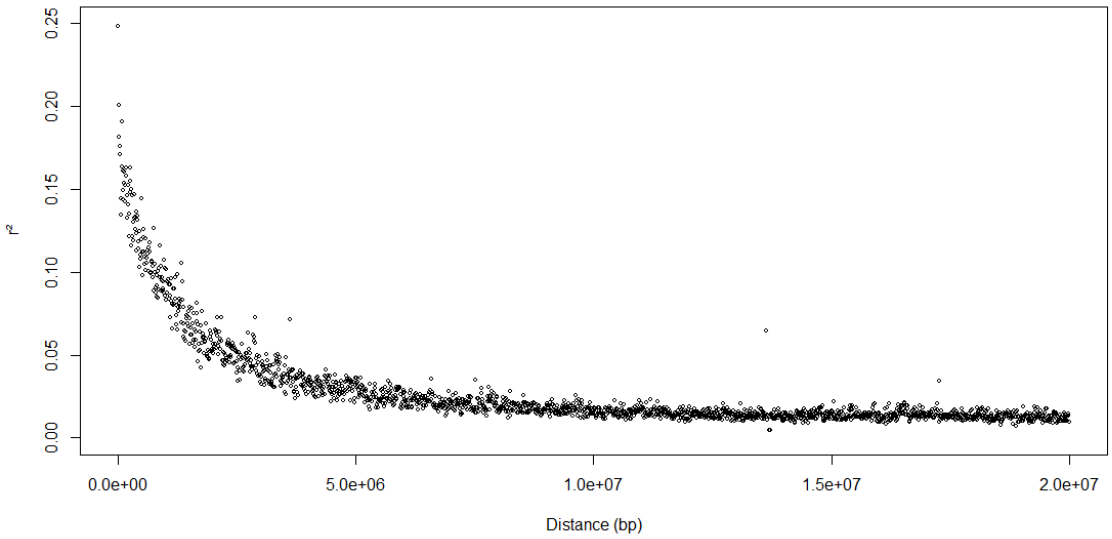


Figure 24 : Decay of average linkage disequilibrium (measured as r^2) versus physical distance in increments of 10,000 bp

both core collections, cider and dessert, were included in the analysis since no difference was observed when correcting for structure and/or kinship

2.3. Results

2.3.1. SNP genotyping

After visually screening the 7,867 SNPs of the IRSC apple 8K SNP array v1 on GenomeStudio, a set of 4,234 polymorphic SNPs evenly spread across the apple genome was obtained; after removing potential paralogous SNPs, the number of markers was reduced down to 3,704. The number of markers per linkage group was approximately proportional to their length. The average distance between two adjacent SNPs was 140 kb, with the maximum distance separating markers ranging from 1.26 Mb on LG17 to 4.25 Mb on LG15. The distribution of the SNP minor allele frequencies (MAF) was quite uniform across the different possible MAF values (Figure 23) whether considering the cider or the dessert cultivars. Overall, few data were missing in the dataset, with 2,981 markers having no missing data at all and the maximum percentage of missing data being 5.2% and 10.2% per marker and per individual, respectively. This made the inferences using fastPHASE 1.2 highly reliable.

2.3.2. Estimation of linkage disequilibrium

The nonlinear regression model used to analyze the decay of linkage disequilibrium (LD) with the physical distance showed that the squared allele correlation parameter r^2 decayed below 0.2 within 100 kb (Figure 24). When analyzed separately, the cider and the dessert core collections showed very similar behaviors. The results obtained on the whole dataset when taking into account kinship or/and population structure were very similar too. We therefore assumed that loci distant from more than 100 kb were not in LD and considered windows of 100 kb on both sides of outlier SNPs for finding candidate genes possibly evolving under divergent selection.

2.3.3. Differentiation between cider and dessert apples - Analysis of population structure

Pairwise F_{ST} between cider and dessert apples ranged from 0 to 0.24, with a mean value of 0.014, confirming the weak differentiation between the two core collections. ADMIXTURE analyses revealed a minimum value of the cross-validation error for K=2. Only a quarter of the individuals actually showed a clear assignment (membership probability > 0.9) to any cluster, supporting the lack

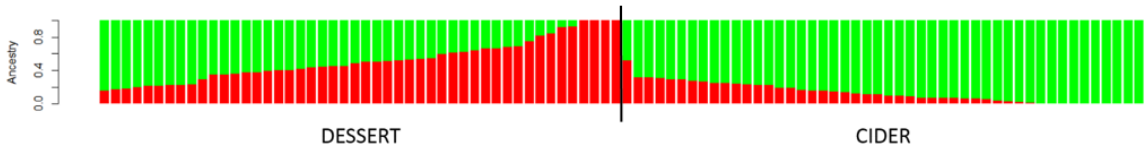


Figure 25 : Population structure of 96 apple cultivars from the cider and dessert genetic pools.
membership probabilities were obtained with ADMIXTURE for K=2
the bar plot, generated using the qqman package in R, shows each individual as a vertical bar

Tableau 4 : SNPs showing significant levels of F_{ST} detected by BayeScan 2.1

SNP Name	LG[a]	Position	F_{ST}
GDsnp01132	8	11,328,418	0.23
RosBREEDSNP_SNP_CA_29926704_Lg15_RosCOS1232_MAF50_MDP0000283141_exon1	15	26,329,550	0.23
RosBREEDSNP_SNP_AG_33667246_Lg15_01897_MAF40_151341_exon1	15	29,068,827	0.24
RosBREEDSNP_SNP_CT_10901071_Lg17_00918_MAF10_1668766_exon3	17	10,334,128	0.19
RosBREEDSNP_SNP_CT_10898449_Lg17_00918_MAF10_466062_exon6	17	10,336,750	0.19

[a]: Linkage Group

Tableau 5 : SNPs showing significant association with the cider/dessert or bitter/sweet phenotypes when taking into account structure and kinship between individuals using GEMMA

SNP Name	LG[a]	Position	p-value
RosBREEDSNP_SNP_CT_22024068_Lg5_RosCOS3072_MAF30_MDP0000753788_exon2[b]	5	19238624	3.83×10^{-4}
RosBREEDSNP_SNP_TC_15251985_Lg8_00354_MAF10_753213_exon1[b]	8	13053086	8.98×10^{-5}
RosBREEDSNP_SNP_TG_23835076_Lg8_RosCOS3331_MAF40_488673_exon1[b]	8	19848379	1.99×10^{-4}
RosBREEDSNP_SNP_GA_33077622_Lg9_01200_MAF20_MDP0000613052_exon1[b]	9	29701351	2.24×10^{-4}
RosBREEDSNP_SNP_GA_1240623_Lg12_RosCOS3293_MAF40_1686868_exon1[b]	12	1033191	4.12×10^{-4}
RosBREEDSNP_SNP_AG_27056933_Lg15_02084_MAF30_1677692_exon1[b]	15	23859694	2.06×10^{-4}
RosBREEDSNP_SNP_AG_32748739_Lg1_RosCOS2753_MAF10_520680_exon1[c]	1	26153648	1.86×10^{-5}
RosBREEDSNP_SNP_AC_33325153_Lg1_01951_MAF10_132337_exon1[c]	1	26730062	2.10×10^{-4}
GDsnp01850[c]	15	23124410	7.20×10^{-4}
RosBREEDSNP_SNP_AC_1452699_Lg16_MDP0000303483_MAF50_MDP0000303483_exon2[c]	16	1452699	2.78×10^{-4}
RosBREEDSNP_SNP_CT_8827345_Lg17_01842_MAF30_MDP0000891106_exon4[c]	17	8427545	7.87×10^{-5}
RosBREEDSNP_SNP_CT_17294445_Lg17_01964_MAF10_1662340_exon9[c]	17	15881764	7.14×10^{-4}

[a]: Linkage Group

[b]: SNP associated with the cider/dessert phenotype

[c]: SNP associated with the bitter/sweet phenotype

of further structure in the dataset. The Q-matrix for K=2 (Figure 25) confirms the lack of strong differentiation according to the cider/dessert classification, even using genome-wide markers.

2.3.4. Detection of F_{ST} outlier loci and genotype-phenotype associations

Out of the 3,704 SNPs tested for their probability to have been under divergent selection using Bayescan 2.1, five exhibited significant genetic differentiation. These five outlier SNPs were located as follows: one SNP on LG08 at 11.32 Mb, two SNPs on LG15 at 26.38 Mb and 29.20 Mb and two SNPs on LG17 at 10.30 Mb (Table 4). The GWAS testing SNP association with cider/dessert variety types revealed six SNPs with significant P-values (*i.e.*, -Log10 P-value ≥ 3). These markers were located as follows: one SNP on LG05 at 19.23 Mb, two SNPs on LG08 at 13.05 Mb and 19.85 Mb, one SNP on LG09 at 29.70 Mb, one SNP on LG12 at 1.03 Mb, and one SNP on LG15 at 23.86 Mb (Table 5 and Figure 26). The bitter/sweet trait was found significantly associated with six SNPs located as follows: 2 SNPs on LG01 respectively located at 2.61 Mb and 2.67 Mb, one SNP on LG15 at 23.12 Mb, one SNP on LG16 at 1.45 Mb and two SNPs on LG17 respectively located at 8.42 Mb and 15.88 Mb (Table 5 and Figure 26).

2.3.5. Genes around candidate SNPs associated with phenotypes

We looked at the gene predictions available on the first version of the genome of apple within 200 kb around the seventeen SNPs detected above as putatively under diversifying selection. In the regions containing the five F_{ST} outlier loci detected by BayeScan, 85 predicted genes were found, whose main classes of putative functions are reported in Supplementary Data (Annexe 4). In the 12 regions carrying the markers found to be associated with the cider/dessert or bitter/sweet phenotypes, 179 and 156 predicted genes were found, respectively, whose main classes of putative functions are shown in Supplementary Data. Among these genes, the most represented biological processes were: 1) amino acid metabolism and starch and sugar metabolism for the F_{ST} outliers, 2) nucleotide metabolism and glycerolipid metabolism for the variety type associated regions and 3) purine metabolism and thiamine metabolism for the bitterness associated regions. The enrichment test made using the entire predicted gene set as reference did not yield any significant result.

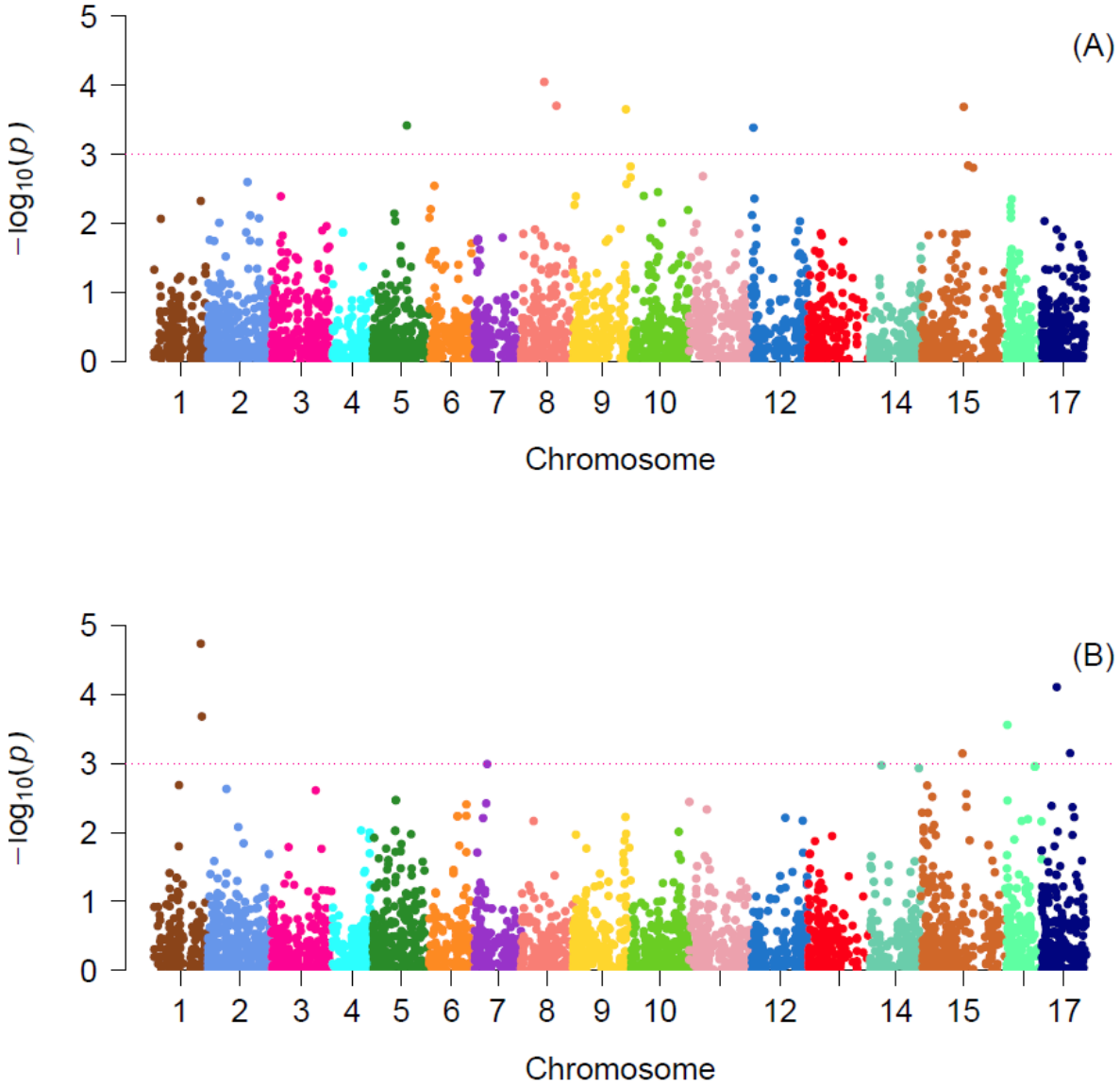


Figure 26 : Manhattan plot of the GWAS testing for association between genotypes and the cider/dessert (A) or bitter/sweet phenotype (B)

the –log10 of the P-value of 3704 SNPs after correction for structure and kinship is plotted against the physical position ; SNPs above the blue line are those exhibiting significant P-values and thus associated with the cider/dessert or bitter/sweet phenotype

2.4. Discussion

2.4.1. Possible biases due to sample and marker choices

We used in this study core collections that maximize the genetic diversity present in larger initial collections, which may have generated biases in allelic frequencies. However, F_{ST} outliers and GWAS methods should be robust to such biases, and even conservative. Indeed, core collections balance the initial extreme allelic frequencies, so that association should be valid across even more diverse genotypes in order to be significant in core collections. The use of core collections instead of random sampling may in addition have led to an underestimation of linkage disequilibria. Indeed, the increased distances between accessions within a core collection reflect an increased number of generations from the most recent common ancestor, and thus a higher number of crossing-overs between linked loci (Nordborg and Tavare 2002). The LD values in the core collection are however again conservative and are actually the appropriate estimates to consider for the definition of the windows size around the significant SNPs in the core collections.

Possible ascertainment biases in the SNP array design result from the choice of the markers among the most polymorphic SNPs based on genome re-sequencing of 27 dessert apple cultivars (Chagne, Crowhurst, et al. 2012). The direct consequence is a more uniform distribution of the MAF spectrum than generally observed for resequencing data (Pe'er et al. 2006). This SNP ascertainment bias most probably led to overestimating the r^2 values (Nielsen and Signorovitch 2003; Nielsen 2004; Lachance and Tishkoff 2013) and thus the LD extent. In the end, the combined impact of the core collection sampling and the SNP ascertainment bias on the LD estimation is difficult to assess. In addition, SNPs exhibiting contrasted frequency in the dessert and cider apple pools may have been discarded from the 8K apple array, even if they had a higher frequency in the cider apple gene pool, thus restricting the chance of detecting the corresponding genomic regions. These ascertainment biases however are again conservative: they may have led us to miss some genomic regions involved in cider *versus* dessert cultivars, but should not have yielded false positives. The regions detected here should therefore be considered as interesting candidates, but not an exhaustive list.

2.4.2. Low level of genomic differentiation between cider and dessert variety types

A previous study had reported a lack of population genetic structure between cider and dessert apples, using only a couple of dozen of microsatellite markers (Cornille et al. 2012). Our results confirm this result using a much higher number of markers of a different type (*i.e.*, SNPs instead of SSR) along the genome, with no clear assignment of most of the different cultivars to either one or the other of the two inferred clusters according to their variety type. The low mean value of F_{ST} between cider and dessert apples (0.014 in our study) also supports the lack of genome-wide differentiation and is consistent with the mean F_{ST} value of 0.02 found by Cornille et al. 2012. Actually, some cultivars, discarded from the present study, are known to be used for both cider and dessert (*e.g.*, Bagué Petit, Raccroupi, Cazo Jaune), which means the phenotypic classification in cider and dessert apples is not morphologically clear-cut either.

2.4.3. Long distance LD in the cultivated apple

The r^2 was found here to decrease below 0.2 within 100 kb. In previous studies on apples, r^2 was found decaying below 0.2 within 500kb in a population of 7 full-sib families genotyped with 2,500 SNPs (Kumar et al. 2012) and within 1 cM (corresponding to approximately 500 kb considering that the apple genome is 750 Mb and that the genetic map is 1,500 cM long) in a collection of 132 apple cultivars genotyped with 238 SNPs (Micheletti et al. 2008). Such discrepancies with our study may be explained by a sampling of siblings in the former study, therefore implying fewer recombination events than in a core collection encompassing a high diversity and several generations between individuals. In the latter study, a fewer number of SNPs, not spanning the entire genome, is also an explanation for a larger range of LD.

LD has also been studied in other *Rosaceae* crops like *Prunus persica*, where r^2 reached 0.1 within 1,200 kb in an Oriental peach germplasm (Li et al. 2013), and *Pyrus pyrifolia*, where r^2 fell below 0.2 at approximately 1,800 kb in a population of old and modern cultivars, considering the pear genome is 600 Mb and 1,100 cM long (Iwata et al. 2013). Studies performed on other allogamous tree species showed lower values of distances above which the LD decayed below 0.2: 200 bp in *Populus tremula* (Ingvarsson 2005) and approximately 2 kb in *Pinus taeda* L. (Brown et al. 2004). These levels of LD appear low compared to our results, probably because the studies were conducted on wild populations of forest trees, in which a much higher number of recombination events probably occurred since the

last population bottleneck. In addition, the rather high average distance between our markers may have led to miss some occurrences of short-distance LD.

2.4.4. Differentiated genomic regions between cider and dessert apples

We identified here a total of 17 regions potentially bearing genes responsible for phenotypic differences between cider and dessert apples. Five of these regions harbored F_{ST} outlier loci that exhibited high differentiation levels between cider and dessert cultivars while the other twelve showed significant associations between the genotypic information and the variety type or the bitter trait while accounting for structure and kinship. According to the results on LD decay, 200 kb-windows around the significant SNPs were investigated. The enrichment test performed on the three set of genes around significant SNPs, *i.e.*, F_{ST} outliers and the two association analyses results, did not detect any particularly over-represented pathway. No genes known to be involved in the traits differentiating cider and dessert cultivars, such as the polyphenol pathway, were identified around the F_{ST} outliers located on LG08 and LG15. Two genes having high sequence similarity with UDP-glycosyltransferases were found around the two outlier markers on LG17. These genes can play a role in the synthesis pathways of several polyphenol compounds such as flavonoids or anthocyanidins, as exemplified by the *MdPT1* gene (Jugdé et al. 2008) involved in the glycosylation of phloretin into phlorizin, a major dihydrochalcone of apple known to have a bitter taste that may contribute to the peculiar flavor of cider (Whiting and Coggins 1975).

Regarding the results of the association between the genotypic information and the variety type, the two SNPs located on LG08 colocalized with QTLs linked to biennial bearing and yield (Guitton et al. 2012). Cider apples are in fact known to be more subject to biennial bearing than dessert apples (Dapena, Minarro, and Blazquez 2005). However, no gene known to control any traits *a priori* differentiating cider and dessert apples was found within the genomic regions examined around the six SNPs detected as significantly associated with the variety type. This may be due to lack of knowledge on these genes, and actually 13% of the genes did not have any predicted function. Alternatively, this may be because selection targeted the regulatory elements in the pathways. In fact, several genes coding for transcriptional regulation elements were found in these candidate regions. Finally, the estimation we made on the extent of LD may not reflect reality in these particular regions (since it is a genome-wide mean value we calculated) and could lead the causative factors for our outliers to be located outside of the windows examined.

All the six genomics regions identified when testing the association between the genotypes and the bitter/sweet phenotype were found to colocalize with QTLs responsible for the content of several

polyphenolic compounds, either measured in the flesh or in the peel of the fruits (Chagne, Krieger, et al. 2012; Khan, Chibon, et al. 2012; Kumar et al. 2012; Verdu et al. 2014). The two SNPs located on LG01 colocalized with three QTLs responsible for *p*-coumaroyl quinic acid, hydroxycinnamic acid and flavanols contents. The SNP located on LG15 colocalized with two QTLs responsible for flavonols and flavanols contents and the two SNPs on LG17 respectively colocalized with QTLs responsible for quercetin 3-*O*-rutinoside and chlorogenic acid contents. The last area located on LG16 colocalized with a region well known to host several strong effect QTLs responsible for numerous polyphenolic compounds such as catechin, epicatechin and procyanidins, all belonging to the flavanol class of polyphenols (Chagne, Krieger, et al. 2012; Khan, Chibon, et al. 2012). A gene coding for a *LeucoAnthocyanidin Reductase* (*LAR*) was identified underlying this QTL hotspot and is thought to be the gene responsible for the numerous QTLs in this area (Khan, Schaart, et al. 2012). The *LAR* gene is indeed the one in the polyphenol pathway leading to the formation of the flavanols from leucocyanidin. Interestingly, the SNP significantly associated with the bitter/sweet phenotype and located on LG16 at 1.43 Mb was close to the *LAR* gene (MDP0000376284) located at 1.53 Mb, which makes our result highly consistent with this particular QTL hotspot and the *LAR* candidate gene. Altogether, the six SNPs associated to the bitter/sweet phenotype were located very close (less than 1 Mb on average) to the markers exhibiting the highest LOD score in the QTL analyses.

2.4.5. Applications in cider apple breeding

This study is a first step for the identification of the genetic bases of phenotypic traits that differentiate cider and dessert apple varieties. In addition, our markers will be useful for marker-assisted selection (MAS) for breeding cider varieties carrying both traits already present in cider varieties (such as high polyphenol content) and traits mainly present in dessert apple varieties (such as annual bearing, high yield or disease resistance). Our markers can indeed guide both the choice of the cider apple progenitors and the selection of seedlings from crosses between dessert and cider varieties, and thus segregating for the favorable haplotypes. By genotyping the seedlings of a cross between a cider and a dessert variety type at the loci we identified as linked with traits of interest, one could choose the individuals bearing the favorable alleles and keep the individuals combining traits from the two variety types. Another application of the information we described here could be the inventory of the several traits and genomic regions responsible for them in order to better manage germplasm diversity in the cultivated apple (Prada 2009).

2.5. Conclusions

Unravelling the genomic bases of quantitative trait variation is essential for understanding evolution and for accelerating plant breeding (Alonso-Blanco and Méndez-Vigo 2014). Furthermore, the question of sustainable management of germplasm resources is increasingly recognized as a fundamental goal to achieve in many crops (see the DivSeek initiative, http://www.divseek.org/). Recently, it has been suggested that an international consortium for the sustainable management of apple genetic diversity in particular is timely (Volk et al. 2014). Our results on the detection of a few key genomic regions involved in the phenotypic differentiation between cider and dessert apples emerging from an otherwise homogeneous genomic background, should be very useful for designing such sustainable apple program. The identified outlier genomic regions will indeed be good targets for screening important genetic variation for conserving both cider and dessert apples specific traits. These programs should also focus on the sustainable conservation of the wild apple gene pools. Wild-to-crop introgressions have indeed been a key driver of the cultivated apple evolution, particularly through introgression from the European crabapple *M. sylvestris* (Cornille et al. 2012). It would be interesting to assess whether the outliers detected here in the cultivated apple have originated from such introgressions from the bitter crabapples. It would feature wild gene pools as sources of key genes for cultivated apple breeding in cider and dessert apples. Overall, our results thus illustrate how genomic can help to feed breeding and conservation programs, and a similar approach could be developed for detecting the genomic basis of other key traits, such as resistance to pathogens or climate adaptation.

Acknowledgments

We thank Philippe Guardiola and Anne Coutolleau of CHU Angers for the genotyping of the individuals using the International RosBREED SNP Consortium (IRSC) apple 8K SNP array v1. We thank Laurence Feugey and Arnaud Guyader for providing access to the genetic resources of INRA Angers, UE HORTI for taking care of the plant materiel and the ANAN platform for DNA quantification. Diane Leforestier also thanks Thibault Leroy for discussions that helped improve this paper. TG, AC and AB thank the BASC labex, the Région Ile de France (PICRI) and the Institut Diversité Ecologie et Evolution du Vivant (IDEEV).

Data Archiving Statement

The genotypic data have been deposited on http://www.rosaceae.org/search/diversity: Accession Number tfGDR1016.

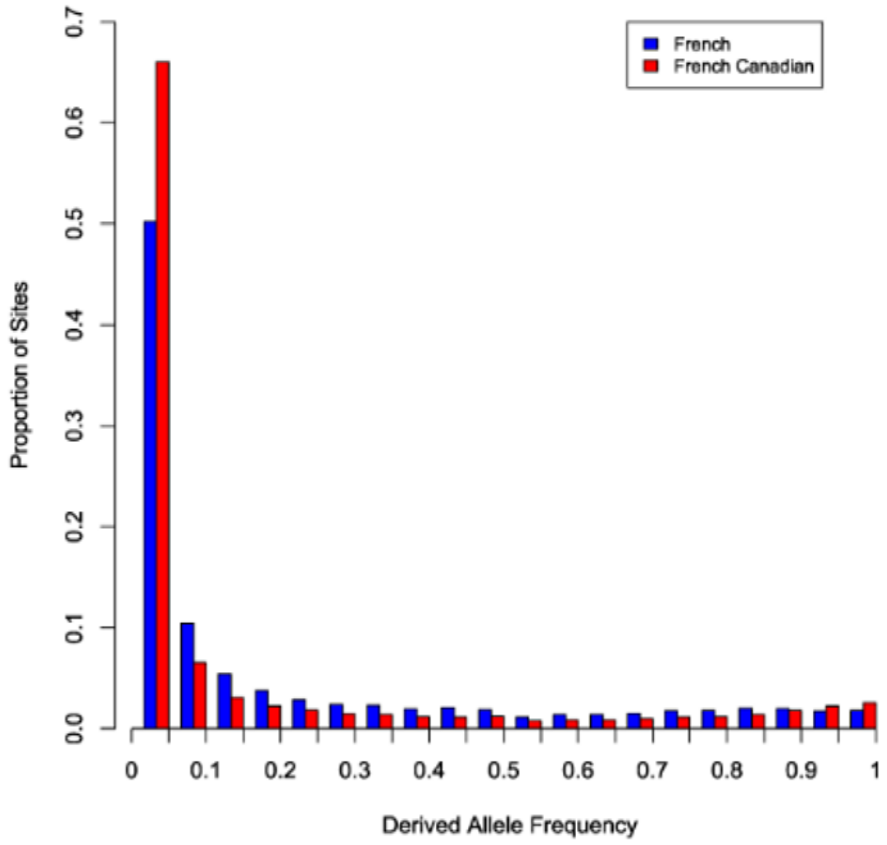


Figure 27 : Spectre de fréquence de site dans deux populations d'hommes français et franco-canadiens à partir de données de re-séquençage d'exome, Casals et al. (2013)

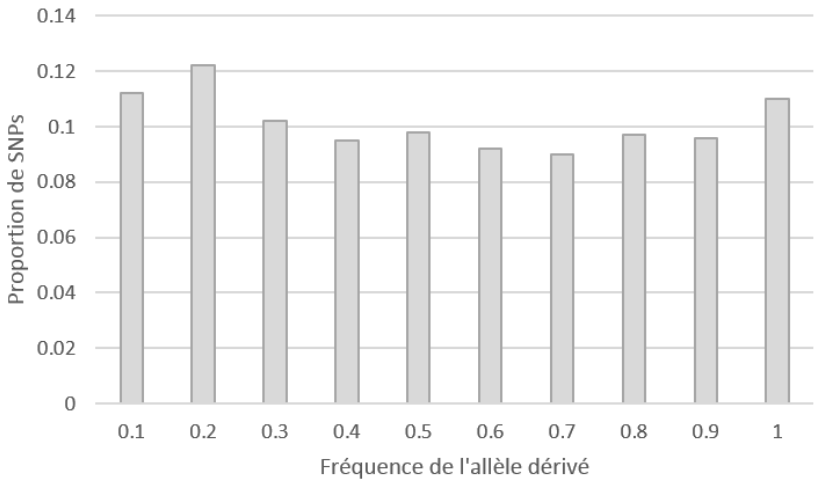


Figure 28 : Spectre de fréquence de site dans les deux core collections de pommiers à cidre et à couteau

3. Analyses complémentaires

En supplément des analyses de différenciation génétique et de génétique d'association de l'article ci-dessus, des analyses de génétique des populations ont été réalisées sur les données de génotypage et portent en particulier sur les estimateurs D de Tajima, F de Fu et Li et H de Fay et Wu (voir Chapitre 1, 1.3.2 et Chapitre 2, 5.1.3). Ces analyses avaient au départ été incluses dans le manuscrit de l'article mais ont été retirées lors de la re-soumission. Les raisons de ce retrait, et notamment les biais auxquels nous avons été confrontés, sont expliqués ci-dessous. Les résultats tels qu'ils avaient été écrits dans le manuscrit sont ensuite présentés.

3.1. Le spectre de fréquences de sites

La plupart des analyses de génétique des populations reposent sur l'hypothèse de neutralité (Kimura 1968) et dépendent entièrement de la distribution des fréquences alléliques, représentées par le Spectre de Fréquences de Sites (SFS). Sous l'hypothèse de neutralité, la représentation graphique du SFS a une forme de L, résultant de la présence de nombreux polymorphismes à basse fréquence et de peu de polymorphismes à haute fréquence (Figure 27), tel que représenté dans Casals et al. (2013). De nombreuses études ont démontré l'importance tant de la distribution des fréquences alléliques que de la population choisie dans des études de génétique des populations (Clark et al. 2005; Albrechtsen, Nielsen, and Nielsen 2010). Les données de génotypage dont nous disposions présentent deux biais importants concernant le SFS et le choix des individus de la population.

La première observation que nous pouvons faire sur les données porte sur la distribution des fréquences alléliques qui ne correspond pas à ce que l'on attend lorsque l'on entreprend une étude de génétique des populations. Au lieu de la forme en L caractéristique de sites évoluant sous l'hypothèse de neutralité, la représentation graphique du SFS montre que toutes les classes de fréquences alléliques sont représentées de façon équivalente (Figure 28). Comme indiqué dans l'article ci-dessus, la conception de la puce 8k SNPs (Chagne, Crowhurst, et al. 2012) a été réalisée à partir de données de re-séquençage à faible profondeur de 27 variétés de pommiers à couteau. Les marqueurs présents sur la puce ont été choisis notamment en fonction de leur fréquence à l'intérieur du pool d'individus re-séquencés, et les SNPs ayant une trop faible fréquence ont été mis de côté. Un grand nombre de marqueurs à faible fréquence manquent donc dans le jeu de données que nous avons analysé, ce qui a conduit à la forme particulière du SFS que l'on obtient ici. De ce fait, les calculs des différents estimateurs de génétique des populations sont largement biaisés puisqu'ils reposent tous sur les fréquences alléliques.

Tableau 6 : Statistiques générales des cinq estimateurs de génétique des populations calculés séparément pour les core collections de pommiers à cidre et de pommiers à couteau

Pommes à couteau	H Fay & Wu	D Tajima	F Fu & Li	H_e[a]	H_o[b]
Minimum	-3,74	0,04	0,82	0,22	0,21
Maximum	1,09	3,46	2,88	0,43	0,47
Moyenne	-0,86	1,69	2,02	0,33	0,34

Pommes à cidre	H Fay & Wu	D Tajima	F Fu & Li	H_e[a]	H_o[b]
Minimum	-5,33	-0,68	0,26	0,17	0,18
Maximum	3,06	2,98	2,67	0,43	0,49
Moyenne	-0,87	1,63	1,96	0,32	0,33

[a] Hétérozygotie attendue (H_e)
[b] Hétérozygotie observée (H_o)

Tableau 7 : Valeurs des différents estimateurs de génétique des populations dans les régions potentiellement sous sélection dans la core collection de pommiers à cidre

N°	LG	Position début	Position fin	D Tajima	F Fu & Li	H Fay & Wu	H_e/H_o[a]
1	6	18 664 338	22 106 755	-0,19	0,27	-0,71	0,22/0,24
2	12	19 386 503	25 092 903	-0,68	0,76	-0,95	0,17/0,18
3	17	9 003 320	11 333 375	0,53	0,52	-1,13	0,22/0,34

[a] Hétérozygoties attendue (H_e) et observée (H_o)

Le deuxième biais auquel nous avons été confrontés pour ces analyses concerne l'échantillonnage de la population d'étude. Les core collections ne sont en effet pas appropriées pour mener des études de génétique des populations qui portent en général sur des populations naturelles. Il existe, il est vrai, des études menées sur des espèces domestiquées telles que le bœuf ou le chien, mais les individus étudiés ne sont généralement pas choisis pour maximiser le niveau de diversité au sein de la population d'étude. Ce choix conduit donc à une surestimation des paramètres calculés ci-dessous, notamment parce qu'ils évaluent la diversité génétique au sein des populations.

3.2. Résultats

3.2.1. Analyses de génétique des populations

Les valeurs moyennes, maximales et minimales des différents estimateurs (D, F, H et H_e/H_o), calculés afin de détecter la présence de signatures de sélection dans le génome du pommier, sont reportées dans le Tableau 6. En comparant visuellement les variations de ces estimateurs (Figure 29) dans les génomes des pommes à cidre et des pommes à couteau, trois régions montrent des tendances différentes. Ces régions, localisées sur les LG06, LG12 et LG17 montrent les valeurs les plus basses de D de Tajima et F de Fu et Li, la plupart du temps inférieures à 0 pour le D de Tajima, et ce uniquement dans la CC48 Cidre. La région localisée sur le LG12 montre également des faibles valeurs d'H_e/H_o, ainsi que des valeurs négatives de H de Fay et Wu (Tableau 7).

3.2.2. Recherche de gènes candidats

Les gènes localisés dans les régions potentiellement sous sélection ont été étudiés. Les fenêtres dans lesquelles les gènes ont été trouvés ont été calculées sur un nombre fixe de SNPs et non pas sur une distance fixe. Cela a conduit à l'identification de 560, 695 et 247 gènes respectivement pour les LG06, LG12 et LG17. Les classes de gènes les plus représentées sont les classes des gènes ayant une fonction dans le métabolisme des protéines, des ARN et dans les voies de signalisation (Figure 30). Parmi les gènes de la classe Divers, 10 gènes présentent une homologie de séquence avec des gènes de la voie des polyphénols.

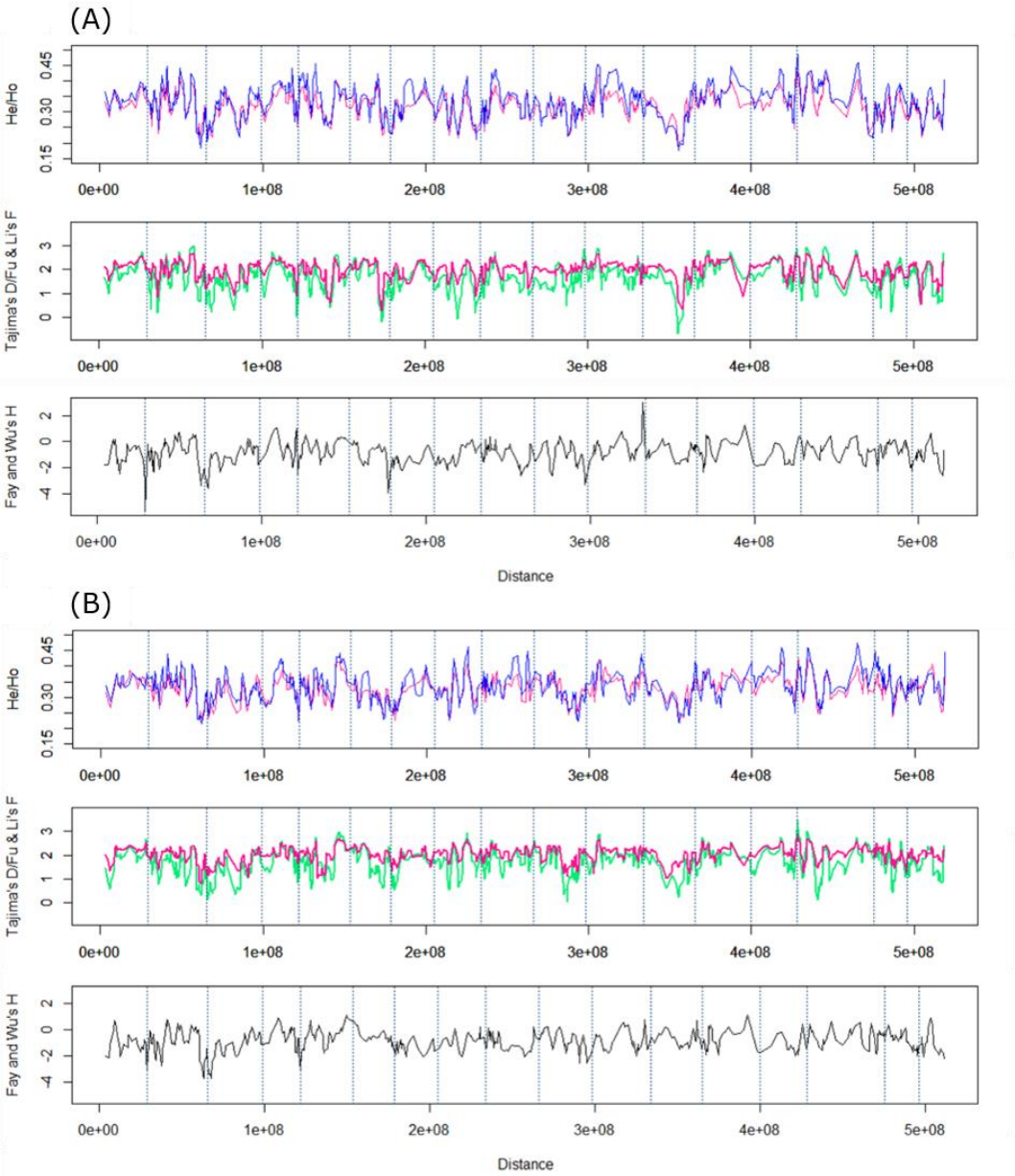


Figure 29 : Variation des niveaux d'hétérozygoties observée et attendue, D de Tajima, F de Fu et Li et H de Fay et Wu le long du génome de deux core collections de pommiers à cidre (A) et à couteau (B) en partant du haut sur chaque figure : hétérozygoties attendue (bleu foncé) et observée (rose) ; D de Tajima (vert) et F de Fu et Li (fuschia) ; H de Fay et Wu (noir)

3.3. Discussion

3.3.1. Estimateurs statistiques de génétique des populations

Les différents estimateurs de génétique des populations calculés montrent comme attendu des valeurs moyennes trop élevées ou trop faibles. La valeur moyenne du D de Tajima dans les CC48 Cidre et Couteau, 1,63 et 1,69 respectivement, ne correspond pas aux attendus sous le modèle de neutralité. Le D de Tajima est en effet attendu à une valeur proche de 0 lorsque les allèles évoluent de façon neutre, au-dessus de 0 pour les loci sous sélection balancée et en dessous de 0 pour les loci sous sélection positive, dans une population où aucun événement démographique n'a eu lieu. Une valeur positive de D de Tajima reflète un excès de variants de fréquence intermédiaire qui résulte très vraisemblablement ici de l'utilisation de core collections qui maximisent la diversité, et donc conduisent à la surreprésentation des allèles rares. Plusieurs études ont permis l'identification de gènes candidats présentant des valeurs positives ou négatives de D de Tajima. Chez l'épicéa *Picea abies* par exemple, des valeurs négatives de D de Tajima ont été détectées dans des gènes potentiellement impliqués dans la vernalisation et les voies de signalisation de la photopériode (Heuertz et al. 2006). Des valeurs positives de D de Tajima ont quant à elles été détectées chez le maïs *Zea mays*, avec une valeur moyenne de 0,04 (Wright et al. 2005), résultant soit d'une sélection balancée soit d'un goulot d'étranglement récent dans la population. Ici, la valeur moyenne de D de Tajima est 40 fois supérieure à celle trouvée chez le maïs. Les goulots d'étranglement sont connus, au même titre que la sélection balancée, pour générer des valeurs positives de D de Tajima. Cette hypothèse est cependant peu probable puisqu'aucun événement démographique de ce genre n'a été détecté chez le pommier cultivé (Cornille et al. 2012).

Le F de Fu et Li reflète globalement les mêmes processus que le D de Tajima, une valeur positive reflétant un excès de polymorphismes de fréquence intermédiaire et une valeur négative reflétant un excès de polymorphismes à basse fréquence. Le F de Fu et Li est cependant plus sensible aux expansions de population et aux autostops génétiques. Les valeurs moyennes trouvées ici, respectivement de 2,88 et 2,67 dans les CC48 Cidre et Couteau, reflètent tout comme le D de Tajima un excès de polymorphismes de fréquence intermédiaire imputable à l'utilisation des core collections. Les régions génomiques dans lesquelles le F de Fu et Li et le D de Tajima sont les plus faibles simultanément reflètent pourtant probablement l'existence d'un événement de sélection purifiante. Ces trois régions ont donc été étudiées et parmi elles, celle localisée sur le LG12 montre, en plus des faibles valeurs de D de Tajima et de F de Fu et Li, une valeur très basse d'H_e/H_o ainsi qu'une valeur négative de H de Fay et Wu ce qui augmente la vraisemblance pour cette région d'avoir été sous sélection négative.

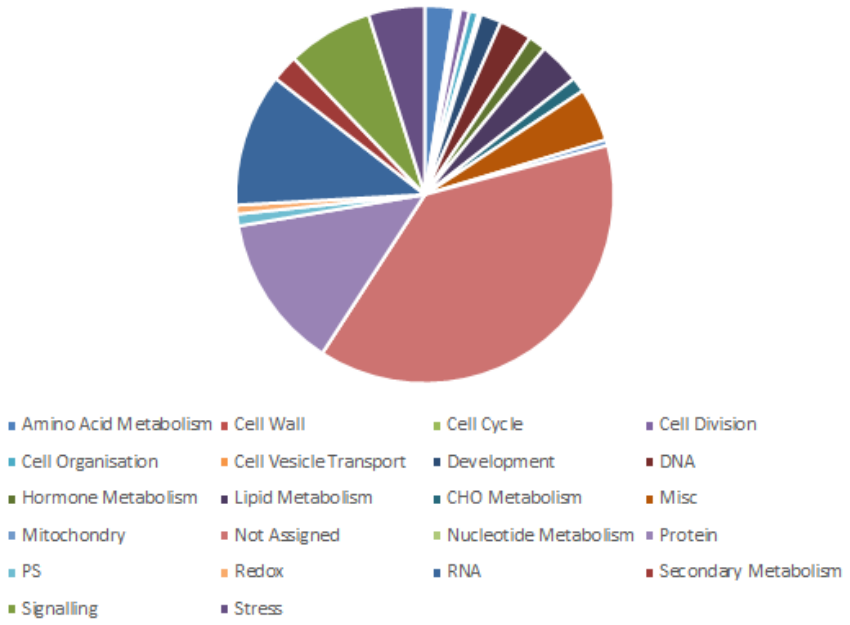


Figure 30 : Distribution des gènes identifiés lors de l'étude de génétique des populations entre pommiers à cidre et à couteau dans les principales classes de gènes

Aucune différence significative n'a été observée dans la variation du H de Fay et Wu qui représente le taux d'allèles ancestraux (obtenus à partir de *Prunus persica*) et dérivés. Les valeurs moyennes dans les CC48 Cidre et Couteau, respectivement -0,87 et -0.86, sont très semblables. On observe cependant des valeurs plus extrêmes dans la CC48 Cidre. Les régions montrant des valeurs extrêmes de H de Fay et Wu ne co-localisent pas avec les trois régions citées précédemment mais H prend des valeurs négatives dans ces régions ce qui suggère soit un événement de sélection soit un événement démographique récent.

Aucune signature de sélection n'a été trouvée dans le génome des pommes à couteau. Ce résultat est surprenant puisque l'on supposait des pressions de sélection assez importantes pour évoluer depuis les pommes petites et amères vers les grosses pommes sucrées consommées actuellement. Une explication possible pour ces résultats est l'utilisation de la puce de génotypage. Comme mentionné précédemment, le choix des marqueurs de la puce a été fait sur des données de re-séquençage de variétés à couteau en gardant les marqueurs hautement polymorphes et en éliminant les variants rares. L'estimation du D de Tajima est de ce fait biaisée et les potentielles signatures de sélection dans le génome des pommes à couteau indétectables. Les régions détectées dans le génome des pommiers à cidre dépendent quant à elles potentiellement de marqueurs ayant des fréquences intermédiaires chez les variétés de pommiers à couteau et des fréquences basses chez les variétés de pommiers à cidre. Une autre explication pour cette absence de signal serait que les variétés à couteau ont été sélectionnées sur beaucoup plus de caractères d'intérêt que les variétés à cidre ce qui participerait à la dilution du signal.

3.3.2. Recherche de gènes candidats

Parmi les gènes identifiés dans les régions ayant les plus faibles valeurs de D de Tajima et de F de Fu et Li, 10 gènes ont été identifiés comme ayant une haute homologie de séquence avec des gènes de la voie des polyphénols. Parmi ces gènes on retrouve : un gène homologue d'une 4-coumarate:CoA ligase responsable de la synthèse des molécules précurseurs de la voie des phénylpropanoïdes, deux gènes homologues de flavanone 3β-hydroxylases qui convertissent les flavonones en dihydroflavonols, trois gènes homologues d'UDP-glucose:flavonoïdes 3-O-glucosyltransférases et deux homologues d'UDP-glucose:anthocyanidines 3-O-glucosyltransférases qui forment les précurseurs des anthocyanidines, un gène homologue d'une isoflavone réductase responsable de la réduction des hydroxisoflavones en isoflavones et un gène homologue d'une dyhydroflavonol 4-réductase qui catalyse la réaction de synthèse des leucoanthocyanidines (Andersen and Markham 2006). Toutes ces enzymes sont des enzymes de la voie de biosynthèse des polyphénols, molécules connues pour être présentes en plus grande quantité dans les pommes à cidre que dans les pommes à couteau (Sanoner et al.

1999). Il est intéressant de souligner que la région potentiellement sous sélection et localisée sur le LG17 (9 Mb à 11,3 Mb) coïncide à nouveau avec un des QTLs détectés dans une étude sur les polyphénols de la pomme (Chagne, Krieger, et al. 2012). Ce QTL, dont le marqueur présentant le plus fort LOD (GDsnp02075, LOD = 3,73 et r^2 = 14,7%) est localisé à 11,59 Mb, est responsable de la concentration en quercetine 3-O-xyloside dans le fruit. Les intervalles de confiance des QTLs sont généralement larges (plusieurs cM correspondant chez le pommier à plusieurs Mb) et il est possible ici que la zone identifiée recouvre en partie l'intervalle de confiance du QTL. Ces deux signaux pourraient alors correspondre à l'effet d'un même gène candidat qui serait polymorphe et dont un des allèles aurait été préférentiellement sélectionné dans les variétés à cidre. Il est également intéressant de noter que cette région potentiellement sous sélection contient les deux SNPs localisés sur le LG17 et identifiés lors de l'analyse de différenciation génétique grâce au F_{ST}.

En dehors des gènes de la voie des polyphénols, aucune autre voie métabolique caractéristique des pommes à cidre ou à couteau n'a été identifiée. Certains gènes sans fonction prédite ou facteurs de transcription localisés dans ces régions pourraient être responsables de cette absence de résultat.

4. Discussion générale

L'analyse visant à étudier la différenciation entre les deux core collections de variétés de pommiers à cidre et à couteau a permis l'identification de différentes régions génomiques potentiellement impliquées. Au total, quatre régions contenant cinq marqueurs ont été identifiées lors de l'analyse de différenciation génétique, douze régions lors des analyses testant d'une part l'association du génotype avec le phénotype cidre-couteau et d'autre part avec le phénotype doux-amer, et trois régions lors des analyses visant à détecter les régions sous sélection. Parmi ces régions, celle localisée sur le LG17, détectée comme étant potentiellement sous sélection et contenant deux marqueurs présentant de fortes valeurs de F_{ST}, contient des gènes présentant une forte homologie de séquence avec des gènes de la voie des polyphénols. De la même façon, toutes les régions identifiées lors de l'analyse d'association sur le caractère doux-amer des fruits co-localisent avec des QTLs de la teneur en polyphénols des fruits (Chagne, Krieger, et al. 2012; Khan, Chibon, et al. 2012; Kumar et al. 2012; Verdu et al. 2014). Ces résultats soulignent l'importance probable de la voie des polyphénols dans la différenciation entre les pommes à cidre et les pommes à couteau, et sont cohérents avec la quantité plus élevée en moyenne (toutes les pommes à cidre ne sont en effet pas amères) de polyphénols dans les premières en comparaison avec les secondes (Dapena, Minarro, and Blazquez 2005).

Les analyses portant sur les potentielles signatures de sélection dans le génome des pommiers à cidre et à couteau, bien que biaisées par l'utilisation de la puce de génotypage et de core collections,

ont permis l'identification de trois régions génomiques. Ces régions s'étendent sur de grandes distances (plusieurs Mb) et n'ont pas permis l'identification précise de gènes candidats, tout comme l'étude d'association sur le phénotype cidre-couteau. Une connaissance plus fine des différences phénotypiques entre les pommes à cidre et les pommes à couteau pourrait permettre une meilleure analyse des gènes identifiés dans ces régions. Les analyses de différenciation génétique et d'association ont conduit à l'identification de régions du génome différentes. Ces différences s'expliquent d'abord par le fait que lors des analyses d'association, la structure et l'apparentement entre individus sont pris en compte, contrairement aux analyses de différenciation. Ensuite, l'analyse de différenciation et les deux analyses d'association ont été réalisées sur des données phénotypiques très différentes. L'analyse du F_{ST} et la première analyse d'association ont en effet été effectuées sur le phénotype cidre-couteau des individus, qui regroupe probablement de nombreux caractères, tandis que le phénotype doux-amer testé lors de la deuxième analyse cible principalement le caractère de teneur en polyphénols.

L'utilisation de la génétique des populations pour comparer deux groupes à l'intérieur d'une espèce est un élément prometteur pour la découverte des bases génétiques de la différenciation entre ces deux groupes et pour mieux comprendre les événements de sélection ayant mené à la formation des variétés. Des approches similaires ont été développées chez l'animal dans des espèces telles que le bœuf et le porc (Rothammer et al. 2013; Wang, Chen, et al. 2014), mais l'identification précise des traits sous sélection se révèle difficile, en particulier lorsqu'un grand nombre de caractères ont été soumis à sélection. La présente étude s'est focalisée sur les gènes impliqués dans la teneur en polyphénols, plus élevée chez les pommes à cidre que chez les pommes à couteau. Les régions génomiques détectées par l'analyse du F_{ST} ou de l'association cidre-couteau ne contenant pas de gènes appartenant à cette voie peuvent contenir des gènes contrôlant la variation de traits phénotypiques ne présentant que peu ou pas d'intérêt du point de vue de la sélection moderne. C'est le cas par exemple du port plus buissonnant d'un certain nombre de variétés anciennes de pommiers à cidre qui aurait pu donner un signal de sélection. L'identification des régions soumises à sélection peut donc aider à l'identification des caractères soumis à sélection lors de la différenciation des variétés (Rothammer et al. 2013).

Cette étude est une première étape vers l'identification des bases génétiques de la différenciation entre les pommiers à cidre et les pommiers à couteau. Elle est basée sur un nombre limité d'individus maximisant la diversité génétique au sein des deux groupes. Une étude portant sur un nombre plus élevé d'individus et de marqueurs permettrait un gain de puissance pour les analyses statistiques ainsi qu'un gain de précision dans la localisation des loci impliqués. L'utilisation des puces 20k (Bianco et al. 2014) et 480k (Bianco, Durel, Troggio, et al., in prep.) permettrait d'augmenter le nombre de marqueurs utilisés mais pas de s'affranchir du biais lié au SFS. A l'inverse, un génotypage de type re-séquençage tel que le séquençage RAD (Elshire et al. 2011) pourrait être une solution à ce problème bien qu'il faille garder à l'esprit que le pommier est une espèce hautement hétérozygote ce

qui pourrait être une limite dans l'identification précise des génotypes aux SNPs révélés après restriction par des enzymes, si celles-ci ciblent des sites polymorphes (Myles 2013).

Chapitre 4

Variation du déséquilibre de liaison dans une core collection de pommiers à couteau

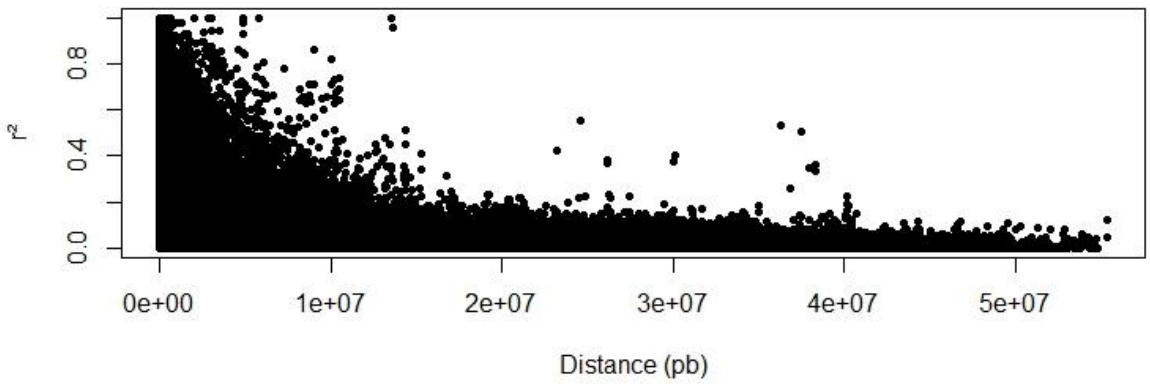


Figure 31 : Graphique du déséquilibre de liaison estimé grâce au r² calculé entre toutes les paires de marqueurs

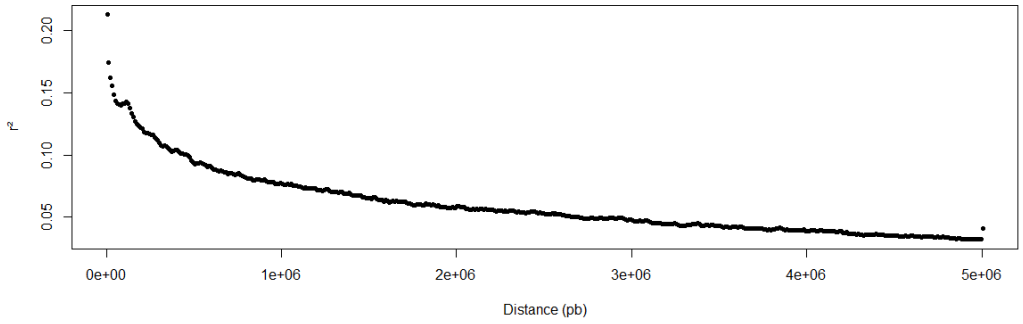


Figure 32 : Courbe de décroissance du déséquilibre de liaison (exprimé par le paramètre r²), en fonction de la distance physique en utilisant les données de génotypage issus de la puce 480k SNPs sur la CC278

Tableau 8 : Estimation de la distance physique à partir de laquelle le r² descend en-dessous de 0,2 par groupe de liaison dans les CC278 et CC48

Groupe de liaison	r² CC278	r² CC48
LG01	25 kb	25 kb
LG02	< 5 kb	< 5kb
LG03	10 kb	10 kb
LG04	5 kb	10 kb
LG05	< 5 kb	< 5 kb
LG06	10 kb	10 kb
LG07	10 kb	5 kb
LG08	10 kb	10 kb
LG09	10 kb	10 kb
LG10	5 kb	5 kb
LG11	5 kb	5 kb
LG12	< 5 kb	< 5 kb
LG13	25 kb	35 kb
LG14	5 kb	10 kb
LG15	15 kb	20 kb
LG16	15 kb	10 kb
LG17	< 5kb	<5 kb

1. Introduction

Le déséquilibre de liaison est un outil important dans les études de cartographie par génétique d'association et dans les études de génétique des populations. Son étude permet en effet d'estimer le nombre minimum de marqueurs à utiliser lors d'une analyse de génétique d'association afin de couvrir (en théorie) la totalité du génome, mais également de déterminer une fenêtre de distance à l'intérieur de laquelle les gènes candidats devront être préférentiellement recherchés à la suite de détections d'associations phénotype-génotype significatives. L'analyse de la puce 8k SNPs nous a permis dans le chapitre précédent d'estimer l'étendue du DL dans deux core collections de pommiers à couteau et de pommiers à cidre. La valeur obtenue de 100 kb pour un r^2 inférieur à 0,2 semble élevée lorsque l'on prend en compte le système de reproduction du pommier ainsi que la population étudiée, censée regrouper un grand nombre de recombinaisons. Nous avons donc d'abord utilisé les données de génotypage issues d'une puce 480k SNPs (Bianco, Durel, Troggio, et al., in prep.) pour chercher à mieux évaluer le DL à l'échelle globale du génome. Des données de re-séquençage de gènes candidats positionnels obtenues sur la CC278 ont ensuite permis d'analyser plus en détails l'étendue et la structure du DL en zoomant sur certaines régions génomiques choisies au préalable.

2. Résultats

2.1. Données de puce 480k SNPs

2.1.1. Statistiques générales

Les données de puce 480k SNPs ont été utilisées afin de mesurer l'étendue du DL moyen au niveau du génome entier. La représentation graphique du r^2 pour la totalité des combinaisons entre les différents marqueurs ne permet pas une bonne visualisation de la décroissance du DL (Figure 31). Il a donc été choisi de représenter la décroissance du DL en fonction de la distance physique en calculant le r^2 moyen par incréments de 10 kb (Figure 32), et d'estimer visuellement la valeur en dessous de laquelle le r^2 devient inférieur à 0,2. Le r^2 moyen à l'échelle du génome entier sur la CC278 (en réalité 268 individus retenus *in fine* à cause du manque de données phénotypiques pour 10 des individus de la CC278), calculé entre toutes les paires de marqueurs d'un même groupe de liaison et n'étant pas distants de plus de 500 kb, décroit en-dessous de 0,2 au-delà de 10 kb. Les valeurs calculées par LG vont de moins de 5 kb (LG02, LG05, LG12 et LG17) à 25 kb (LG01 et LG13 ; Tableau 8 et Annexe 5).

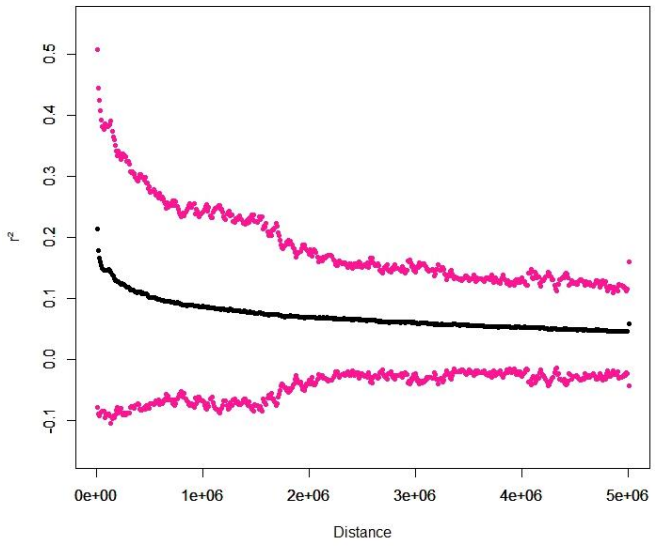


Figure 33 : Représentation d'une courbe de décroissance du déséquilibre de liaison entourée de courbes correspondant aux valeurs du r² +/- l'écart type calculé par incrément de distance

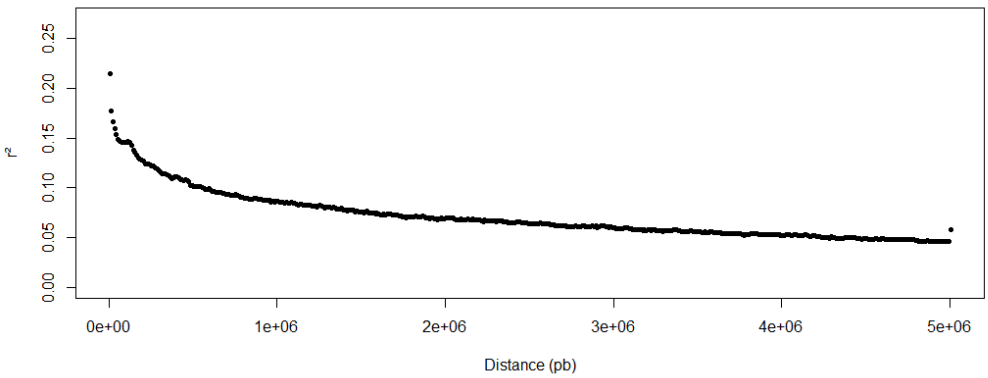


Figure 34 : Courbe de décroissance du déséquilibre de liaison (exprimé par le paramètre r²), en fonction de la distance physique en utilisant les données de génotypage issus de la puce 480k SNPs sur la CC48

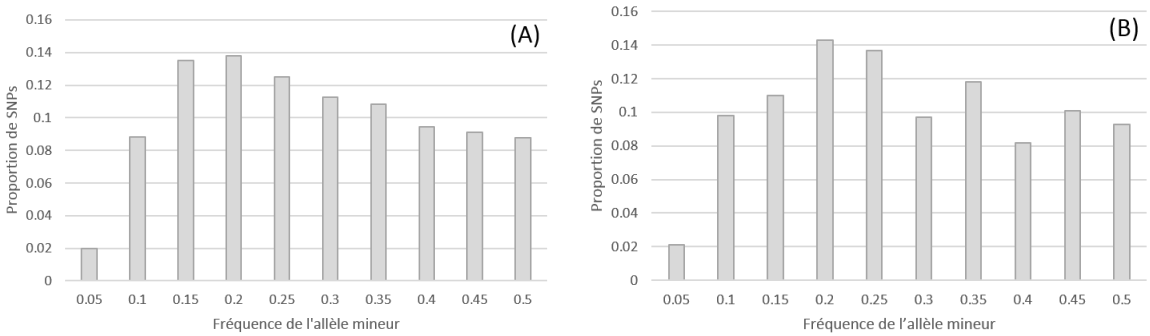


Figure 35 : Spectre de fréquences alléliques dans les deux core collections de pommiers à couteau CC278 (A) et CC48 (B) génotypées avec la puce 480k SNPs

Un très grand écart-type est observé pour chaque point de la courbe de décroissance du DL (Figure 33). La valeur de l'écart-type pour chaque incrément est toujours supérieure à celle du r^2 moyen.

Afin d'apprécier l'impact potentiel de la taille (et de la composition) de la core collection, la même évaluation a été réalisée sur la CC48 Couteau. Dans ce cas, le r^2 moyen à l'échelle du génome entier décroit également en-dessous de 0,2 à partir de 10 kb (Figure 34) et les valeurs par LG varient de moins de 5 kb (LG02, LG05, LG12 et LG17) à 35 kb pour le LG13 (Tableau 8 et Annexe 6). Les histogrammes de distribution des fréquences alléliques mineures (MAF) pour les deux core collections génotypées avec la puce 480k sont présentés en Figure 35. On observe très peu de marqueurs (1,8% et 2,2% respectivement) avec une MAF faible (< 0,05) et une représentation équivalente des autres classes de MAF dans les deux core collections. La classe de MAF majoritaire est la classe des fréquences comprises entre 0,15 et 0,20, les autres classes ayant des fréquences similaires dans les deux cas.

2.1.2. Etude de régions ciblées

La visualisation des blocs haplotypiques grâce au logiciel Haploview 4.2 (Barrett et al. 2005) le long de régions génomiques ciblées montre des patterns très différents de l'attendu. Les diagrammes du DL montrent en effet habituellement des blocs de SNPs en fort DL, mais aucun ou peu de DL entre blocs de SNPs (plus la distance entre deux blocs est importante, moins ils sont susceptibles d'être en fort DL). Sur le haut du LG01 par exemple (Figure 36), on observe, dans le haut du diagramme, des blocs de marqueurs présentant des fortes valeurs de r^2 entre eux. Le bas du diagramme pose quant lui problème : on observe en effet deux groupes de blocs en fort DL entre eux, avec des valeurs de r^2 entre marqueurs relativement élevées (> 0,8), mais aucun DL entre les blocs des deux groupes. Les SNPs des blocs n'étant pas en DL entre eux appartiennent à des scaffolds différents, laissant supposer une mauvaise position de ces scaffolds. A l'intérieur d'un même groupe de blocs de SNPs en fort DL entre eux, le r^2 garde des valeurs très élevées (> 0,8) même à une distance supérieure à 1 Mb.

C'est le cas également dans une région d'environ 3 Mb sur le haut du LG16. La figure 37 montre le pattern de DL entre un sous-ensemble de marqueurs de cette région. Quatre zones du diagramme montrent des valeurs élevées de r^2 : une première zone au début de la région étudiée et s'étendant sur 63 kb (en haut à gauche du diagramme), une seconde zone au milieu de la région étudiée et s'étendant sur 289 kb (en haut et au milieu du diagramme), une troisième zone à l'extrémité de la région étudiée, qui s'étend sur 1 217 kb mais qui présente de nombreux endroits où le DL chute soudainement (en haut à droite du diagramme) et une quatrième zone dénotant la présence d'un DL élevé entre la première et la troisième région (en bas du diagramme). La présence de cette quatrième zone localisée dans le bas du diagramme pose question à une distance si grande (2,7 Mb), d'autant

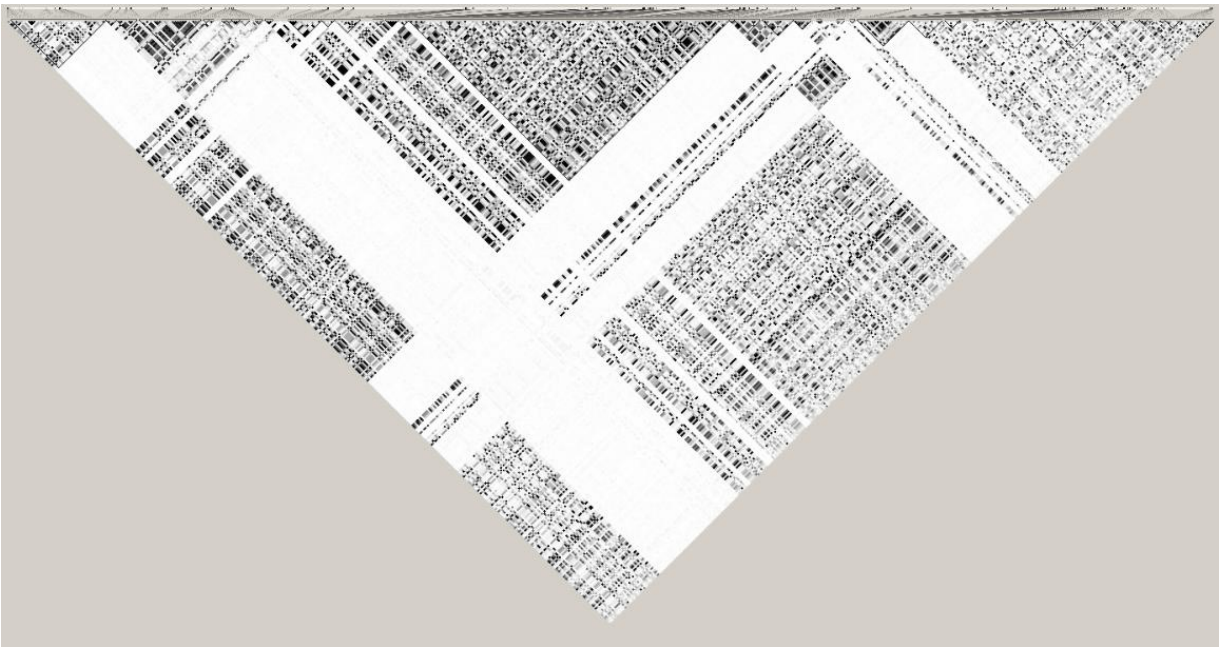


Figure 36 : Représentation graphique des blocs de SNPs en déséquilibre de liaison sur le haut du LG01 sur une distance de 2,3 Mb
les niveaux de gris correspondent à la valeur du r^2, blanc pour une valeur de 0
et noir pour une valeur de 1

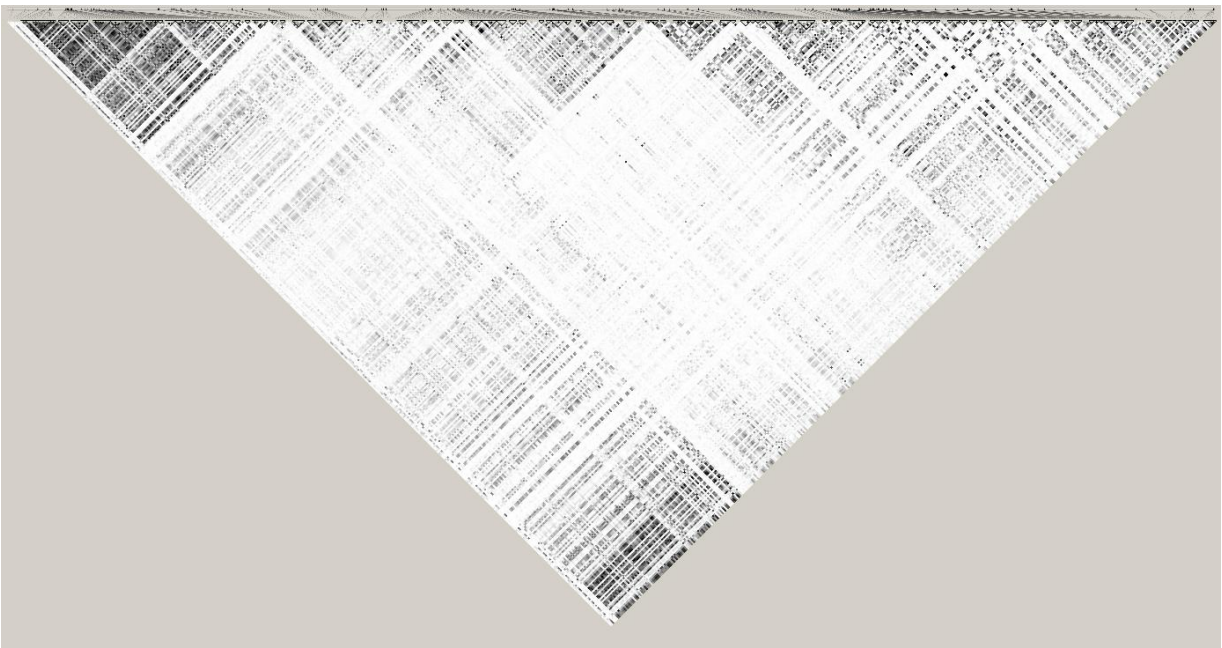


Figure 37 : Représentation graphique des blocs de SNPs en déséquilibre de liaison sur le haut du LG16 sur une distance de 2,7 Mb dans la CC278

plus que le niveau de DL est faible entre la seconde zone et les zones une et trois. L'attendu lorsque l'on examine le pattern du DL est de voir apparaître des blocs en fort DL le long des régions mais ces blocs ne sont habituellement pas en DL entre eux.

Cette situation a été retrouvée dans plusieurs endroits du génome choisis au hasard ou choisis dans les régions d'intérêt de la génétique d'association (Annexe 7). Les différentes régions, analysées plus en détails et contenant des SNPs localisés sur un seul ou plusieurs scaffolds, présentent toutes le même pattern que la région localisée sur le LG01 dans lequel on peut distinguer des contigs ou scaffolds mal placés (pas de DL avec les contigs ou scaffolds adjacents) mais également un DL qui reste fort (> 0,8) même entre des marqueurs éloignés de plusieurs Mb.

2.2. Données de re-séquençage de gènes

Un total de 96 fragments de gènes, répartis dans six régions du génome du pommier, a été re-séquencé chez 251 individus de la CC278. Certains des fragments, choisis au départ pour être localisés dans des régions exoniques, se sont malheureusement révélés être dans des régions introniques voire inter-géniques à cause d'erreurs d'orientation de certains contigs dans la version 1 du génome du pommier. Au final, 59 fragments sur les 96 étudiés ont conservé leur position initiale.

Une dizaine d'individus ont été génotypés deux fois afin de contrôler la qualité du séquençage. Un taux d'erreurs de 6‰ sur la totalité des fragments retenus pour ces 10 individus a été observé, ce qui indique une très bonne répétabilité des données disponibles et en conséquence des haplotypes inférés.

Suite au séquençage sur MiSeq, chaque fragment a été lu avec une profondeur moyenne de 2 000X. En moyenne, 21 SNPs ont été trouvé par fragment, les valeurs allant de un SNP à 73 SNPs (Annexe 1). Les fragments localisés sur le bas du LG01 présentaient le plus grand nombre de SNPs (28,8 SNPS en moyenne) et ceux du LG07 le plus faible nombre (16,2 SNPs en moyenne). En regardant les fragments contenus à l'intérieur d'un même gène, on remarque que le nombre de SNPs peut être très différent entre les deux fragments, sans influence de la position du fragment en partie 5' ou 3' du gène. La localisation des fragments dans des régions codantes ou non codantes ne semble pas non plus avoir eu d'effet sur le nombre de SNPs identifiés.

Le pommier étant diploïde, l'attendu était d'observer un ou deux haplotypes majoritaires correspondant aux situations d'homozygotie ou d'hétérozygotie respectivement. Cependant, la combinaison de la PCR par Fluidigm® et du re-séquençage MiSeq peut avoir entraîné des erreurs de séquence générant par la suite de faux haplotypes au sein d'un individu donné. Selon les cas, la distribution du nombre d'haplotypes variait fortement, certains fragments pour lesquels la fréquence du premier haplotype était supérieure à 95% et d'autres pour lesquels aucun haplotype majoritaire ne

Tableau 9 : Tableau récapitulatif du nombre moyens d'haplotypes, de la fréquence la plus élevée et de la fréquence la plus faible de l'haplotype le plus fréquent dans la CC278 pour 38 fragments de gènes re-séquencés

LG	Nombre moyen d'haplotypes	Fréquence 1er haplotype la plus élevée	Fréquence 1er haplotype la plus faible
01	12,5	0,89	0,28
03	NA	NA	NA
07	9,7	0,91	0,47
10	11,7	0,94	0,41
11	11,7	0,91	0,25
17	9,3	0,78	0,41

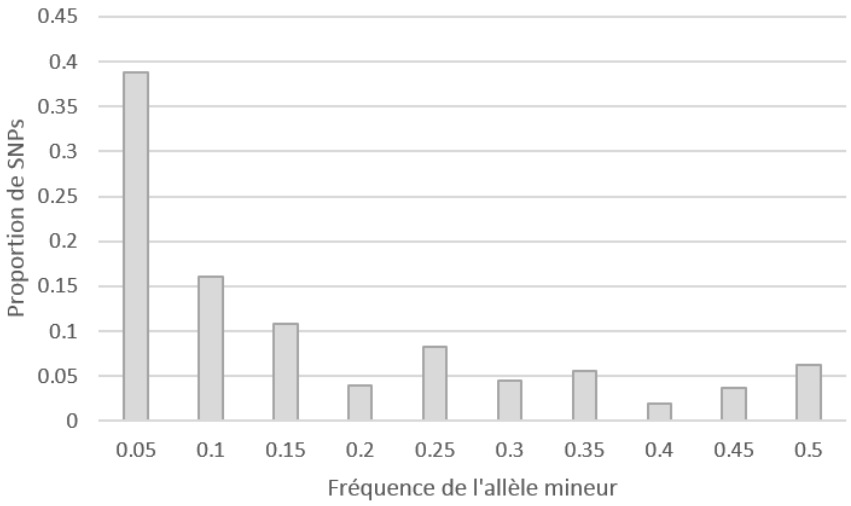


Figure 38 : Spectre de fréquences alléliques dans la core collection de pommiers à couteau CC278 sur les données de re-séquençage de fragments de gènes

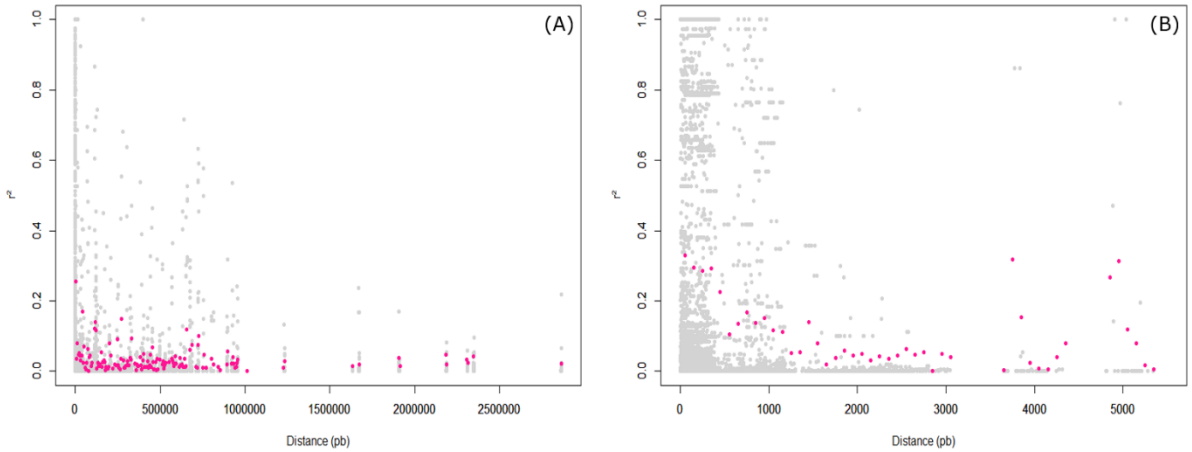


Figure 39 : Courbes de décroissance du déséquilibre de liaison (exprimé par le paramètre r^2), en fonction de la distance physique en utilisant les données de génotypage issus des données de reséquençage de fragments de gènes sur la CC278 sur les données poolées des 6 régions génomiques en inter- (A) et intra-génique (B)

les points gris représentent la totalité des comparaison entre toutes les paires de marqueurs ; les points roses représentent les r^2 moyens par incrément de 1 250 pb

se distinguait. Des haplotypes ont été attribués à la majorité des fragments : sur la totalité des 24 096 fragments individuels séquencés, plus de 6 000 fragments étaient clairement homozygotes avec la fréquence du premier haplotype supérieure à 90%, et plus de 10 000 étaient clairement hétérozygotes. Les fragments pour lesquels aucun haplotype majoritaire n'a été identifié ont été supprimés des analyses. Sur les 96 fragments re-séquencés, 86 ont été analysés (Annexe 8). Certains de ces fragments présentent des données manquantes au niveau des marqueurs situés à l'extrémité des reads, donc localisés au milieu des fragments. Parmi les fragments ne comprenant pas de données manquantes (38 sur 86), le nombre d'haplotypes varie de quatre à 37, avec la fréquence du premier haplotype comprise entre 25% et 94% (Tableau 9). Aucune des six régions ne semble contenir un nombre plus élevé de fragments présentant de nombreux haplotypes.

L'histogramme de distribution des MAF des SNPs identifiés dans les données de re-séquençage chez les individus de la CC278 est présenté en Figure 38. La plupart des marqueurs (~ 40%) présentent une MAF faible (< 0,05) et les autres classes de MAF sont de moins en moins représentées.

L'estimateur du DL, r^2, calculé entre toutes les paires de marqueurs, soit au sein d'un même gène (deux fragments considérés), soit au sein d'une même région génomique, a donné les résultats présentés dans le Tableau 10, la Figure 39 et l'Annexe 9. Certains fragments des LG03 et LG11 étaient localisés sur des scaffolds différents sur la version 1 du génome. Les analyses ont donc été réalisées seulement sur les fragments d'un même scaffold puisque la distance entre les scaffolds, fixée arbitrairement à 200 kb dans la version 1 du génome, était en réalité inconnue, ce qui aurait engendré un biais dans les calculs du r^2. Les valeurs de l'étendue du DL intra-génique varient de 250 pb à 750 pb et celles du DL inter-génique (à l'intérieur d'une région génomique) de 1 000 pb à 5 000 pb, les deux valeurs semblant corrélées entre les deux échelles (Tableau 10). La visualisation des blocs haplotypiques le long des six régions étudiées est montrée pour le LG01 en Figure 40 et pour les autres régions en Annexe 10. Contrairement aux diagrammes obtenus avec les données de la puce 480k SNPs, les blocs de SNPs en fort DL ne sont ici observés que sur le haut du diagramme, les blocs de SNPs n'étant pas en fort DL à longue distance.

3. Discussion

3.1. Comparaison des estimations de déséquilibre de liaison obtenues avec les deux puces de génotypage

Dans le Chapitre 3, en utilisant les données de la puce 8k SNPs (Chagne, Crowhurst, et al. 2012), nous avons montré que la valeur du r^2 passait en dessous de 0,2 au-delà de 100 kb (Figure 24).

Tableau 10 : Estimation de la distance physique à partir de laquelle le r^2 descend en-dessous de 0,2 par région génomique abritant les gènes re-séquencés dans la CC278

LG	LD intra	LD inter
01	750 pb	5000 pb
03	650 pb	2000 pb
07	400 pb	1000 pb
10	250 pb	1000 pb
11	600 pb	2000 pb
17	350 pb	1000 pb
Total	500 pb	2000 pb

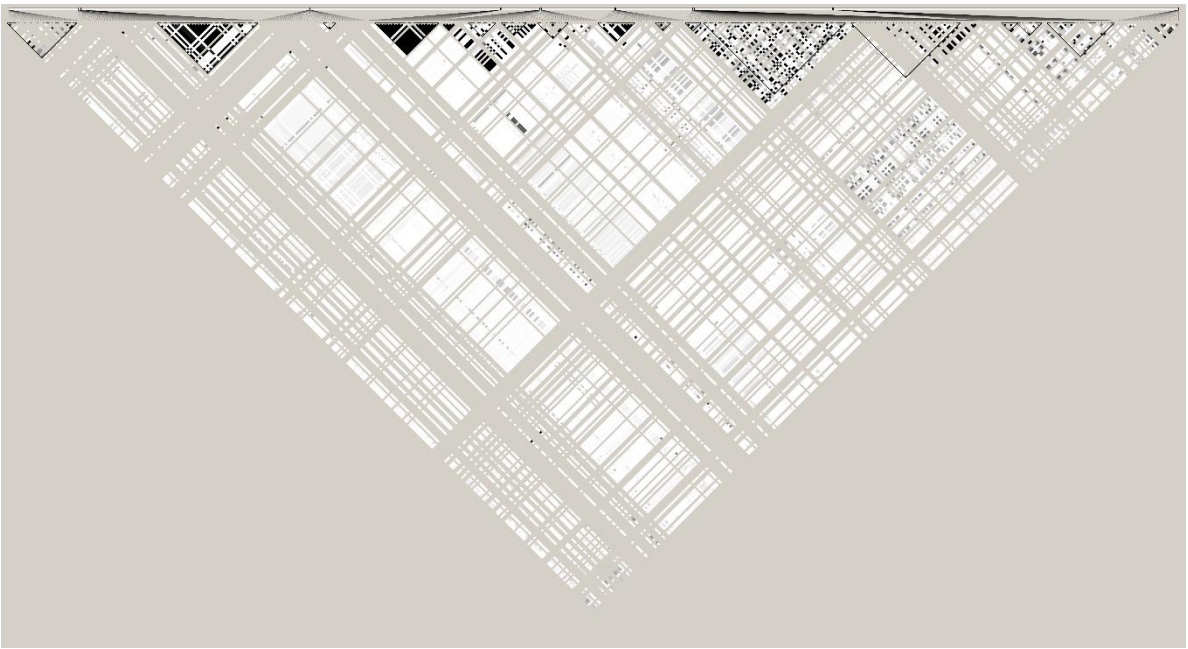


Figure 40 : Représentation graphique des blocs de SNPs en déséquilibre de liaison sur le LG01 à partir des données de re-séquençage sur la CC278

Les données de la puce 480k SNPs ont permis de réévaluer significativement cette distance, estimée à seulement 10 kb. Plusieurs éléments peuvent être invoqués pour expliquer cette grande différence :

(i) la densité de marqueurs sur la puce : les SNPs de la puce 8k sont en effet espacés en moyenne de 140 kb tandis que ceux de la puce 480k le sont seulement de 3 kb ; l'estimation du DL entre des SNPs espacés par de très faibles distances a donc été nettement améliorée en utilisant la puce 480k ; sur la puce 8k, l'estimation de la valeur de 100 kb pour un r^2 de 0,2 était largement sujette à caution au regard du très faible nombre de points disponibles dans la partie de la courbe située entre 0 et 200 kb (Figure 24) ;

(ii) le choix des marqueurs : comme indiqué précédemment, les SNPs répertoriés sur la puce 8k proviennent du re-séquençage à très faible profondeur (~ 1X) de 27 variétés de pommes à couteau notoirement utilisées dans les programmes récents d'amélioration ; ceux de la puce 480k proviennent du re-séquençage beaucoup plus profond (~ 15X en moyenne) d'une large gamme de variétés anciennes européennes (Bianco, Durel, Troggio, et al., in prep.) ; la représentativité et l'adéquation de ces derniers marqueurs au génotypage de populations anciennes (comme les deux core collections étudiées ici) sont donc plus appropriées ; par ailleurs, une place plus importante a été laissée aux SNPs identifiés comme ayant une MAF faible lors du design de la puce 480k, afin de se rapprocher de la forte proportion de SNPs à très faible fréquence observée dans les données de re-séquençage ; cette proportion peut notoirement influencer l'estimation du DL comme indiqué dans le Chapitre 1, 1.2.3 (Lachance and Tishkoff 2013) ;

(iii) les populations génotypées : les individus génotypés dans les deux populations sont en partie différents ; le génotypage avec la puce 8k SNPs a été réalisé sur 48 variétés de pommiers à couteau alors que celui avec la puce 480k a été réalisé sur 268 variétés de pommiers à couteau incluant ceux de la CC48 Couteau; du fait de la faible taille de la CC48, la population génotypée avec la puce 8k totalise un nombre moyen de générations (et donc de recombinaisons) depuis le dernier ancêtre commun entre deux individus supérieur à celui de la CC278 ; cela aurait dû conduire à une estimation plus faible de l'étendue du DL dans la première population (liaison négative entre nombre de recombinaisons et étendue du DL) ; les résultats obtenus sont donc paradoxaux et la différence de population lors du génotypage n'est sans doute pas un facteur important au regard des deux précédents ; cependant, du fait de leurs tailles respectives, la diversité globale explorée dans les deux core collections est plus importante

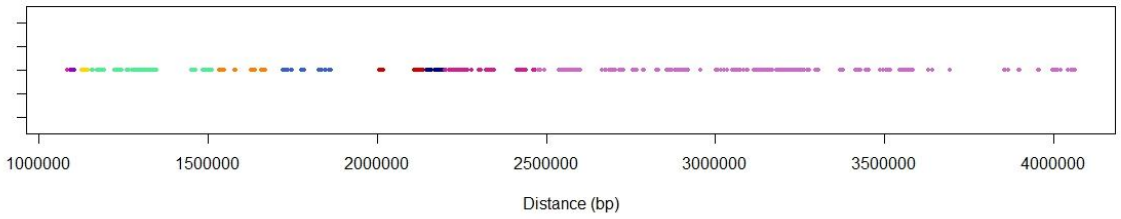


Figure 41 : Distribution des SNPs des différents scaffolds localisés sur le haut du LG16 selon la version 3 du génome du pommier

les différents scaffolds sont représentés par les différentes couleurs

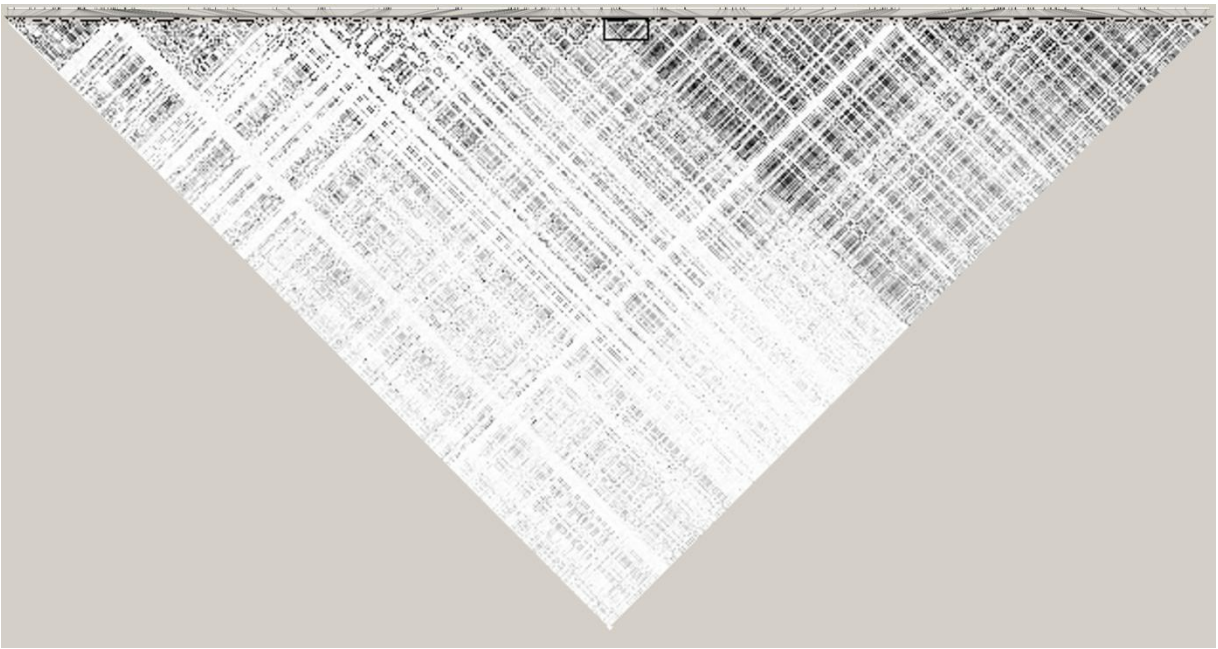


Figure 42 : Représentation graphique des blocs de SNPs en déséquilibre de liaison sur le haut du LG16 sur une distance de 2,7 Mb, après déplacement du scaffold n°4 au niveau du scaffold n°10

dans la seconde que dans la première : cela se mesure par exemple au travers du nombre moyen d'allèles identifiés par marqueur SSR (16,3 dans la CC278 contre 12,8 en moyenne par locus dans la CC48 couteau estimés avec 24 SSR, (Lassois et al. 2015) ; cette différence de diversité a pu impacter l'estimation de l'étendue du DL.

L'effet de la population génotypée sur l'étendue du DL a été étudié en comparant les valeurs de r^2 obtenues pour la CC48 Couteau génotypée avec la puce 480k SNPs (Tableau 8). Les valeurs par LG sont du même ordre de grandeur dans la CC48 et la CC278 ce qui amène à la conclusion que la densité de la puce de génotypage a une grande influence sur l'estimation de l'étendue du DL, contrairement au choix de la population.

L'estimation de l'extension du DL obtenue avec la puce 480k SNPs (10 kb) est donc celle qui a été par la suite considérée lors de la recherche de gènes candidats au voisinage des SNPs identifiés comme étant les plus significativement associés aux caractères étudiés dans l'approche de génétique d'association, présentée dans le Chapitre 5. Il est clair que cette valeur est une estimation potentiellement très imprécise de l'extension du DL sur le génome et que cette estimation varie probablement très fortement d'une région à l'autre du génome, ce qui est illustré par les très grands écarts-types observés autour de la courbe de décroissance du DL (Figure 33).

3.2. Etude d'un cas particulier sur la puce 480k SNPs

En étudiant de plus près les patterns de DL le long des chromosomes, de nombreuses régions sont apparues comme étant en fort DL bien qu'éloignées de plusieurs Mb. Reprenons ici l'exemple du haut du LG16 (Figure 37), identifié comme étant une région d'intérêt dans des analyses ultérieures. L'hypothèse la plus probable dans ce cas de figure est que le positionnement des scaffolds de la région étudiée sur la version 3 du génome ne correspond pas à la réalité. La région de 3 Mb étudiée ici comprend 10 scaffolds (Figure 41), les blocs en fort DL à distance correspondant aux scaffolds numéros 4 et 10 (vert et lilas). Les SNPs localisés sur le scaffold 10 ne sont pas répartis de façon homogène puisqu'un gap d'environ 100 kb est observé aux trois quarts de la longueur du scaffold. Les marqueurs du scaffold 4 ont donc été déplacés au niveau du gap du scaffold 10 pour évaluer si le pattern global de DL ne serait pas amélioré. La figure 42 représente le pattern de DL sur la région de 3 Mb après remaniement des scaffolds 4 et 10. Le pattern du DL observé est beaucoup plus cohérent avec ce qui est attendu lorsque l'on étudie le DL, des blocs de marqueurs présentant un fort DL entre eux et aucun DL entre les marqueurs de différents blocs situés à des distances plus grandes. La région étudiée ici n'est pas un cas isolé sur la version 3 du génome du pommier, comme l'attestent les représentations des différentes régions génomiques présentées dans la partie 2.1.2. Le séquençage du génome a en

effet été réalisé sur un individu hétérozygote ce qui complique l'assemblage des reads en un génome correct. L'estimation moyenne de la décroissance du DL le long du génome en utilisant les données de puce 480k SNPs est donc à considérer avec beaucoup de précautions, les distances entre marqueurs pouvant être erronées. Une visualisation détaillée de la totalité du génome serait nécessaire de manière à retravailler la carte physique. Les analyses de cette thèse, notamment la comparaison de la localisation des SNPs identifiés comme significativement associés avec un caractère phénotypique et des QTLs identifiés pour ce même caractère, ont été réalisées en parallèle sur les données des versions 1 et 3 du génome, grâce aux contigs ayant persisté lors du passage à la version 3.

3.3. Déséquilibre de liaison à fine échelle

L'étendue du DL intra- et inter-génique présente des variations entre les différentes régions du génome étudiées. Le DL intra-génique s'étend en effet sur près de 1 000 pb dans certaines régions comme le LG01 et sur seulement 250 pb dans d'autres régions, comme le LG10. Il en est de même pour le DL inter-génique qui s'étend de 1 000 pb pour les LG07, LG10 et LG17, et jusqu'à 5 000 pb pour le LG01. Il convient tout d'abord de rappeler que ces valeurs ont été estimées visuellement donc une incertitude non négligeable existe, en particulier pour le DL inter-génique. En considérant que ces valeurs sont fiables, ces différences observées peuvent s'expliquer par le fait que l'étendue du DL n'est pas homogène dans toutes les régions du génome et est fortement impactée par la présence de « hotspots » de recombinaison (Kim et al. 2007), et par la localisation des gènes étudiés, à proximité ou non des centromères (régions de faible recombinaison ; Comadran et al. 2011). Les valeurs calculées ici sont consistantes avec la population étudiée et le mode de reproduction de l'espèce, à savoir une core collection d'individus appartenant à une espèce allogame. La distance sur laquelle s'étend le DL est en effet relativement faible par rapport aux valeurs observées chez des espèces autogames telles qu'*Arabidopsis thaliana* ou le maïs *Zea mays* (Nordborg et al. 2002; Stich et al. 2005), mais comparable à celles estimées chez des espèces allogames telles que le peuplier tremble *Populus tremula* (Ingvarsson 2005), le pin à torches *Pinus taeda* (Brown et al. 2004) et l'épicéa commun *Picea abies* (Heuertz et al. 2006). De plus, les gènes sélectionnés ici sont des candidats positionnels et leurs fonctions n'en font pas nécessairement des gènes soumis à sélection. Celle-ci aurait eu pour effet d'augmenter l'étendue du déséquilibre de liaison sur des distances plus grandes que celles observées ici, comme observé chez Palaisa et al. (2003).

Les différentes régions étudiées ne semblent pas présenter des niveaux différents de diversité haplotypique au sein des gènes re-séquencés, les régions présentant toutes des fragments possédant un nombre élevé et un faible nombre d'haplotypes. Cela peut laisser penser que les six régions étudiées n'ont pas été soumises à sélection, auquel cas des niveaux très élevés de diversité (cas d'un événement

de sélection diversifiante) ou au contraire très faibles (cas d'un événement de sélection négative) auraient été observés.

Au final, les éléments les plus robustes à l'intérieur du génome, que ce soit la version 1 ou la version 3, sont les contigs. Il aurait donc été préférable d'estimer la décroissance du DL en fonction de la distance physique en se limitant aux SNPs présents sur le même contig.

La cartographie génétique fine des contigs permettrait d'améliorer leur positionnement relatif au sein des scaffolds mais cela nécessiterait des tailles de populations de cartographie très élevées afin que la résolution de la cartographie génétique soit suffisante. Cela serait malheureusement difficilement réalisable compte-tenu du prix de génotypage d'un individu avec la puce 480k SNPs (~ 160€). La visualisation fine du DL entre les SNPs d'une même région génomique telle que présentée ci-dessus serait une solution pour résoudre localement les erreurs de positionnement des contigs, mais très longue à mettre en place au niveau du génome entier.

En définitive, seuls un re-séquençage, en particulier de fragments de grande longueur, du génome d'un individu homozygote de pommier et un réassemblage *de novo* semblent être une solution efficace pour améliorer à court terme la carte physique du génome du pommier. Ce travail est engagé à FEM sur un haploïde doublé de la variété « Golden Delicious » sélectionné à l'INRA d'Angers il y a plus de 20 ans.

3.4. Comparaison des données de puce et de données de re-séquençage pour l'estimation du déséquilibre de liaison

Deux techniques de génotypage ont été utilisées sur les mêmes individus de la CC278, à quelques individus près. L'estimation de l'étendue du DL grâce aux données résultant de ces deux techniques de génotypage peut donc être comparée. Le DL moyen prend des valeurs de r^2 inférieures à 0,2 au-delà de 10 kb selon les données de génotypage de la puce 480K SNPs, et au-delà d'environ 750 pb selon les données de re-séquençage de gènes. Cette différence d'un facteur 10 ou plus peut être expliquée par plusieurs facteurs. La première explication dépend à nouveau de la densité de marqueurs utilisés. Les marqueurs de la puce 480k SNPs sont espacés en moyenne de 3 kb alors que les SNPs détectés dans les données de re-séquençage sont espacés en moyenne de 25 pb. L'utilisation des données de re-séquençage permet donc en toute logique une estimation plus précise de l'étendue du DL pour de très faibles distances physiques. La seconde explication réside dans le choix des marqueurs de la puce de génotypage. De même que pour la 8k SNP array v1 (Chagne, Crowhurst, et al. 2012), les SNPs de la puce 480k sont issus de données de re-séquençage d'une soixantaine de génomes de variétés de pommiers à couteau et les SNPs ayant une MAF trop faible n'ont majoritairement pas été inclus sur la puce. De ce fait, le spectre de distribution des fréquences alléliques

est biaisé vers les fortes valeurs, et, en conséquence, l'estimation du DL l'est également. Il a en effet été montré que la distance sur laquelle décroit le DL est surestimée dans les données provenant d'un génotypage réalisé sur une puce SNP en comparaison de celles issues d'un génotypage réalisé par re-séquençage (Lachance and Tishkoff 2013). Lors de la réalisation d'une étude d'association avec des données de puce SNP, l'estimation de la distance sur laquelle s'étend le DL, souvent réalisée sur les mêmes données de génotypage, sera donc biaisée puisque la plupart des puces de génotypage SNP sont construites de la même façon, avec les mêmes critères de sélection des SNPs. La distance estimée avec ces données de puce sera donc plus élevée qu'une distance estimée avec des données de re-séquençage. Il ne faut cependant pas oublier que cette distance est calculée sur l'ensemble du génome et n'est donc pas représentative de l'étendue du DL dans une zone spécifique. Il a été montré par ailleurs que le DL de liaison n'est pas homogène dans toutes les régions du génome (Drouaud et al. 2006; Comadran et al. 2011) et qu'un événement de sélection par exemple peut entrainer une augmentation de la distance sur laquelle il s'étend (Palaisa et al. 2003). Enfin, lors du design de la puce 480k, lorsque plusieurs marqueurs étaient en trop fort DL ($r^2 > 0,85$), il a été décidé de ne garder qu'un tag-SNP (un SNP choisi pour représenter un groupe de SNPs ou haplotype en fort DL), représentatif du groupe selon une démarche similaire à celle du logiciel Tagger (de Bakker et al. 2005 ; Durel, communication personnelle). Le DL à l'échelle du génome entier a donc été cassé et cela pourrait expliquer le pattern inhabituel de DL observé ici.

La recherche de gènes candidats dans une fenêtre de distance égale à deux fois la distance de décroissance du DL (de part et d'autre du marqueur associé) est donc à réaliser avec précaution en gardant à l'esprit qu'en fonction du DL local, le locus causal peut se trouver au-delà des limites fixées.

3.5. Discussion générale

Il est important de bien garder en mémoire que ces estimations d'étendue du DL pour une valeur fixée de $r^2 = 0,2$ sont toutes entachées d'une très forte variation qui n'est que l'expression du fait que le DL entre deux marqueurs est la résultante de l'histoire des recombinaisons entre ces deux loci, et de l'histoire des mutations à l'origine du polymorphisme observé à ces deux loci (Nordborg and Tavare 2002). Si une mutation est apparue beaucoup plus tardivement que l'autre dans l'arbre généalogique global de la population étudiée, l'ancêtre commun le plus récent de chaque mutation sera situé à des niveaux très différents de cet arbre généalogique, ce qui entrainera un DL très différent d'une situation où deux mutations apparaîtraient à des moments proches, donc avec un nombre de générations voisin depuis l'ancêtre commun le plus récent. Dans une situation théorique où tous les polymorphismes seraient apparus en une seule fois à un seul niveau de l'arbre généalogique, le DL serait principalement lié à la recombinaison, donc au nombre de générations depuis l'ancêtre commun

le plus récent ; il traduirait alors clairement la distance physique entre les loci, comme dans le cas d'une descendance de cartographie ou d'un pedigree connu. Le fait que les mutations puissent apparaître régulièrement à des moments différents le long de l'arbre généalogique vient largement perturber la liaison DL – distance physique et génère une grande variation de situations de DL pour une même distance physique entre deux loci, d'où la très grande variance observée sur ce paramètre, quelle que soit la distance physique entre les deux marqueurs. D'autres évènements évolutifs, comme la démographie (expansion, goulot d'étranglement) ou la dérive suivie d'admixture, viennent eux aussi flouter davantage cette liaison.

Les valeurs observées ici, que ce soit en utilisant les données de puce SNP ou les données de re-séquençage de gènes, sont consistantes avec la population et le modèle étudié. Globalement, on peut considérer que la distance sur laquelle s'étend le DL est faible chez le pommier, avec une valeur moyenne de 10 kb selon les données de puce et de 1 à 4 kb selon les données de re-séquençage. Lors de l'analyse d'association, la distance de 10 kb de part et d'autre des marqueurs présentant une association significative avec le génotype sera utilisée.

Les mêmes analyses de génétique des populations basées sur le SFS que dans le Chapitre 3 peuvent être réalisées sur les données de re-séquençage. La distribution des MAF sur ces données correspond en effet à l'attendu qui est d'observer une majorité de marqueurs avec une MAF très faible (< 0,05). Des pommiers sauvages ont également été re-séquencés lors de ce projet et pourront servir à déterminer l'état ancestral des allèles identifiés pour tracer le SFS. Après vérification que le SFS ne présente pas de biais comme pouvaient en présenter les données de génotypage issues des deux puces, des estimateurs tels que le D de Tajima, le F de Fu et Li, le H de Fay et Wu, et même le dN/dS pour les fragments effectivement localisés dans des gènes, pourront être calculés et permettre de déterminer si les régions dans lesquelles sont localisées les gènes ont été soumises à sélection. Le travail de thèse s'est concentré sur les analyses de GWAS, pour lesquelles les données sont arrivées tardivement, et ces analyses de génétique des populations n'ont donc pas encore été réalisées.

Chapitre 5

Génétique d'association dans une core collection de pommiers à couteau

Tableau 11 : Valeurs des BIC calculés pour les analyses de variance sur les différents caractères phénotypiques en utilisant des modèles prenant en compte différents effets

	BIC modèle 1[a]	BIC modèle 2[b]	BIC modèle 3[c]
Résistance EUB04	**10102,59**	10106,86	**-**
Résistance 104	11822,8	**11815,47**	**-**
Résistance mélange de souches	7811,069	**7804,958**	**-**
Résistance feu bactérien	**31609,85**	31790,32	**-**
Acidité	3270,197	**3247,173**	3338,13
Ratio sucre/acidité	3298,16	**3253,14**	3360,74
Croquant	3155,47	**3150,29**	3314,57
Fermeté	3205	**3195,94**	3420,78
Teneur en fibres	**3469,46**	3482,94	3736,19
Granulosité	**3210,02**	3235,86	3464,19
Jutosité	**3127,94**	3132,45	3270,78
Fondant	**3323,85**	3334,01	3637,4
Farinosité	2428,01	**2384,02**	3051,28
Russeting	2897,53	**2840,79**	3349,73
Sucre	3030,15	**3012,14**	3098,36
Goût	3445,3	**3414,48**	3480,55

[a] pour les tests de résistance : Trait ~ 1 + Inoc + (1|Clone) + (1|Clone:Inoc) ; pour les tests de qualité du fruit : Trait ~ 1 + Année + Année : Testeur + (1|Clone) + (1|Clone:Année)
[b] pour les tests de résistance : Trait ~ 1 + Inoc + (1|Clone) ; pour les tests de qualité du fruit : Trait ~ 1 + Année + (1|Clone) + (1|Clone:Année)
[c] pour les tests de qualité du fruit : Trait ~ 1 + Année + (1|Clone)

1. Introduction

La recherche de gènes candidats associés aux caractères phénotypiques d'intérêt agronomique par des approches de cartographie en descendances, bien qu'efficace, ne permet d'explorer qu'une infime partie de la diversité génétique disponible. Chez le pommier, une très grande diversité génétique est disponible, dénotée par le très grand nombre de variétés et de phénotypes différents, mais également par les études de diversité menées dans l'équipe ResPom de l'INRA d'Angers. Le séquençage du génome et le développement récent d'une puce de génotypage haut débit font du pommier un bon candidat pour une étude de génétique d'association sur génome entier. Nous disposions ici d'une core collection de variétés anciennes de pommiers à couteau ainsi que des données de génotypage sur une puce 480k SNPs pour ces variétés. A cela s'ajoute un éventail de données phénotypiques concernant la résistance du pommier à deux maladies, la tavelure et le feu bactérien, et la qualité du fruit avec des données sensorielles acquises par un panel d'experts. Ce dernier chapitre porte donc sur l'étude de ces données de phénotypage et sur l'analyse des résultats de l'étude d'association menée dans le but d'identifier de nouvelles sources génétiques de résistance aux maladies et de qualité du fruit chez le pommier.

2. Résultats

2.1. Analyses statistiques des données phénotypiques

2.1.1. Résultats des tests de résistances

Au total, trois tests de résistance à la tavelure et un test de résistance au feu bactérien ont été réalisés. Les analyses de variance effectuées sur les AUDPC (tavelure) ou sur le pourcentage de tige nécrosée (feu bactérien) montrent un effet significatif à la fois du génotype et des différentes modalités de test (date d'inoculation et/ou position dans le bloc), et pour deux tests (résultats des tests avec la souche EU-B04 de *Venturia inaequalis* et avec *Erwinia amylovora*), une interaction significative entre ces deux facteurs. Le modèle le plus vraisemblable a été choisi selon les valeurs de BIC calculés pour chaque modèle (Tableau 11).

Lors du test de résistance à la tavelure effectué au printemps 2012, la majorité des génotypes présentaient les symptômes d'une forte sensibilité à la souche EU-B04 (race 1,10), comme l'indique la forte valeur moyenne d'AUDPC de 20,39 observée pour les génotypes communs aux trois tests. La

Tableau 12 : Héritabilités au sens large (H^2) et au sens strict (h^2) calculées respectivement suite aux ANOVA sur les données phénotypiques de résistance et aux analyses d'association

Test de résistance	H^2	h^2
EU-B04	0,45	0,33
104	0,79	0,68
Mélange de souches	0,51	0,46
CFBP1430	0,86	0,58

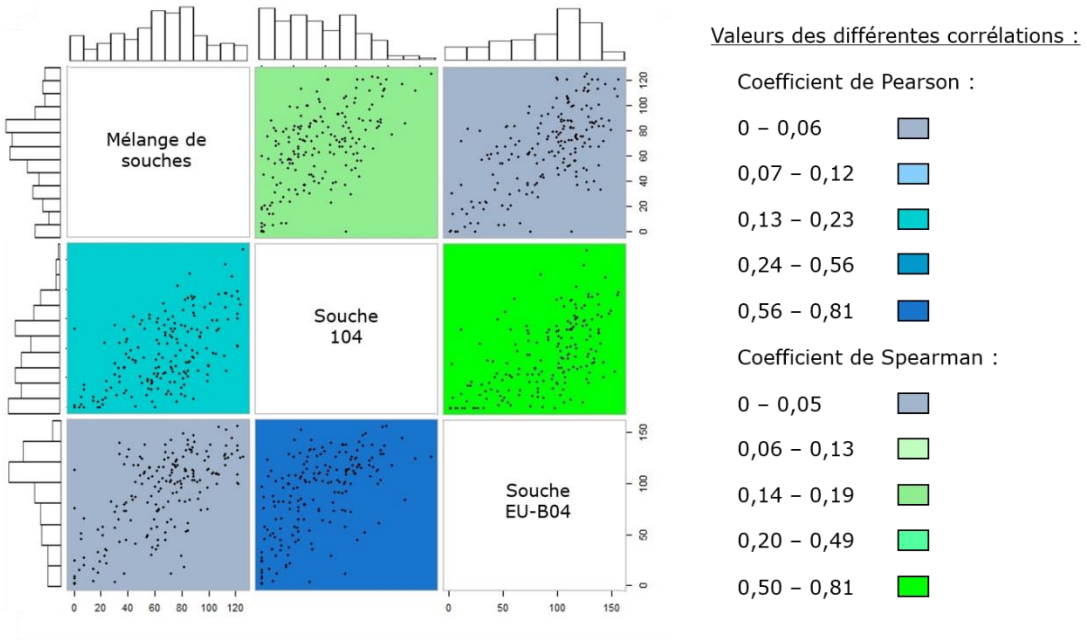


Figure 43 : Représentation graphique des corrélations de Pearson et Spearman calculées entre les valeurs moyennes d'AUDPC, et des distributions de ces valeurs pour les trois tests de résistance à la tavelure

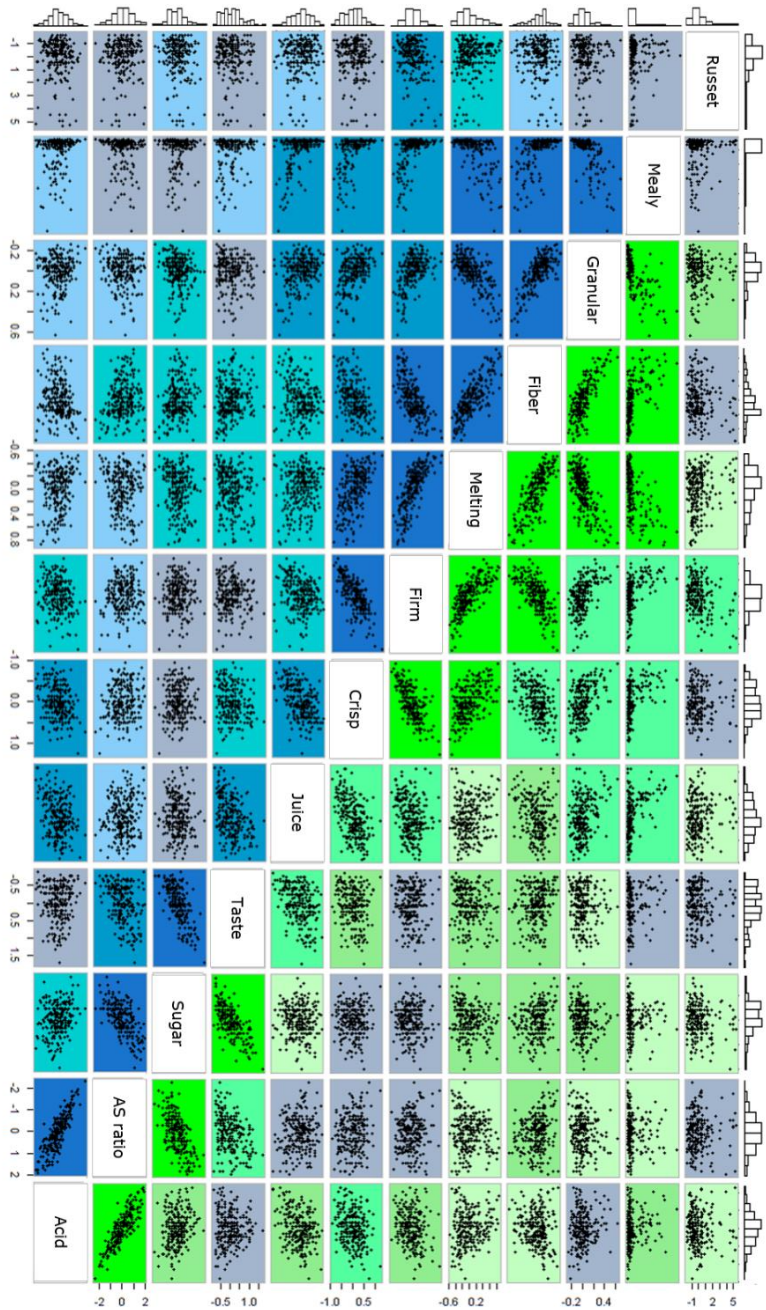
distribution des moyennes ajustées du test de l'automne 2013 réalisé avec la souche 104 (race 1) montre, inversement au test de 2012, une distribution centrée sur les faibles valeurs, dénotant une plus grande résistance des génotypes vis-à-vis de cette souche (valeur moyenne d'AUDPC de 15,27 pour les génotypes communs aux trois tests). Lors du test réalisé au printemps 2013 en utilisant le mélange de souches, la distribution des moyennes ajustées est similaire à une distribution normale avec peu d'individus très résistants et très sensibles et une majorité de phénotypes intermédiaires. La valeur moyenne d'AUDPC pour les génotypes communs aux trois tests était de 20,21 (Figure 43).

Les héritabilités au sens large (H^2) ont été calculées pour les trois tests de résistance à la tavelure et le test de résistance au feu bactérien (Tableau 12) en utilisant la formule adaptée selon le modèle d'analyse de variance choisi. Les valeurs d'héritabilité obtenues sont de moyennes à fortes (de 0,45 à 0,86), la valeur la plus faible étant obtenue pour le test avec la souche EU-B04, et la valeur la plus forte avec la souche 104, en ce qui concerne la résistance à la tavelure.

Afin de corriger les différents effets « bloc » ou « date d'inoculation », les BLUP des valeurs génotypiques propres à chaque test (BLUP_G par test) les prenant en compte ont été calculés. Le modèle utilisé a été vérifié en réalisant les graphiques des projections de la valeur génotypique estimée avec le BLUP contre la valeur réelle des phénotypes (pour tous les caractères on observe une forte corrélation entre les deux valeurs), ainsi qu'en vérifiant la normalité de la distribution des résidus du modèle (Annexe 11).

Pour les génotypes communs aux trois tests tavelure, les valeurs de corrélation observées entre les valeurs d'AUDPC moyennes des différents tests révèlent une liaison plus élevée entre les AUDPC observées pour la souche EU-B04 et la souche 104 que pour les deux autres comparaisons impliquant le mélange de souches (Figure 43). En marge de ces corrélations, l'observation des graphiques de distribution des valeurs d'AUDPC pour les souches prises deux à deux révèle de nombreuses situations d'interactions spécifiques entre génotypes et souches testées. C'est en particulier le cas pour la comparaison entre les deux souches monoconidiales EU-B04 et 104, pour lesquelles on observe de nombreux génotypes avec une AUDPC nulle ou très faible vis-à-vis de la souche 104 alors que leur AUDPC est moyenne à élevée vis-à-vis de la souche EU-B04. Une liste d'une dizaine génotypes présentant des AUDPC très faibles à nulles vis-à-vis des 3 inocula a pu être dressée.

Le test de résistance vis-à-vis du feu bactérien a montré une réponse majoritaire de sensibilité avec près d'un quart des individus présentant une nécrose sur plus de 90% de la tige. L'héritabilité calculée pour ce test est très forte (0,86), en accord avec le nombre moyen élevé de copies phénotypées par génotype (6,2).



Valeurs des différentes corrélations :

Coefficient de Pearson :

- 0 – 0,06
- 0,07 – 0,12
- 0,13 – 0,23
- 0,24 – 0,56
- 0,56 – 0,81

Coefficient de Spearman :

- 0 – 0,05
- 0,06 – 0,13
- 0,14 – 0,19
- 0,20 – 0,49
- 0,50 – 0,81

Figure 44 : Représentation graphique des corrélations de Pearson et Spearman calculées entre les BLUP des différents caractères de qualité du fruit, et des distributions de ces valeurs pour les trois tests de résistance à la tavelure

2.1.2. Résultats des notations de qualité du fruit

Les analyses de variance effectuées sur les données de qualité du fruit montrent un effet significatif du génotype, de l'année de notation, de l'interaction entre ces deux facteurs et, pour certains caractères, du facteur « notateur dans année ». Le modèle le plus vraisemblable a été choisi selon les valeurs de BIC calculés pour chaque modèle (Tableau 11).

Afin de tenir compte de l'effet année et éventuellement de l'effet notateur, les BLUP des valeurs génotypiques le(s) prenant en compte ont été calculés (BLUP_G). Les BLUP des valeurs génotypiques propres à chaque année (BLUP_G2012, BLUP_G2013 et BLUP_G2014) ont également été calculés. Le modèle utilisé a été vérifié en réalisant les graphiques des projections de la valeur génotypique estimée avec le BLUP_G contre la valeur réelle des phénotypes (pour tous les caractères on observe une forte corrélation entre les deux paramètres), ainsi qu'en vérifiant la normalité de la distribution des résidus du modèle (Annexe 12). La majorité des caractères présente une distribution globalement normale des BLUP, à l'exception du caractère farinosité, et dans une moindre mesure des caractères fondant et russeting (Figure 44). Ces trois caractères ont par la suite été considérés avec précaution du fait de ces distributions non normales.

Afin d'évaluer l'ampleur relative des effets d'interactions entre les facteurs génotype et année de notation, le rapport de la variance d'interaction sur la somme des variances génétique et d'interaction a été calculé pour chacun des caractères (Tableau 13). Les valeurs de ces rapports sont comprises entre 0,28 pour le russeting et 0,96 pour la farinosité, traduisant un très fort effet de l'interaction pour ce dernier caractère. Les valeurs génotypiques annuelles des individus communs aux trois années de notation ont également été représentées sur un graphique afin de visualiser ces interactions (Annexe 13). Pour la plupart des caractères, on observe un déclassement important des individus entre les années de notation ainsi qu'un changement d'échelle, ce qui participe à l'interaction, mais l'ampleur du déclassement et du changement d'échelle est variable d'un caractère à l'autre. De ces résultats dépendent les résultats du test d'association. Il est en effet peu probable que des résultats significatifs soient trouvés avec les BLUP_G (interannuels) pour les caractères présentant une forte interaction génotype-année. Pour ces caractères, les BLUP annuels seront plus représentatifs d'un comportement génétique propre à chaque année.

Les héritabilités au sens large (H^2) ont été calculées pour les douze caractères sur les trois années de notation prises en compte simultanément, puis séparément (Tableau 13), en utilisant la formule adaptée selon le modèle d'analyse de variance choisi. Parmi les valeurs d'héritabilité obtenues, celles des caractères perçus en premier par le consommateur tels que l'acidité et le ratio sucre/acidité sont fortes (~ 0,7 à 0,8) et celle du sucre est moyenne (~ 0,4). L'héritabilité du goût global, lié aux caractères précédents, apparaît relativement élevée (> 0,5). Concernant les caractères liés à la texture du fruit (croquant, fermeté et jutosité), les héritabilités sont moyennes (~ 0,5). Les caractères détectés

Tableau 13 : Héritabilités au sens large (H^2) et au sens strict (h^2) calculées respectivement suite aux ANOVA sur les données phénotypiques de qualité du fruit et aux analyses d'association les héritabilités ont été estimées soit sur l'ensemble des 3 années de notation, soit année par année

Caractère évalué	Int.[a]	H^2	h^2	H^2 2012	h^2 2012	H^2 2013	h^2 2013	H^2 2014	h^2 2014
Acidité	0,32	0,76	0,58	0,88	0,33	0,84	0,50	0,87	0,56
Ratio sucre/acidité	0,39	0,72	0,59	0,85	0,51	0,84	0,69	0,85	0,14
Croquant	0,67	0,48	0,55	0,86	0	0,82	0,52	0,71	0,57
Fermeté	0,66	0,52	0,65	0,87	0	0,86	0,57	0,86	0,72
Teneur en fibres	0,85	0,28	0,20	0,88	0	0,85	0,44	0,81	0,03
Granulosité	0,89	0,20	0	0,94	0	0,80	0,10	0,78	0
Jutosité	0,65	0,50	0,53	0,85	0,48	0,82	0	0,73	0,53
Fondant	0,81	0,35	0,18	0,88	0	0,89	0,35	0,87	0,21
Farinosité	0,96	0,08	0,51	0,90	0,02	0,96	0,41	0,93	0,99
Russeting	0,28	0,86	0,86	0,97	0,92	0,97	0,79	0,98	0,94
Sucre	0,66	0,44	0	0,76	0,12	0,72	0	0,74	0
Goût	0,54	0,54	0	0,67	0,02	0,79	0	0,67	0

[a] rapport de la variance d'interaction sur la somme des variances génétique et d'interaction

Tableau 14 : Valeurs des BIC calculés pour les analyses d'association sur les différents caractères en utilisant un modèle ne prenant en compte que l'apparentement ou un modèle prenant en compte la structure et l'apparentement

Caractère	BIC modèle K	BIC modèle K + Q
Résistance EUB04	2450.98	**1323.01**
Résistance 104	2290.27	**1174.09**
Résistance mélange de souches	2048.12	**1582.07**
Résistance feu bactérien	2383.27	**2214.75**
Acidité	841.60	**670.38**
Ratio sucre/acidité	836.22	**623.52**
Croquant	773.71	**252.17**
Fermeté	782.27	**353.57**
Teneur en fibres	830.16	**22.52**
Granulosité	827.67	**33.52**
Jutosité	732.92	**277.31**
Fondant	865.77	**132.53**
Farinosité	683.26	**-750.15**
Russeting	935.00	**835.85**
Sucre	667.07	**124.37**
Goût	782.78	**353.21**

en second lors de la dégustation des fruits (granulosité, fondant et teneur en fibres) présentent quant à eux des héritabilités plus faibles (~ 0,2 à 0,35). Le caractère farinosité présente une héritabilité très faible (< 0,1) en lien avec la distribution très peu variable de ce caractère. Enfin, le caractère russeting, évalué visuellement sur l'extérieur du fruit, présente une héritabilité très élevée (~ 0,87).

Des corrélations ont été calculées entre les BLUP_G des différents caractères et sont présentées dans la Figure 44. Des corrélations élevées entre les caractères acidité et ratio sucre/acidité (0,83), les caractères sucre, ratio sucre/acidité et goût (de 0,6 à 0,7) et entre les caractères croquant, fermeté, fondant, granulosité, teneur en fibres et farinosité (de 0,45 à 0,80) ont été trouvées. On peut également noter des corrélations assez élevées entre les caractères acidité et jutosité (0,28) ou croquant (0,25), ces deux derniers étant aussi fortement corrélés entre eux (0,51).

2.2. Analyses d'association

Le modèle le plus approprié pour les analyses d'association a été choisi en calculant le BIC pour chacun des tests, lorsque seuls les effets de l'apparentement étaient pris en compte, ou lorsque les effets de structure et d'apparentement étaient pris en compte simultanément. Les valeurs des BIC pour chacun des tests sont reportées dans le Tableau 14. Pour toutes les analyses, que ce soit pour les caractères de résistance ou de qualité du fruit, le modèle ayant le BIC le plus faible était celui prenant en compte les effets de structure et d'apparentement. C'est donc ce modèle qui a été utilisé pour les analyses d'association.

2.2.1. Résistances à la tavelure et au feu bactérien

Les associations entre les valeurs génotypiques des individus (BLUP) et les génotypes aux SNPs disponibles grâce à la puce 480k ont été testées sur les données des quatre tests de résistance. De manière surprenante, seuls quelques SNPs ayant un effet faible (*p*-value proche de la valeur seuil choisie, ici *p*-value = 10^{-5}) ont pu être identifiés. D'autres SNPs, pour lesquels la *p*-value était supérieure à 10^{-5} ($-\log(p) < 5$), présentaient cependant des profils intéressants. Certaines régions génomiques abritaient en effet des « pics » formés par plusieurs SNPs très regroupés présentant des *p*-values décroissantes puis croissantes (avec un SNP central ayant une valeur minimale de *p*-value). En fixant un seuil de significativité correspondant à une *p*-value ~ 3.10^{-5} (Figure 45 et Annexe 14), la plupart de ces groupes de SNPs deviennent significatifs. Il a été choisi ici de n'étudier que les SNPs les plus significatifs de ces pics de SNPs regroupés avec des significativités décroissantes. Au total, huit régions du génome ont été identifiées. Parmi celles-ci on retrouve cinq régions pour les tests de résistance à la

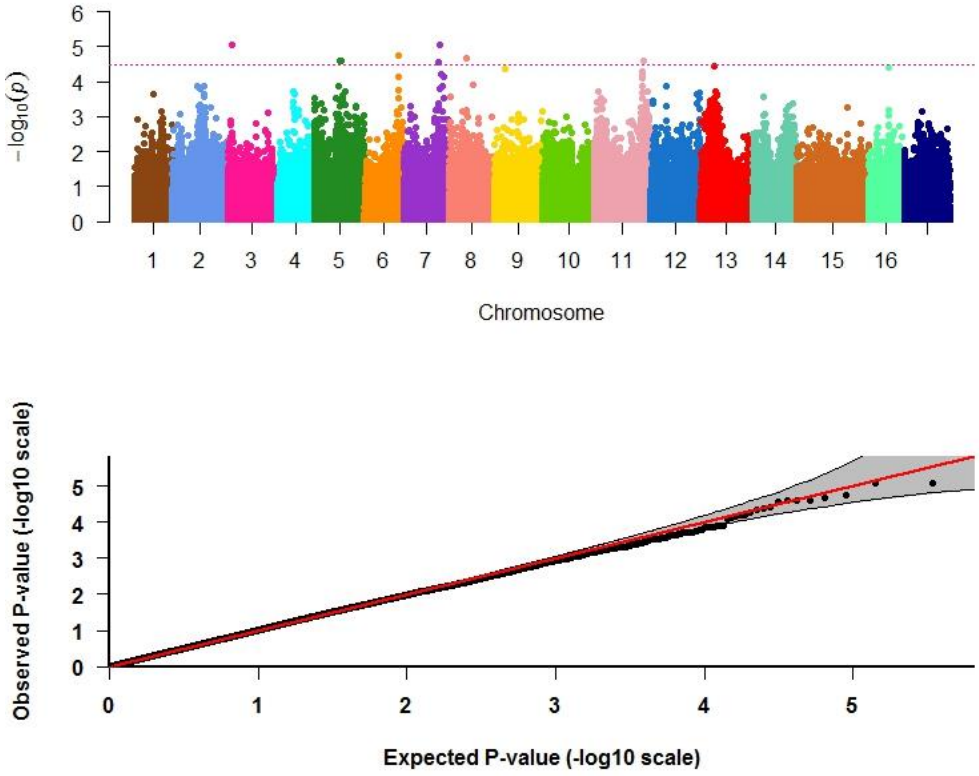


Figure 45 : Manhattan plot et QQ plot des résultats de l'analyse d'association réalisée sur les données de résistance au feu bactérien en corrigeant de la structure et de l'apparentement

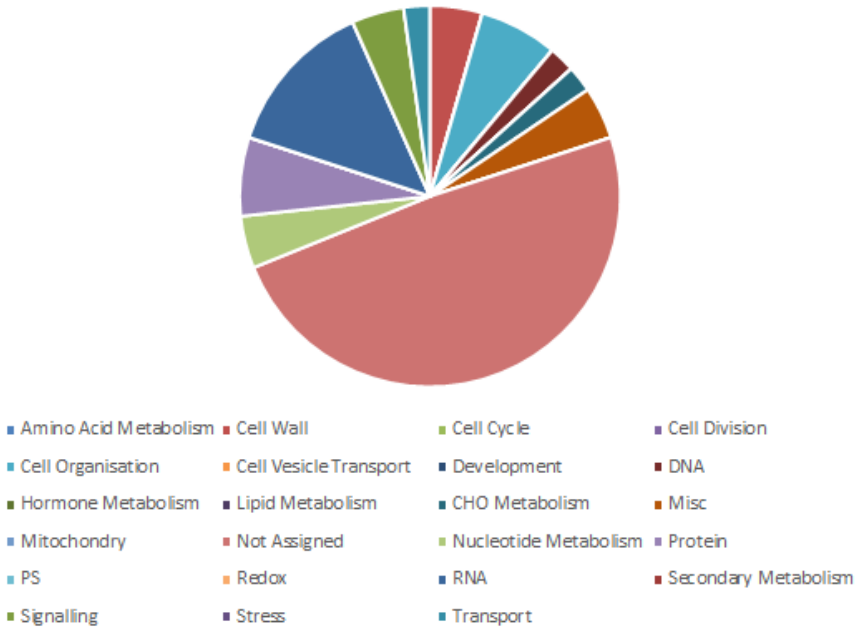


Figure 46 : Distribution des gènes candidats identifiés au voisinage des associations significatives lors de l'étude de génétique d'association portant sur les caractères de résistance à la tavelure et au feu bactérien dans les principales classes de gènes

tavelure, localisées sur les LG01 et LG02 à 11,6 Mb et 22,1 Mb respectivement (souche 104), sur le LG15 à 30,5 Mb (souche EU-B04) et sur les LG08 et LG13 à 2,2 Mb et 18,8 Mb respectivement (mélange de souches). Pour les tests d'association avec la résistance au feu bactérien, trois régions génomiques localisées à 27 Mb sur le LG06, 27,8 Mb sur le LG07 et 38,5 Mb sur le LG11 ont été localisées. En complément de ces détections d'associations significatives, les héritabilités au sens strict (h^2 = « chip heritability » ; Speed et al. 2012) ont été calculées pour chacun des tests (Tableau 12). Les valeurs obtenues pour les tests tavelure, comprises entre 0,33 et 0,68, montrent des héritabilités moyennes à fortes. L'héritabilité calculée pour le test de résistance au feu bactérien est de 0,58 et peut être considérée comme forte.

Parmi les 60 gènes répertoriés dans des fenêtres de +/- 10kb autour des SNPs significativement associés avec les caractères de résistance, 31 n'ont pas de fonction prédite ou ne s'alignent avec aucune séquence de la base de données. Les autres gènes sont répartis dans différentes classes fonctionnelles telles que l'organisation cellulaire et les processus impliquant l'ARN (Figure 46). On note cependant la présence d'un homologue du gène *NPR1*, caractérisé chez *Arabidopsis thaliana* et différentes autres espèces de plantes pour son implication dans l'activation de gènes de défense, identifié lors de l'analyse sur les données de résistance au feu bactérien.

2.2.2. Caractères de qualité du fruit

Concernant les données de qualité du fruit, les résultats des analyses d'association entre les valeurs génotypiques des individus sur les trois années (BLUP_G et BLUP annuels) et les génotypes aux SNPs montrent, pour certains caractères, de fortes associations (Figure 47 et Annexe 15). Il a été choisi ici de n'étudier que les SNPs les plus significatifs des pics composés de plusieurs SNPs avec des significativités décroissantes.

Concernant les caractères ayant attrait au goût des fruits, l'analyse interannuelle sur le caractère sucre ne détecte pas d'association significative. On observe cependant des SNPs significatifs lors des analyses sur les années prises séparément, même si les diagrammes QQplots incitent à une grande prudence (Annexe 15): quatre régions localisées à 32,8 Mb et 6,7 Mb sur les LG05 et LG12 en 2012 et à 38,3 Mb sur le LG05 et 51,1 Mb sur le LG15 en 2014 exhibent ainsi chacune un pic de SNPs avec des *p*-values décroissantes puis croissantes en 2012. Les résultats du caractère acidité, qui présente les plus faibles *p*-values, montrent des associations significatives pour des SNPs localisés à 14,6 Mb sur le LG08, à 30,6 Mb sur le LG14 et à 1,2 Mb et 3,8 Mb sur le LG16 dans l'analyse interannuelle. Les résultats des associations par année confirment ces quatre régions et permettent l'identification de trois régions supplémentaires, sur le bas du LG04 (24,6 Mb) et le haut du LG05 (6 Mb) en 2012 et au milieu du LG13 (28,2 Mb) en 2014. Les résultats pour le caractère ratio sucre/acidité

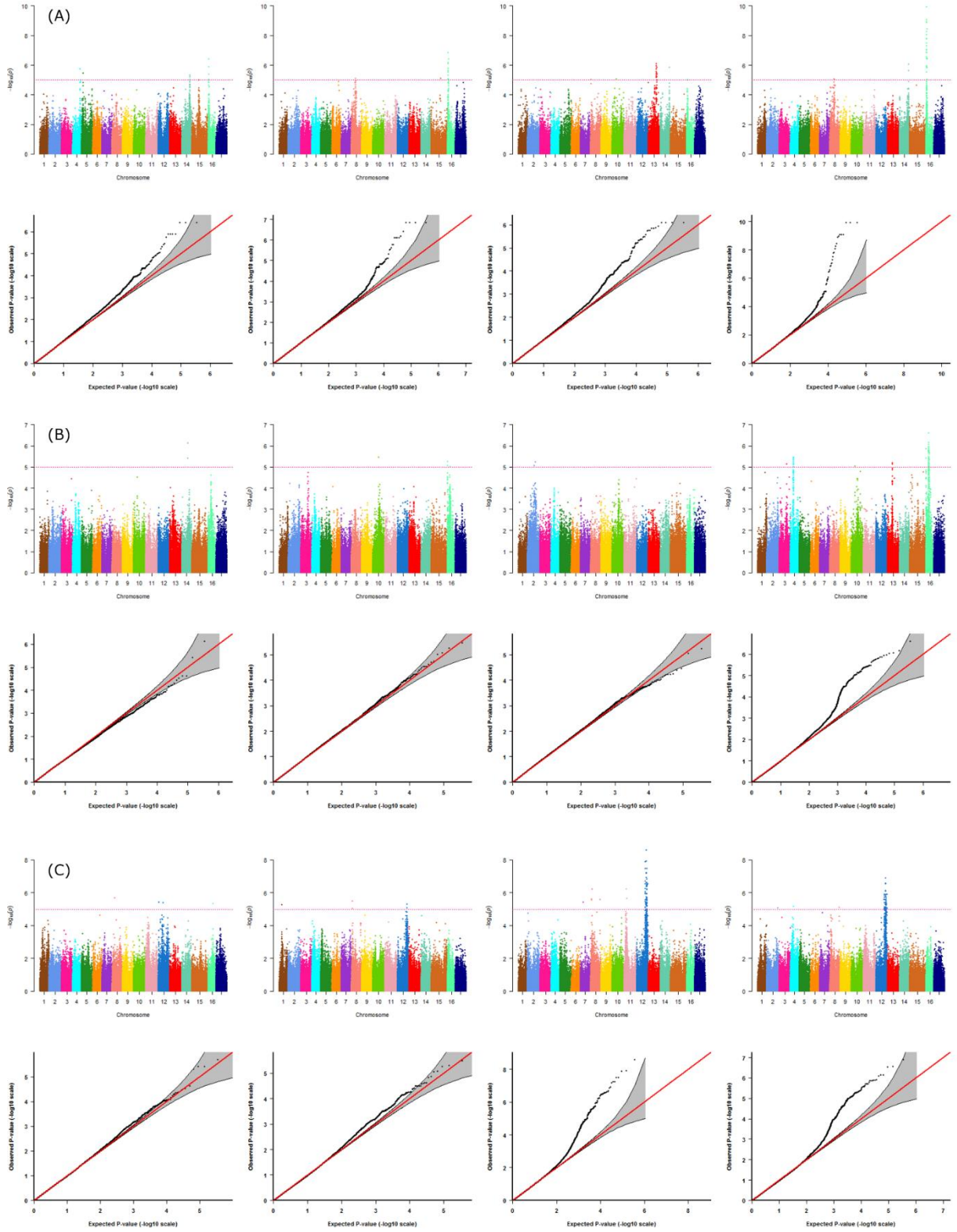


Figure 47 : Manhattan plots et QQ plots des résultats des analyses d'association réalisée sur les données d'acidité (A), de fermeté (B) et de russeting (C) en corrigeant de la structure et de l'apparentement les 3 premiers diagrammes correspondent respectivement aux années 2012, 2013 et 2014 (BLUP_G2012, BLUP_G2013, BLUP_G2014) ; Le dernier diagramme correspondant à l'analyse réalisée sur les données ajustées à l'ensemble des 3 années (BLUP_G)

confirment les résultats obtenus sur le caractère acidité avec de nouveau les mêmes SNPs significatifs sur le haut du LG16 et au milieu du LG08 en interannuel. Une autre région localisée à 38,5 Mb sur le LG05 présente des SNPs en association avec le caractère ratio sucre/acidité en 2014. Le test d'association avec le caractère goût a permis l'identification d'une région localisée sur le bas du LG09 à 33,7 Mb en interannuel et d'une autre localisée sur le haut du LG08 à 14 Mb en 2012, mais les diagrammes QQplots incitent à la plus grande prudence (Annexe 15).

Concernant les caractères liés à la texture du fruit, une région a été identifiée pour le caractère croquant dans l'analyse interannuelle, localisée à 10,4 Mb sur le LG16. L'analyse de l'année 2013 a permis l'identification de quatre régions supplémentaires sur le haut du LG04 à 8,8 Mb, le bas du LG10 à 30,2 Mb, le milieu du LG13 à 17,3 Mb et à nouveau le haut du LG16 à 7,7 Mb, soit 3 Mb en amont de la région identifiée dans l'analyse interannuelle. Lors de l'analyse interannuelle du caractère fermeté, les mêmes régions que pour le caractère croquant localisées à 17,4 Mb sur le LG13 et à 10,4 Mb sur le LG16 et trois régions localisées sur les LG04 à 9,3 Mb et LG16 à 0,5 Mb et 9 Mb ont été identifiées. L'analyse de l'année 2013 a permis l'identification d'une région localisée à 20,3 Mb sur le LG10 mais le diagramme QQ plot n'est pas probant (Annexe 15). Les résultats de l'analyse interannuelle du caractère jutosité révèlent des SNPs significativement associés avec le phénotype seulement sur le LG16 à 1,2Mb, c'est-à-dire au même endroit que pour le caractère acidité. Les analyses par année n'apportent pas d'information supplémentaire.

Les résultats des analyses sur les caractères secondaires de texture ne montrent aucun SNP significatif pour le caractère teneur en fibres hormis quelques-uns isolés. L'analyse sur le caractère fondant a permis l'identification de trois régions localisées à 9,4 Mb sur le LG04 et à 9,1 Mb et 10,1 Mb sur le LG16 en interannuel ainsi que quatre régions localisées sur le bas du LG03 à 26,1 Mb et le milieu des LG05 à 20,7 Mb et LG13 à 17,6 Mb en 2013, et sur le bas du LG03 à 28,6 Mb en 2014. L'analyse sur le caractère granulosité ne montre pas de SNPs significatifs en interannuel mais la même région identifiée précédemment sur le bas du LG03 à 28,6 Mb pour le caractère fondant présente des associations significatives en 2014, ainsi qu'une région localisée sur le LG16 à 9,7 Mb en 2013. L'analyse du caractère farinosité n'a permis l'identification d'aucune région génomique contenant des SNPs en association avec le phénotype. Lors de l'analyse sur l'année 2014, une multitude de SNPs répartis sur la majorité des LG présentaient une *p*-value significative mais tous étaient des SNPs isolés, et n'ont donc pas été retenus dans la suite de l'étude.

Enfin, concernant le caractère du russeting, une région génomique située à 32,5 Mb sur le LG12 a été identifiée lors de l'analyse interannuelle, et très clairement confirmée lors de l'analyse de 2014. Deux régions localisées sur le haut des LG08 à 4,1 Mb et LG11 à 7,6 Mb ont également été identifiées lors de l'analyse des données de 2014.

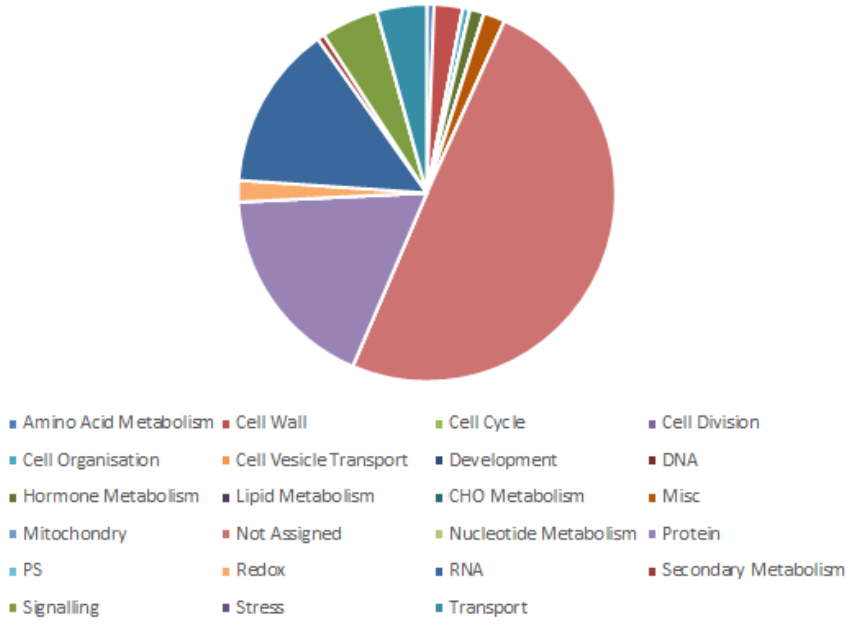


Figure 48 : Distribution des gènes candidats identifiés lors de l'étude de génétique d'association portant sur les caractères de qualité du fruit dans les principales classes de gènes

Tableau 15 : R^2 représentant la part de variation phénotypique expliquée calculé pour les SNPs significativement associés avec les caractères acidité, fermeté et résistance au feu bactérien

Caractère étudié	Marqueur	R^2 individuel	R^2 global
Acidité	LG08 (14,6 Mb)	6,9 %	31 %
	LG14 (30,6 Mb)	4,6 %	
	LG16 (1,2 Mb)	19,5 %	
Fermeté	LG04 (9,3 Mb)	5,1 %	19,6 %
	LG13 (17,3 Mb)	4,5 %	
	LG16 (0,5 Mb)	4,3 %	
	LG16 (9 Mb)	5,5 %	
Résistance au feu bactérien	LG06 (27 Mb)	3,8 %	9,9 %
	LG07 (27,8 Mb)	2,6 %	
	LG11 (38,5 Mb)	3,5 %	

Les résultats des analyses d'association sur les BLUP des interactions génotype-année (BLUP_I2012, BLUP_I2013, BLUP_I2014) ont permis de confirmer les résultats obtenus avec les BLUP annuels mais n'ont pas permis l'identification d'autres régions génomiques.

En complément de ces détections d'associations significatives, les héritabilités au sens strict (h^2 = « chip heritability » ; Speed et al. 2012) ont été calculées pour chacun des caractères (Tableau 13). Les valeurs obtenues varient grandement entre les caractères pour les BLUP_G, de 0 pour la granulosité, le sucre et le goût à 0,86 pour le russeting, mais sont plus constantes pour les BLUP annuels (entre 0,67 et 0,97 en 2012, entre 0,72 et 0,97 en 2013 et entre 0,67 et 0,98 en 2014).

Parmi les 163 gènes détectés autour des SNPs présentant une association significative avec les caractères de qualité du fruit, 80 n'ont pas de fonction prédite ou ne s'alignent avec aucune séquence de la base de données. Les autres gènes sont répartis dans différentes classes fonctionnelles telles que l'organisation cellulaire et les processus impliquant l'ARN (Figure 48). Un gène codant pour un transporteur de malate (MDP0000252114) a été identifié sur le haut du LG16 (1,2 Mb) lors du test d'association réalisé sur le caractère acidité.

2.3. Analyses de variances sur les SNPs significatifs

Des analyses de variance ont été effectuées sur les BLUP_G de quelques caractères (acidité, fermeté et feu bactérien) en retenant comme facteurs les SNPs ayant été précédemment détectés comme significativement associés avec les phénotypes. Les R^2 calculés pour ces trois tests varient de 9,9% à 31% en prenant en compte l'ensemble des marqueurs et de 2,6% à 19,5% pour les marqueurs pris un par un (Tableau 15).

Pour le caractère acidité, sur les quatre marqueurs localisés sur les LG08, LG14 et LG16, trois ont montré un effet significatif lors de l'analyse de variance. Le marqueur situé en position 3,8 Mb sur le LG16 n'a pas montré d'effet significatif. Séparément, les trois marqueurs significatifs représentent respectivement 6,9%, 4,6% et 19,5% de la variation phénotypique et, conjointement, 31,4% de cette variation.

Pour le caractère fermeté, sur les cinq marqueurs détectés comme étant significativement associés au phénotype, quatre ont été trouvés comme ayant un effet significatif lors de l'analyse de variance. Le marqueur non significatif lors de cette analyse simultanée des cinq marqueurs est celui localisé sur le LG16 à 10,1 Mb. L'analyse par marqueur montrait cependant un effet fort de ce marqueur dans la variation phénotypique. Les quatre autres marqueurs localisés sur les LG04, LG13 et LG16 représentent respectivement 5,1%, 4,5%, 4,3% et 5,5% pour un total de variation phénotypique expliquée de 19,6%.

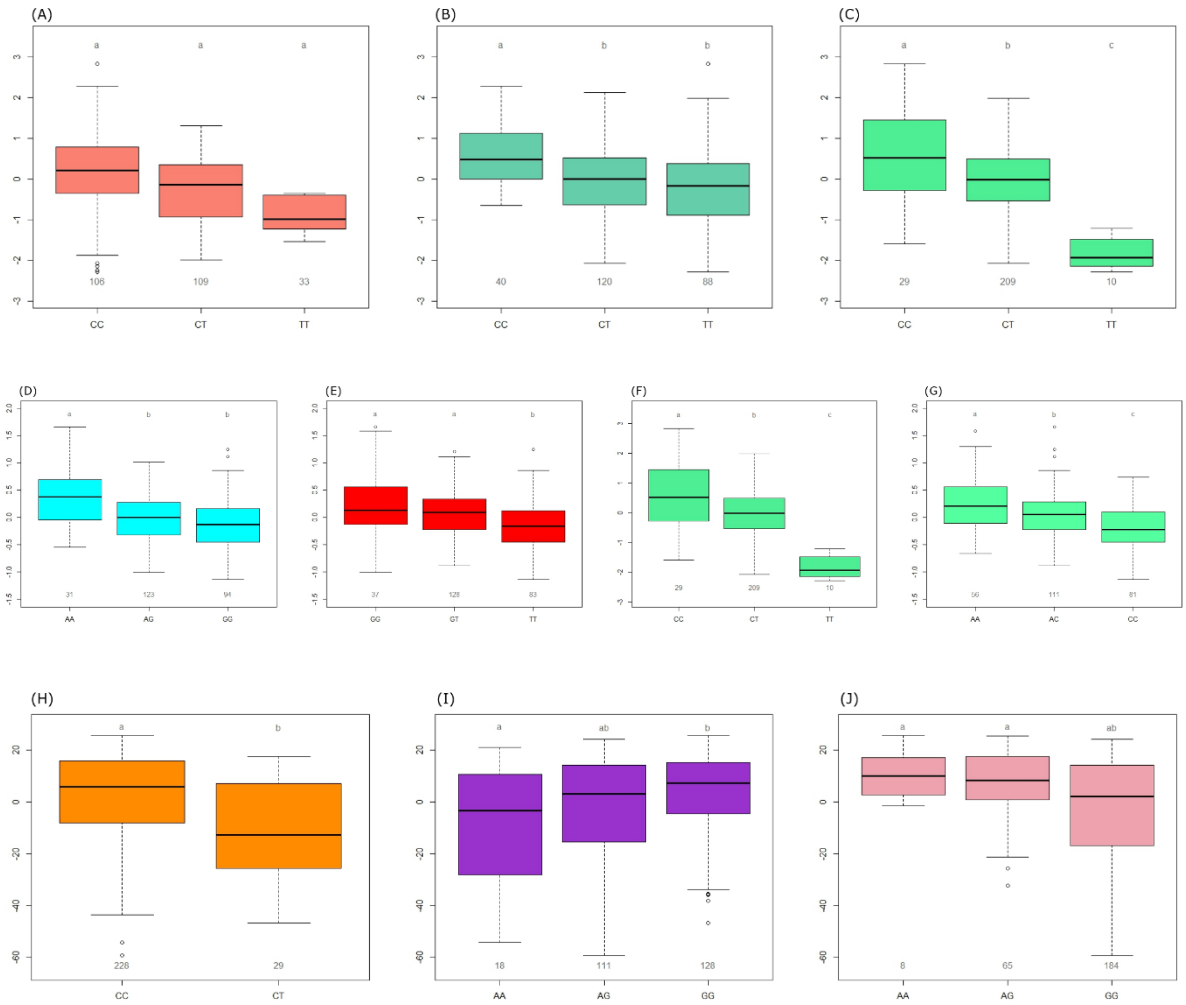


Figure 49 : Représentation graphique du test de comparaison des moyennes des BLUP par classe génotypique (AA, AB et BB) pour les SNPs significativement associés avec les caractères acidité, fermeté et résistance au feu bactérien

(A) LG08 Acidité ; (B) LG14 Acidité ; (C) LG16 Acidité ; (D) LG04 Fermeté ; (E) LG13 Fermeté ; (F) LG16 Fermeté ; (G) LG16 Fermeté ; (H) LG06 Résistance au feu bactérien ; (I) LG07 Résistance au feu bactérien ; (J) LG11 Résistance au feu bactérien

Tableau 16 : Effets d'additivité et de dominance calculés pour les SNPs significativement associés aux caractères acidité, fermeté et résistance au feu bactérien

Caractère étudié	Marqueur	Effet |a|	Effet |d|
Acidité	LG08 (14,6 Mb)	1,17	0,65
	LG14 (30,6 Mb)	0,41	0,17
	LG16 (1,2 Mb)	1,17	0,65
Fermeté	LG04 (9,3 Mb)	0,26	0,17
	LG13 (17,3 Mb)	0,20	0,03
	LG16 (0,5 Mb)	0,23	0,01
	LG16 (9 Mb)	0,22	0,02
Résistance au feu bactérien	LG06 (27 Mb)	-	-
	LG07 (27,8 Mb)	6,61	2,26
	LG11 (38,5 Mb)	6,20	2,60

Pour le caractère de résistance au feu bactérien, les trois marqueurs identifiés ont montré un effet significatif lors de l'analyse de variance. Ces marqueurs localisés sur les LG06, LG07 et LG11 représentent respectivement 3,8%, 2,6% et 3,5%, pour un total de 9,9% de variation phénotypique expliquée.

Les résultats des tests de Newman-Keuls réalisés sur les trois classes génotypiques (AA, AB et BB) observées pour chacun des SNPs identifiés comme significatifs dans les analyses ci-dessus sont présentés en Figure 49. Les effets des différentes classes génotypiques sur les BLUP sont proportionnels aux R^2 calculés précédemment, le SNP localisé sur le haut du LG16 pour le caractère acidité ayant le plus fort effet. Les effets d'additivité et de dominance pour ces marqueurs sont présentés dans le Tableau 16. Pour certains de ces marqueurs, l'effet de dominance est très faible en comparaison de l'effet d'additivité et pour d'autres SNPs, un fort effet de dominance est observé bien que l'effet d'additivité soit toujours supérieur.

3. Discussion

3.1. Analyse des données de phénotypage

3.1.1. Phénotypages

Avant d'explorer les résultats des calculs d'héritabilités et des analyses d'association, certaines précisions doivent être apportées à propos des données de phénotypage et notamment à propos de la façon dont elles ont été acquises.

Tout d'abord, l'échelle de notation des individus pour les différents tests (sauf le test de résistance au feu bactérien pour lequel la note correspond à une mesure physique en centimètres) est souvent imprécise. Lors des tests de résistance à la tavelure, une note de 0 à 7 représentant le pourcentage de surface foliaire recouvert de sporulation a été donnée (adaptée de Croxall et al. 1952); pour les notations des caractères de qualité de fruit, une note de 1 à 9 a été attribuée. Ces notations sont réalisées selon une échelle ordinale qui est, par définition, discontinue et qui est utilisée dans le cadre d'une évaluation visuelle ou sensorielle, donc avec une part de subjectivité (Yoshioka and Fukino 2010). Les notations ne peuvent donc pas être aussi précises que s'il s'agissait de mesures instrumentales.

Viennent ensuite les conditions dans lesquelles se trouve le notateur et qui peuvent grandement impacter les notes qu'il va attribuer aux individus. Parmi ces conditions se trouvent par exemple l'état

de fatigue du notateur, l'influence de l'individu n – 1 sur la notation de l'individu n (une variété acide ne sera pas perçue exactement de la même manière si elle est notée à la suite d'une autre variété acide ou à la suite d'une variété douce), ou encore le moment de la journée pendant laquelle a lieu la notation (après ou avant le repas pour les tests de qualité du fruit). Il convient néanmoins de préciser que les notateurs (« panel de dégustation ») ont acquis au cours des ans une bonne expérience dans l'évaluation sensorielle de la qualité des fruits.

Les conditions environnementales de croissance des arbres ont également une forte influence sur les notes finales données aux individus. Lors des tests de résistance, cela correspond à l'hétérogénéité de la température, de la luminosité ou encore de la disponibilité en eau, malgré des conditions globales contrôlées, pour les copies de génotypes inoculées à des dates différentes, mais également à l'intérieur de la serre en fonction de la position des individus sur les tablettes.

Pour les données de qualité du fruit, cette hétérogénéité d'environnement correspond par exemple à l'emplacement des arbres dans les parcelles qui peut différer pour les conditions d'ensoleillement ou d'hygrométrie du sol (Ebel, Proebsting, and Patterson 1993), certains génotypes ayant même été récoltés dans des parcelles différentes, même si elles sont voisines. Ce facteur « parcelle » n'a cependant pas pu être pris en compte dans l'analyse statistique (pas de répétition des variétés entre les parcelles). L'âge de l'arbre sur lequel le fruit dégusté a été récolté peut également avoir eu une influence importante. Un autre point important est la date de récolte des fruits, théoriquement à maturité physiologique qui correspond à la maturité optimale pour une dégustation « à la récolte », c'est-à-dire sans période de conservation préalable. Ce stade, déterminé grâce au dosage de réduction de l'amidon (Smith et al. 1979), peut en effet être difficile à identifier et il est probable que certains génotypes aient été récoltés quelques jours avant ou après la date correspondant à ce stade. Un autre biais dû à la période de dégustation découle directement du point précédent. Les variétés anciennes de pommes à couteau arrivent en effet à maturité à des dates très étalées dans le temps et il a été montré que cette date a une influence sur les qualités organoleptiques des fruits. Les fruits des variétés précoces et tardives ne subissent en effet pas les mêmes conditions environnementales avant leur récolte ce qui peut jouer sur certains des caractères notés ici (Blankenship 1987; Nava, Dechen, and Nachtigall 2007).

Une autre difficulté lors de l'échantillonnage des fruits est de réussir à choisir des pommes représentatives de l'arbre entier. Hors, il a été montré qu'en fonction de la position du fruit sur l'arbre, les qualités organoleptiques des fruits pouvaient être grandement impactées, et ce en raison de la différence d'ensoleillement aux différents niveaux de l'arbre et de la différence d'alimentation des fruits en carbohydrates et en eau (Tustin, Hirst, and Warrington 1988). Ce choix des fruits sur l'arbre a de plus pu être impacté par la disponibilité en fruits, dépendante de l'année de récolte en raison du phénomène d'alternance observé chez le pommier, qui correspond à une charge élevée de fruits une année et très faible l'année suivante (Jonkers 1979). La représentativité des échantillons dégustés peut

en effet être discutée puisque, en moyenne, les notateurs n'ont dégusté qu'un seul fruit (deux morceaux différents pour les deux notateurs), sauf dans le cas où les deux notateurs avaient des avis très différents sur les notes attribuées, auquel cas ils dégustaient un deuxième fruit. Enfin, les analyses de variance ont montré un effet significatif de l'effet année et ce pour tous les caractères. Un effet d'interaction a également été montré par le déclassement des variétés communes aux trois années de notation d'une année sur l'autre, bien qu'il soit vraisemblablement dû à un très fort effet de l'échantillonnage. Le fruit récolté et dégusté une année peut se révéler avoir de bonnes qualités organoleptiques et celui de l'année suivante des qualités seulement suffisantes, et cela ne sera dû qu'au choix du fruit sur l'arbre. L'année de récolte a donc une influence sur les notes attribuées aux fruits à travers le caractère relativement aléatoire de l'échantillonnage réalisé.

3.1.2. Héritabilités

Les héritabilités au sens large (H^2) calculées pour les différents caractères de résistance et de qualité du fruit sur les trois années sont globalement élevées. Dans le cas du test de résistance au feu bactérien, l'estimation de l'héritabilité a cependant pu être surestimée par la présence d'un nombre significatif de génotypes avec une même note maximale (100% de la tige nécrosée) pour toutes les copies testées, ce qui aura entrainé une sous-estimation de la variation intra-classe. Quelques caractères de qualité du fruit tels que la farinosité ou la granulosité présentent des héritabilités au sens large de faible valeur qui peuvent s'expliquer par la distribution non normale des notes attribuées aux individus pour ces deux caractères (valeurs centrées sur les faibles classes), et tout particulièrement pour la farinosité. Lors des analyses de variance réalisées sur les données de qualité du fruit par année de notation, les valeurs des héritabilités annuelles obtenues sont toutes supérieures à 0,65. Ces valeurs sont très fortes et probablement surestimées du fait du mode de dégustation des fruits, c'est-à-dire un même fruit dégusté par deux notateurs, ce qui génère un effet d' « environnement commun » très important qui entraîne une sous-estimation de la variation résiduelle et, en conséquence, une surestimation de l'héritabilité. Dans l'idéal, il aurait fallu qu'un fruit différent récolté sur un arbre différent de la même variété soit dégusté par chaque notateur, ou au moins que deux fruits différents du même arbre soient dégustés par les notateurs. L'ampleur du décalage entre l'héritabilité interannuelle et les héritabilités annuelles peut également être interprétée comme une indication de l'impact de l'échantillonnage, variable selon les caractères : par exemple, pour les caractères acidité, ratio sucre/acide et russeting, les héritabilités interannuelles et annuelles sont du même ordre de grandeur (0.8-0.9), ce qui traduit potentiellement un faible impact de l'échantillonnage (et une relativement faible interaction génotype-année) ; au contraire, pour des caractères comme la teneur en fibres, la granulosité et le fondant, le fort décalage entre les héritabilités interannuelles (faibles)

d'une part, et les héritabilités annuelles (fortes) d'autre part, dérive très vraisemblablement d'un échantillonnage plus aléatoire entre les années combiné à l'effet d'« environnement commun » issu de la dégustation du même fruit par les notateurs une année donnée.

Les héritabilités au sens strict (h^2 = « chip heritability ») calculées à la suite des analyses d'association sont, de la même manière, relativement élevées pour la plupart des caractères de résistance et qualité du fruit en interannuel. Les héritabilités au sens strict sont du même ordre ou inférieurs aux héritabilités au sens large, ce qui est attendu puisque l'héritabilité au sens strict ne reflète que la part de variance additive alors que l'héritabilité au sens large intègre les effets additif et de dominance. Un seul des caractères, la farinosité, présente une très faible héritabilité au sens large (0,08) et une héritabilité au sens strict assez élevée (0,51). Compte-tenu de la distribution des valeurs phénotypiques du caractère farinosité ainsi que de celle des résidus, les valeurs d'héritabilité et les résultats des analyses d'association ont été considérés avec la plus grande prudence. D'autres caractères, granulosité, sucre et goût, présentent quant à eux des héritabilités nulles en interannuel alors que les héritabilités au sens large varient de 0,21 à 0,54. L'attendu pour ces caractères est donc de n'identifier aucun SNP significativement associé avec le phénotype lors des analyses d'association sur les BLUP_G, ce qui s'est généralement produit.

Pour les héritabilités au sens strict annuelles, plusieurs situations sont observées. Certains caractères comme le goût et la granulosité présentent des héritabilités au sens strict faibles voire nulles pour les trois années de notation. D'autres, comme l'acidité et le russeting, présentent des héritabilités au sens strict élevées pour les trois années. Certains enfin, comme le croquant, la jutosité et le fondant présentent des héritabilités au sens strict élevées une année et faible ou nulle une autre année. Ces différentes situations reflètent l'interaction des effets année et génotype, très forte dans les cas où les héritabilités varient grandement d'une année à l'autre. Les différents points évoqués en 3.1.1 peuvent expliquer ces différences entre les héritabilités annuelles pour un même caractère. A nouveau, les biais potentiels liés à l'échantillonnage annuel et à la dégustation d'un même fruit par les notateurs peuvent donner une certaine « illusion » de la variabilité génétique entre les variétés alors que celle-ci, même si elle existe vraisemblablement *a minima*, est très difficile à « capter » à travers l'analyse sensorielle.

Des analyses de variance sur les marqueurs détectés comme étant associés significativement avec les caractères acidité, fermeté et résistance au feu bactérien ont permis l'estimation du R^2 représentant la part de variation phénotypique expliquée conjointement par ces marqueurs. Les valeurs obtenues lors des analyses de variance sont très inférieures aux valeurs d'héritabilités au sens strict calculées suite aux analyses d'association. Ce phénomène aussi connu sous le terme de « missing heritability » correspond à la part de variation phénotypique expliquée manquante et est commun à de nombreuses études d'association (Maher 2008; Yang et al. 2010; Crow 2011; Ridge et al. 2013).

Plusieurs hypothèses ont été avancées quant aux causes de la « missing heritability ». La première hypothèse repose sur un déterminisme très polygénique des caractères étudiés où chaque

locus impliqué ne contribue que pour une faible part au contrôle du caractère. Les études d'association ne permettent en effet pas l'identification d'une multitude de loci expliquant chacun un trop faible pourcentage de la variation phénotypique totale (Gibson 2012). La seconde hypothèse repose sur le fait que les allèles causaux associés à un phénotype peuvent être présents en trop faible fréquence dans les populations étudiées. Cela s'illustre dans cette étude par exemple dans le cas des tests de résistance à la tavelure. Lors des tests phénotypiques, certains génotypes présentaient des symptômes de résistances assez forts (symptôme de « pin-point », forte crispation de la feuille, tâches de nécrose et de chlorose très marquées…). Le nombre de génotypes présentant ces symptômes et possédant les allèles responsables n'était cependant peut être pas suffisant pour permettre la détection d'un signal significatif pendant l'analyse d'association, surtout s'il s'agissait de gènes différents (situés dans différentes régions génomiques) qui contrôlaient l'apparition d'un même symptôme comme cela peut être le cas pour ce type de symptôme. Les allèles rares ne peuvent en effet être identifiés dans une étude d'association que si la taille de la population est suffisamment grande pour que la puissance statistique du test permette leur détection (Manolio et al. 2009). De plus, les variétés sur lesquelles ont été réalisés les tests de résistance à la tavelure et au feu bactérien font partie d'une core collection qui regroupe une diversité génétique relativement large. Cela revient à regrouper un maximum d'allèles différents qui par conséquent se retrouvent en fréquence faible. Cela soulève également le problème de la taille de la population étudiée, relativement limitée dans notre étude : en augmentant l'effectif total étudié, le nombre d'individus porteurs des allèles causaux rares serait également augmenté. Chez l'homme, cette méthode a permis l'identification de SNPs causaux de faible fréquence dans des populations d'étude de très grande taille (Rivas et al. 2011; Trynka et al. 2011). La troisième hypothèse permettant d'expliquer la « missing heritability » repose sur le fait que les modèles employés ici pour effectuer les analyses d'association sont des modèles « simple marqueur » qui ne prennent pas en compte certaines formes d'épistasie pouvant exister entre plusieurs marqueurs (Zuk et al. 2012). Ainsi, deux marqueurs détectés comme étant non significativement associés avec le phénotype étudié pourraient en réalité participer à la « missing heritability » s'ils étaient considérés non pas séparément, mais en interaction. La quatrième hypothèse concerne la représentativité des polymorphismes étudiés en comparaison avec les polymorphismes totaux dans un génome. Les études d'association sont en effet souvent réalisées sur des données de génotypage issues de puces SNP. Ces données peuvent paraître comme étant représentatives des polymorphismes présents dans les génomes. Cependant, les marqueurs ne couvrent pas toujours la totalité du génome, et la puce 480k SNPs en est un bon exemple. Certaines zones du génome plus ou moins étendues (jusqu'à 3,92 Mb) ne sont en effet pas jalonnées par des marqueurs. Les données utilisées ici sont également exemptes des marqueurs localisés sur le LG0, correspondant à tous les contigs qui n'ont pas pu être mappés sur la version 3 du génome du pommier. Il est donc possible que certains SNPs causaux non détectés ici fassent partie du LG0. Par ailleurs, les marqueurs SNP ne prennent pas en compte la variation possible du nombre de copies de

gènes (en tandem ; « Copy Number Variation »), qui peut avoir un rôle sur l'expression de certains caractères comme c'est le cas pour la résistance du soja à un nématode (Cook et al. 2012). Enfin, une dernière hypothèse concernant la « missing heritability » repose sur le fait que les marqueurs génotypiques utilisés dans bon nombre d'études ne sont pas suffisants pour capter toute la diversité au sein des génomes, et notamment la diversité épigénétique (Slatkin 2009). Un gène monomorphe dans une population peut, par exemple, être méthylé chez certains individus et non chez d'autres, ce qui peut conduire à son inactivation différentielle et à une variation phénotypique sans variation génétique au niveau de la séquence ADN.

Tous les problèmes mentionnés ci-dessus concernant le phénotypage des individus et la « missing heritability » participent au faible nombre de régions génomiques identifiées comme étant significativement associées avec les caractères étudiés ici.

3.2. Analyses d'association

3.2.1. Données de résistance aux maladies

Les tests d'association sur les données de résistance à la tavelure et au feu bactérien ont donné des résultats peu significatifs. Les marqueurs les plus significativement associés avec le phénotype ont en effet une *p*-value à peine plus faible que le seuil fixé ici, et très peu de SNPs, un ou deux par analyse, sont significatifs. En fixant un seuil légèrement inférieur à celui utilisé précédemment, -log(*p*) = 4,5 au lieu de –log(*p*) = 5, d'autres régions génomiques susceptibles de contenir des SNPs causaux ont été identifiées, même si elles sont à considérer avec précaution.

Parmi les cinq régions génomiques identifiées dans les analyses de résistance à la tavelure, deux co-localisent avec des QTLs de résistance à la tavelure détectés dans des études de cartographie précédentes. La région localisée à 22,1 Mb sur le LG02 co-localise avec une région génomique portant différents QTLs identifiés par Calenge et al. (2004). Ces QTLs ont été identifiés lors de l'analyse d'une descendance F1 issue d'un croisement entre les variétés « Discovery » et « TN10-8 » et inoculée successivement avec plusieurs souches monoconidiales de *Venturia inaequalis* (dont la souche 104 utilisée ici). Une autre étude, impliquant un mélange de souches de *V. inaequalis* (dont les souches 104 et EU-B04) inoculé sur la même descendance F1, a également permis d'identifier plusieurs QTLs localisés dans la même région (Lê Van et al. 2013). Il est également intéressant de noter que cette région est connue pour abriter trois gènes majeurs de résistance, *Vbj*, *Vh2* et *Vh8* tous trois identifiés dans des pommiers sauvages (Gygax et al. 2004; Bus et al. 2005). Cette région génomique du LG02 est connue depuis longtemps pour abriter différents facteurs de résistance à la tavelure. Il est donc

intéressant de pouvoir la retrouver par génétique d'association. La plus-value de la démarche GWAS devrait se situer dans une diminution de l'intervalle de confiance contenant le/les facteur(s) de résistance révélé(s) ici, en appliquant par exemple un intervalle de +/- 10kb autour du SNP le plus significatif. Il convient cependant d'être assez prudent dans le cas présent, car il est probable que cette région soit relativement étendue, surtout si elle contient un ensemble de gènes de résistance dont certains pourraient correspondre à des gènes majeurs de type NBS-LRR (Nucleotide Binding Site – Leucine Rich Repeat) chez les espèces sauvages (hypothèse fonctionnelle possible pour les gènes *Vbj*, *Vh2* et *Vh8*). Ce type de gènes est en effet connu pour être assez largement réparti dans le génome avec une distribution en clusters, comme illustré récemment par Perazzolli et al. (2014) qui ont analysé la répartition de l'ensemble des analogues de gènes de résistance de type NBS prédits à l'échelle du génome du pommier. Il est d'ailleurs intéressant de noter qu'un cluster de gènes de type NBS est situé précisément dans la région du LG02 où se situe le SNP le plus significatif ici. Dans ce cadre, l'interprétation fonctionnelle possible d'une telle co-localisation serait que le locus identifié ici par génétique d'association correspondrait à une version « quantitative » des gènes majeurs, à effet qualitatif chez les espèces sauvages. Une telle hypothèse avait déjà été avancée dans le cas des co-localisations entre QTLs et gènes majeurs détectées par Calenge et al. (2004) et Lê Van et al. (2013), mais nécessiterait d'être validée. La région localisée à 30,5 Mb sur le LG15 co-localise également avec plusieurs QTLs de résistance à la tavelure. Calenge et al. (2004), Durel et al. (2003) et Durel et al. (2004) ont ainsi identifié dans cette région plusieurs QTLs de résistance, face à la souche EU-NL24, à deux souches de race (6) et à une souche de race (7) respectivement. Ces QTLs couvraient à chaque fois des intervalles de confiance très grands. Il est donc difficile de savoir si la co-localisation est significative ou fortuite. Par ailleurs, la recherche de co-localisations avec un cluster de gènes de type NBS n'a pas permis de révéler de coïncidence notoire sur ce LG15. Aucune co-localisation avec des QTLs de résistance à la tavelure n'a été observée pour les régions localisées sur les LG01, LG08 et LG13. Seul le haut du LG08 présente une concentration en gènes de type NBS particulièrement élevée (Perazzolli et al. 2014), ce qui pourrait amener aux mêmes interprétations (et précautions) que précédemment. Étonnamment, le LG17 connu pour porter un grand nombre de QTLs de résistance à la tavelure n'a pas été identifié ici. Les études ayant permis la localisation de ces QTLs portaient sur des descendances issues de croisement avec des individus d'origine anglaise (Calenge et al. 2004; Lê Van et al. 2013). Il est donc possible que les QTLs identifiés précédemment proviennent du pool génétique des variétés anglaises, non représentées dans la core collection étudiée ici, d'où cette absence de signal.

Parmi les trois régions génomiques identifiées dans les analyses de résistance au feu bactérien, une seule co-localise avec un QTL de résistance au feu bactérien. Cependant, cette région, localisée sur le LG07, co-localise avec un QTL identifié dans plusieurs études sur des individus descendant de croisements avec la variété « Fiesta » comme parent (Calenge et al. 2005; Khan et al. 2006). Dans

une étude complémentaire, Khan et al. (2007) ont aussi montré que l'allèle de résistance de ce QTL dérivait très vraisemblablement de la variété anglaise « Cox's Orange Pippin » et était hérité dans plusieurs variétés récentes sélectionnées à partir du parent « Cox ». Cette variété n'est pas présente dans la core collection étudiée ici, mais il est possible que des variétés apparentées et portant le même allèle favorable y soient présentes et aient ainsi contribué à l'identification d'un signal significatif à cet endroit du génome. Dans ce cas, le gain de précision dans l'emplacement de ce QTL grâce à l'approche de génétique d'association serait très conséquent si on retient une fenêtre de +/- 10kb autour du SNP le plus significatif. Cependant, les gènes candidats positionnels de cette région n'ont pas permis d'en retenir un (ou certains) sur la base d'une fonction qui aurait du sens au regard de ce qui est connu en matière de résistance du pommier au feu bactérien. Les deux autres régions localisées sur les LG06 et LG11 ne co-localisent avec aucun des rares QTLs identifiés pour la résistance au feu bactérien.

Les QTLs de résistance identifiés jusque-là l'ont été dans des descendances F1 ne représentant qu'une infime partie de la diversité génétique du pommier. Les régions identifiées ici qui ne co-localisent avec aucun QTL connu dans la littérature représentent ainsi peut-être de nouvelles sources de résistance à étudier afin de mieux connaître l'architecture génétique de la résistance du pommier à ces deux maladies. Si elles étaient confirmées dans le cadre de nouvelles expérimentations en serre, et si possible sur le terrain, ces régions génomiques pourraient être exploitables dans une optique de sélection assistée par marqueurs, en ayant au préalable identifié les phases favorables (résistantes) des SNPs retenus.

Les gènes localisés sur le génome autour des SNPs localisés au point le plus haut d'un pic composé d'autres marqueurs ayant des *p*-values plus faibles ont été étudiés. Parmi les gènes identifiés dans les différentes régions génomiques pour les différents tests, la moitié n'ont pas de fonction prédite. Cependant, un gène identifié par l'analyse des données de résistance au feu bactérien et localisé sur le LG06, présente une forte homologie de séquence avec le gène *NPR1* de l'espèce *Pyrus bretschneideri* (Fan et al. 2010). Ce gène, très étudié chez *Arabidopsis thaliana* (Cao et al. 1997; Després et al. 2000; Fan and Dong 2002), code pour une protéine impliquée dans l'induction de la résistance systémique acquise (SAR). Lors de l'infection d'une plante par un pathogène, si la plante reconnait le pathogène, la production d'acide salicylique est activée, puis la protéine NPR1 migre dans le noyau et interagit avec des facteurs de transcription TGA, qui vont induire l'expression des gènes de défense (Durrant and Dong 2004). Hors, il est connu que les bactéries nécrotrophes, dont fait partie *Erwinia amylovora*, sont principalement bloquées par l'activation de la voie de l'acide jasmonique lors de l'infection d'une plante (Glazebrook 2005). Chez le pommier, les études menées au sein de l'équipe ResPom (Dugé De Bernonville et al. 2012) ont montré que la bactérie réprime la voie de l'acide jasmonique dans les feuilles infectées d'une variété sensible (« MM106 ») alors qu'elle ne la réprime pas chez une variété résistante (« Evereste »). La voie de l'acide salicylique n'est quant à elle pas différentiellement modulée entre ces deux variétés. Il a cependant été montré chez la tomate, que certaines bactéries nécrotrophes

activeraient, *via NPR1* et d'autres gènes de la même voie, la voie de l'acide salicylique qui empêcherait la voie de l'acide jasmonique de se mettre en place du fait d'un antagonisme entre ces deux voies et permettrait ainsi la réussite de l'infection (Abd-El Rahman et al. 2012). Le gène identifié ici pourrait donc participer à la modulation de la résistance quantitative du pommier face à *Erwinia amylovora*. selon un mécanisme qui pourrait être différent de celui présent chez la variété résistante « Evereste ». Il a d'ailleurs été montré que la surexpression de ce gène dans des variétés de pommiers sensibles au feu bactérien transformées (« Galaxy » et « M26 ») permettait de réduire la sensibilité à la bactérie *E. amylovora* (Malnoy et al. 2007). De plus, une application d'acibenzoar-S-methyl, un analogue de l'acide salicylique, réduit la sensibilité d'individus de la variété « Golden Delicious » (Brisset et al. 2000). Des études d'expression de ce gène pourraient donc être envisagées sur une petite gamme de variétés de pommiers de la core collection en condition d'infection afin de confirmer/infirmer, et le cas échéant mieux caractériser, le rôle de *NPR1* chez le pommier vis-à-vis d'*E. amylovora*.

3.2.2. Données de qualité du fruit

Selon les caractères étudiés, les analyses d'association ont donné des résultats plus ou moins intéressants. L'étude des caractères teneur en fibre et farinosité n'ont permis l'identification d'aucun marqueur en association significative avec le phénotype. Pour les dix autres caractères de qualité du fruit, au moins un SNP a été détecté, que ce soit lors de l'analyse des données interannuelles (BLUP_G) ou lors de l'analyse des données annuelles. Au total, 41 SNPs répartis dans 27 régions génomiques ont été identifiés lors de cette étude. Sur ces 27 régions, six co-localisent avec des QTLs de qualité du fruit. Les autres régions ne co-localisant avec aucun QTL de qualité du fruit représentent une nouvelle source potentielle de variation et mériteront d'être confirmées dans d'autres études GWAS.

La première région, localisée sur le LG08 à 14,6 Mb et contenant un même SNP lié aux caractères acidité et ratio sucre/acidité, co-localise avec deux QTLs d'acidité représentant une part importante de la variation phénotypique observée dans deux descendances F1 cartographiées (Liebhard et al. 2003; Kenis, Keulemans, and Davey 2008). Pour l'acidité, le SNP du LG08 représentait 6,8% de la variation phénotypique totale observée au sein de la core collection, alors que les QTL captent 7,9% et 33% respectivement de la variation observée. La grande différence avec l'étude de Kenis et al. (2008) est logique puisque le nombre de loci contrôlant la variation du caractère acidité est probablement beaucoup plus important au sein de la core collection étudiée ici qu'au sein de la descendance F1 cartographiée. Il est intéressant de noter la présence d'un SNP associé au caractère goût dans une région située 600 kb en amont sur le LG08 (14 Mb). Les caractères goût et ratio sucre/acidité présentent une corrélation relativement élevée, tout comme les caractères goût et sucre, contrairement aux caractères goût et acidité. Il est donc plus vraisemblable que la corrélation avec le

ratio sucre/acidité soit due au caractère sucre. La présence du SNP associé au goût à 600 kb des deux SNPs associés à l'acidité et au ratio sucre/acidité relève donc plus probablement de la présence d'un second gène responsable du goût des fruits plutôt que d'une co-localisation due à la corrélation entre les caractères.

La seconde région, localisée à 33,7 Mb sur le LG09 et contenant un SNP associé au caractère goût, co-localise avec deux QTLs responsables de la teneur en divers composés détectés dans le cadre d'études de spectrométrie ou chromatographie, réalisées sur des échantillons de fruits à maturité (Zini et al. 2005; Dunemann et al. 2009). Ces composés étaient soit déjà identifiés tels que l'allylanisol, soit non identifiés tels que le composé X51 et le fragment de masse *m/z 61*, et correspondent à des molécules aromatiques susceptibles de jouer un rôle dans la perception du goût des fruits.

La troisième région, localisée à 20,3 Mb sur le groupe de liaison 10 et contenant un SNP lié à la fermeté du fruit, co-localise avec deux QTLs de fermeté identifiés grâce à des mesures instrumentales (mesures physiques) réalisées sur la chair des fruits par pénétrométrie (Kenis, Keulemans, and Davey 2008; Kumar et al. 2012). Une autre étude a par la suite localisé dans cette région un gène candidat codant pour une polygalacturonase (*Md-PG1*) qui serait responsable des QTLs identifiés (Costa et al. 2010). La localisation précise de ce gène a ensuite été proposée à 18,1 Mb à la suite d'une étude d'association sur 77 variétés de pommes à couteau (Longhi et al. 2013). La région identifiée ici se trouve donc environ 2 Mb en aval du gène *Md-PG1* identifié précédemment, soit en dehors de la fenêtre moyenne de distance à l'intérieure de laquelle les marqueurs sont en DL. Trois hypothèses sont alors envisageables :

(i) la première concerne l'agencement du génome du pommier : la localisation de *Md-PG1* à 18,1 Mb sur le LG10 a en effet été réalisée sur la version 1 du génome du pommier tandis que les SNPs de la puce 480k SNPs ont été placés sur la version 3 du génome ; le contig sur lequel est localisé *Md-PG1* (MDC004966.433) ne contient pas de SNP, donc aucune association entre des marqueurs en DL avec le gène n'a pu être testée ; le contig MDC004966.433 n'a de plus pas de position fixe entre les deux versions du génome, étant localisé sur le LG10 dans la version 1 mais non localisé dans la version 3 ; nous ne pouvons donc pas écarter la possibilité que ce gène soit sous-jacent à la région identifiée ici ;

(ii) la deuxième explication concerne la méthode de mesure de la fermeté : les données utilisées lors de l'étude qui a permis d'identifier le gène codant pour la polygalacturonase sont des mesures physiques de la fermeté réalisées avec un appareil dédié ; les données utilisées lors de la présente étude sont des données sensorielles évaluées par un panel de dégustateurs ; il est donc possible que les deux caractères ne reposent pas sur les mêmes bases génétiques, ce qui serait

illustré ici par l'identification de *Md-PG1* en relation avec les mesures physiques, et d'une autre région génomique en relation avec les données sensorielles ;

(iii) la troisième explication pourrait résider dans le fait que la population étudiée lors de l'identification de *Md-PG1* est constituée, pour plus de la moitié des individus, de variétés récentes (Longhi et al. 2013) ; au contraire, les variétés rassemblées ici dans la core collection sont toutes des variétés anciennes, c'est-à-dire obtenues avant 1950 et donc avant que la plupart des variétés récentes ne descendent que de quelques variétés fondatrices ; il est donc possible que la plupart des variétés récentes possèdent un allèle favorable au locus *Md-PG1*, hérité d'une ou quelques-unes des variétés fondatrices ; cet allèle serait alors présent en fréquence élevée dans la population étudiée par Longhi et al. (2013) ; les variétés de notre core collection pourraient inversement ne le posséder qu'en faible fréquence ce qui expliquerait alors l'absence de signal détecté autour du locus *Md-PG1* dans notre collection.

Par ailleurs, il est également intéressant de noter la présence d'un autre SNP associé significativement avec le caractère croquant localisé à 30,2 Mb sur le LG10. Un gène ayant un effet sur la production d'éthylène, et donc associé à la maturation des fruits et à leur perte de fermeté et de croquant (*Md-ACO1*), a été localisé dans la même région (Costa et al. 2010). Ce résultat est cependant à considérer avec prudence, notamment à cause des erreurs d'assemblage observées dans le génome du pommier. Le contig sur lequel le SNP a été identifié était en effet localisé sur le LG04 de la version 1 du génome mais a été relocalisé sur le LG10 dans la version 3 du génome. Il serait donc important de vérifier la position de ce SNP par une approche de cartographie de marqueurs, effectuée sur une descendance par exemple, ou en visualisant plus finement la structure du DL dans cette région comme proposé dans le Chapitre 4.

Les quatrième et cinquième régions, localisées à 28,2 Mb sur le LG13 et 10,4 Mb sur le LG16 et contenant chacune un SNP associé avec les caractères acidité et fermeté respectivement, co-localisent chacune avec un QTL identifié vis-à-vis de ces mêmes caractères (Kenis, Keulemans, and Davey 2008).

La sixième région, localisée tout en haut du LG16 (entre 1,2 et 3,8 Mb) a fait l'objet d'une attention plus particulière. Cette région contient deux SNPs liés aux caractères acidité, ratio sucre/acidité et jutosité et co-localise avec deux QTLs d'acidité ayant un effet fort (Liebhard et al. 2003; Kenis, Keulemans, and Davey 2008). Il est intéressant de noter que le SNP localisé à 1,2 Mb a été identifié lors des analyses des trois caractères acidité, ratio sucre/acidité et jutosité, et que le second localisé à 3,8 Mb a été identifié seulement lors des analyses des deux caractères acidité et ratio sucre/acidité. Un premier point de discussion concerne donc l'identification d'un même SNP lors de l'analyse de deux caractères (acidité et jutosité) ne présentant apparemment pas de lien entre eux. Lors de la dégustation des fruits, le caractère acidité ne peut être détecté que si le jus du fruit est

relâché en bouche lorsque le notateur le croque. Le caractère acidité est donc fortement dépendant du caractère jutosité, un fruit ne pouvant avoir une note élevée pour l'acidité s'il a une note faible pour la jutosité. Un élément permettant de valider cette supposition est la relativement forte corrélation entre les deux caractères (0,28). Dans un second temps, nous nous sommes intéressés au fait que deux régions génomiques proches aient été identifiées lors de l'analyse d'association. Les deux régions étant localisées à 2,6 Mb, deux scenarii étaient envisageables. Le premier aurait vu deux endroits relativement proches exhiber une association avec le même caractère, ici l'acidité. Le second aurait remis en cause l'agencement des contigs sur la version 3 du génome, les deux pics n'étant en fait qu'un seul. L'analyse du DL dans cette région a permis de pencher pour le second scénario (voir Chapitre 4, 3.2). Les deux contigs étant en fort DL entre eux mais sans l'être avec les marqueurs supposément localisés entre les deux pics, cela a conduit à la conclusion que les deux SNPs sont en réalité localisés dans la même petite région génomique. Un autre élément permettant d'étayer ce scénario est le fait que, lors de l'analyse de variance réalisée en prenant comme facteurs les SNPs significativement associés avec le caractère acidité, l'effet du marqueur localisé à 3,8 Mb était annulé par l'effet du marqueur localisé à 1,2 Mb. Il est également intéressant de noter que dans la version 1 du génome, les deux contigs sur lesquels sont localisés ces deux SNPs étaient distants, non pas de 2,6 Mb comme sur la version 3, mais de 45 kb. L'analyse d'association a donc permis l'identification d'une seule région génomique associée avec l'acidité des fruits et la jutosité sur le haut du LG16 à 1,2 Mb.

La recherche de gènes autour du SNP localisé à 1,2 Mb du LG16 a permis l'identification d'un gène candidat codant pour un transporteur de malate (MDP0000252114). L'acide malique représente environ 90% du contenu en acide dans les pommes, le reste étant composé d'acides citrique, succinique et autres à l'état de traces (Ackermann, Fischer, and Amado 1992). Ce gène est donc un excellent candidat et pourrait être responsable de la teneur en acide malique dans les fruits. Ce gène a déjà été pointé dans une étude réalisée par Bai et al. (2012) qui portait sur la recherche de gènes candidats pour le locus *Ma* (pour « malic acid »), cartographié depuis de nombreuses années sur le haut du LG16 (Maliepaard et al. 1998). Ces auteurs ont montré qu'un SNP situé dans ce gène entrainait l'apparition d'un codon stop qui était fortement corrélé à l'absence d'acidité dans les variétés où ce gène était muté. Il se trouve que le SNP (G/A) identifié comme étant le plus significativement associé avec le caractère acidité dans notre étude (et donc présent sur la puce de génotypage) est précisément ce même SNP qui provoque l'apparition du codon stop en position 1455 lorsque l'allèle est la base A (Bai et al. 2012). L'effet de ce SNP est confirmé avec les résultats du test de comparaison de moyennes des différentes classes génotypiques qui montrent un écart de deux points (sur l'échelle 1-9) dans les BLUP entre les individus possédant deux copies de l'allèle G et ceux ayant deux copies de l'allèle A. Notre étude d'association a donc confirmé sur un effectif beaucoup plus conséquent que celui de l'étude de Bai et al. (2012), effectuée sur 29 variétés, l'implication très vraisemblable de ce transporteur de malate dans la variation du niveau d'acidité des pommes au sein d'une collection de variétés anciennes à couteau.

Nous avons par ailleurs pu montrer que ce caractère présentait une légère dominance de l'allèle acide sur l'allèle doux (effet d'additivité et de dominance estimés respectivement à 1,16 et 0,64).

Le haut du LG16 a été identifié ici comme abritant un nombre important de SNPs associés avec un certain nombre de caractères de qualité du fruit (acidité, fermeté, croquant, granulosité et fondant). D'autres études ont également permis l'identification de gènes et de QTLs de qualité du fruit tels que le gène *LAR* à 1,3 Mb, codant pour une leucoanthocyanidine réductase et potentiellement responsable de nombreux QTLs de polyphénols détectés sur le haut du LG16 (Khan, Schaart, et al. 2012), ou un gène codant pour une pectine méthylestérase à 5,5 Mb, potentiellement responsable du maintien de la cohésion cellulaire des cellules de parenchyme, empêchant ainsi le développement de la farinosité (Mikol Segonne et al. 2014). Cette région apparaît donc comme un hotspot incontournable lors des programmes de sélection visant à améliorer la qualité du fruit chez le pommier.

Discussion générale et perspectives

L'exploration des bases génétiques des caractères agronomiques des plantes cultivées représente un enjeu majeur de la génétique quantitative moderne. La connaissance aussi exhaustive que possible des loci impliqués dans la variation phénotypique, de leur contribution relative à cette variation, de leurs éventuelles interactions, de leur organisation spatiale sur le génome, et *in fine* des fonctions des gènes sous-jacents permet de mieux comprendre l'architecture génétique globale de ces caractères et de mieux raisonner leur sélection. Depuis plusieurs années, la génétique d'association est proposée comme une méthode alternative à la cartographie génétique par analyse de liaison pour identifier davantage de loci impliqués et les localiser plus finement sur le génome. La génétique d'association permet en effet l'étude de populations dont les individus ne sont pas forcément apparentés entre eux, ce qui représente un grand avantage en particulier pour l'étude d'espèces pérennes chez lesquelles la cartographie par liaison est assez lourde à mettre en œuvre, et l'accumulation d'un grand nombre d'évènements de recombinaison est assez longue à obtenir par rapport aux espèces annuelles. Chez le pommier, les études de cartographie ont permis la localisation, bien qu'assez imprécise dans la plupart des cas, de nombreux facteurs génétiques responsables de la variation de traits phénotypiques d'intérêt, mais les populations étudiées ne représentaient qu'une faible proportion de la diversité génétique disponible. Le travail réalisé au cours de cette thèse est donc un projet pionnier pour le pommier. Il s'agit en effet de la première étude d'association de cette ampleur réalisée à ce jour sur cette espèce. Ce projet a consisté en l'étude de core collections de variétés anciennes de pommiers dans le but de localiser plus finement, avec potentiellement l'identification de gènes candidats, les régions génomiques associées à des caractères de qualité du fruit tels que l'acidité, le croquant ou encore l'amertume, et à la résistance du pommier à la tavelure et au feu bactérien. Ce travail n'a pu être réalisé que grâce à la construction, bien que tardive, d'une puce de génotypage SNP à haute densité (480k) réalisée dans le cadre du projet européen FruitBreedomics porté par l'IRHS.

L'étude du déséquilibre de liaison dans différentes populations et à l'aide de différentes techniques de génotypage a soulevé plusieurs points de discussion. Premièrement, la visualisation des patterns de déséquilibre de liaison grâce aux données de génotypage issues de la puce 480k SNPs a montré que la représentation actuelle de l'organisation du génome du pommier en scaffolds composés de contigs (carte physique) était encore largement entachée d'erreurs dans les différents endroits du génome que nous avons eu l'occasion d'étudier plus en détails ici. Le séquençage insuffisamment profond et étayé de la variété hétérozygote « Golden Delicious » pour générer les premières versions du génome du pommier s'avère à la longue extrêmement pénalisant pour les études genome-wide comme celles entreprises au cours de cette thèse. Une étude plus approfondie de ces patterns de DL permettrait dans un premier temps d'éliminer les contigs mal positionnés à l'intérieur des scaffolds afin d'obtenir une meilleure base pour la carte physique du génome. Dans une deuxième étape, les données de re-séquençage de longs fragments d'ADN d'un haploïde doublé de « Golden Delicious » devrait

permettre un meilleur assemblage des contigs et scaffolds et un meilleur ancrage sur la carte génétique de manière à obtenir une estimation plus juste des distances physiques entre les SNPs de la puce 480k. Une version significativement améliorée du génome permettra incontestablement une meilleure estimation de l'étendue moyenne du déséquilibre de liaison ainsi que son étude locale dans les zones identifiées au cours des études d'association, allant de pair avec une identification plus précise des facteurs recherchés, des gènes candidats voisins et des haplotypes locaux. Deuxièmement, cette étude a montré le peu d'influence de la taille de la population d'étude sur l'étendue du déséquilibre de liaison lors de l'utilisation d'une même technique de génotypage. L'étude d'une large core collection et d'une mini core collection constituée d'un sous-ensemble d'individus de la première a en effet montré des résultats similaires. L'estimation de l'étendue du déséquilibre de liaison à l'échelle du génome entier peut donc se faire sur un sous ensemble des données de génotypage ce qui permettrait un gain de temps et de moyens. Enfin, l'estimation de l'étendue du déséquilibre de liaison à l'aide de trois techniques de génotypage, une puce 8k SNPs, une puce 480k SNPs et du re-séquençage de gènes candidats positionnels, a montré que la technique influe grandement sur les valeurs estimées, cela en grande partie dû à la distance moyenne entre marqueurs. Les deux puces de génotypage n'ont ainsi permis qu'une estimation imprécise de l'étendue du déséquilibre de liaison. Les données de re-séquençage en revanche ont permis une étude locale à très fine échelle du déséquilibre de liaison dans des régions génomiques choisies et permettent de relever deux conclusions. La première est bien connue et dérive de l'hétérogénéité de distribution des évènements de recombinaison à l'échelle du génome. En effet, l'estimation de l'étendue du déséquilibre de liaison à l'échelle du génome entier ne reflète que très partiellement la réalité, puisque certaines régions montrent des étendues sur lesquelles les marqueurs sont en fort DL plus grandes que d'autres. La recherche de gènes candidats à l'intérieur d'une fenêtre de distance, constante à l'échelle du génome, correspondant à deux fois l'étendue du DL pour un $r^2 < 0,2$ peut donc ne conduire à l'identification d'aucun gène candidat valable si l'on se situe dans une région où l'étendue réelle du DL est plus grande. Une étude plus locale du DL conduirait à une meilleure estimation de la taille de la fenêtre à considérer et permettrait probablement l'identification de plus de gènes candidats fonctionnels. La seconde est que les données de re-séquençage permettent une estimation non biaisée de l'étendue du déséquilibre de liaison, contrairement aux puces de génotypage dont les contraintes lors du design engendrent des biais d'échantillonnage des SNPs, notamment au niveau de la distribution de la fréquence des allèles mineurs ce qui a une grande influence sur l'estimation du déséquilibre de liaison.

L'étude de deux core collections de pommiers à cidre et à couteau a permis grâce à des analyses de différenciation génétique, d'association et de génétique des populations, d'identifier des régions génomiques responsables des différences observées entre les deux types variétaux de pommes. Une partie des régions identifiées par l'approche de génétique d'association co-localisait avec des QTLs de

teneur en polyphénols, ce qui est conforté par les données de la littérature stipulant qu'en moyenne, les pommes à cidre sont plus riches en polyphénols que les pommes à couteau. Mais force a été de constater que c'était surtout l'étude d'association ciblant un caractère précis, ici l'amertume, qui a permis de trouver une bonne concordance entre les associations et les QTLs déjà publiés. Les autres régions génomiques ne co-localisant avec aucun QTL donnent cependant une base pour de futures études permettant de mieux comprendre l'architecture des caractères phénotypiques différenciant les deux types variétaux de pommes et de les identifier. Les résultats des analyses de génétique des populations sont cependant à considérer avec précaution, en raison de l'absence dans les données analysées de bon nombre de polymorphismes à basse fréquence.

Enfin, l'étude de génétique d'association portant sur l'identification de caractères de qualité du fruit et de résistance à la tavelure et au feu bactérien dans une core collection de variétés anciennes de pommiers à couteau a permis la localisation de régions génomiques impliquées dans la variation de certains de ces caractères. C'est le cas notamment d'une région génomique localisée dans le haut du groupe de liaison 16 qui est associée à plusieurs caractères de qualité du fruit tels que l'acidité, l'amertume et la fermeté. Il conviendra donc à l'avenir de décortiquer plus finement cette région génomique et d'envisager des études de phylogénie pour tenter de retracer l'histoire évolutive et de sélection de cette portion du génome. La co-localisation parfaite entre l'association la plus significative pour l'acidité dans notre étude et le gène candidat (transporteur de malate) déjà identifié dans une étude précédente sert de « preuve de concept » à la démarche de génétique d'association telle que nous l'avons mise en œuvre chez le pommier. Une autre région localisée sur le groupe de liaison 6 a été identifiée lors de l'analyse des données de résistance au feu bactérien et a permis d'envisager l'implication potentielle du gène *NPR1* situé précisément dans cette région. Des études d'expression de ce gène ou d'autres gènes de la même voie peuvent être engagées pour valider le rôle de ce gène dans la modulation de la résistance/sensibilité des variétés inoculées avec *Erwinia amylovora*. Les marqueurs identifiés ici ne totalisent qu'une part réduite de la variation phénotypique observée et pour d'autres caractères, au contraire, aucune région n'a pu être identifiée. Ce phénomène, connu sous le nom de « missing heritability », peut être expliqué par plusieurs facteurs. Parmi ceux-ci on retrouve le fait que les marqueurs ne couvrent pas la totalité du génome, le fait que les facteurs génétiques recherchés puissent être en fréquence trop faible dans la population d'étude ou présents en grand nombre, chacun ne comptant que pour une faible partie de la variation phénotypique. Cela pourrait être résolu en utilisant des populations d'étude plus grandes, comme c'est le cas dans le projet FruitBreedomics dans lequel d'autres core collections européennes vont être analysées en plus de celle étudiée ici. Une autre explication possible réside dans le fait que les marqueurs génétiques ne sont peut-être pas suffisants pour capter tout le polymorphisme au sein des individus. Le développement de marqueurs

épigénétiques (bases méthylées) pourra potentiellement à l'avenir permettre de détecter des associations phénotype-marque épigénétique comme cela est envisagé chez l'homme avec les EWAS.

Pour conclure, ce projet de thèse a permis de montrer que l'approche de génétique d'association est une démarche prometteuse pour localiser finement certains facteurs génétiques contrôlant la variation de caractères cibles de la sélection chez le pommier. Outre l'acquisition d'une carte physique du génome de qualité suffisante, l'enjeu va surtout se situer au niveau de l'acquisition de données phénotypiques suffisamment précises et en quantité suffisante pour bien apprécier les caractères dans des conditions environnementales variées, de manière à pouvoir explorer leur déterminisme de manière aussi large que possible. A plus long terme, ce type de données va être utile non seulement pour la connaissance des déterminants génétiques mais aussi pour la mise en œuvre concrète de la sélection assistée par marqueurs, que celle-ci cible ces déterminants ou qu'elle soit réalisée plus globalement à l'échelle du génome *via* des approches de prédiction et sélection génomique qui commencent à être développées sur pommier au sein des équipes ResPom et VaDiPom de l'IRHS.

Références bibliographiques

Abd-El Rahman, T., M. El Oirdi, R. Gonzalez-Lamothe, and K. Bouarab. 2012. Necrotrophic pathogens use the salicylic acid signaling pathway to promote disease development in tomato. *Molecular Plant-Microbe Interactions* 25 (12):1584-1593.

Ackermann, J., M. Fischer, and R. Amado. 1992. Changes in sugars, acids, and amino acids during ripening and storage of apples (cv. Glockenapfel). *Journal of Agricultural and Food Chemistry* 40 (7):1131-1134.

Albrechtsen, A., F. C. Nielsen, and R. Nielsen. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* 27 (11):2534-2547.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*.

Alonso-Blanco, C. and B. Méndez-Vigo. 2014. Genetic architecture of naturally occurring quantitative traits in plants: an updated synthesis. *Current Opinion in Plant Biology* 18:37-43.

Altshuler, D., M. J. Daly, and E. S. Lander. 2008. Genetic mapping in human disease. *Science* 322 (5903):881-888.

Amaral, A. J., L. Ferretti, H. J. Megens, R. P. M. A. Crooijmans, H. Nie, S. E. Ramos-Onsins, M. Perez-Enciso, L. B. Schook, and M. A. M. Groenen. 2011. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS One* 6 (4):e14782.

Andersen, Ø. and K. R. Markham. 2006. *Flavonoids chemistry, biochemistry, and applications*. Boca Raton, FL: CRC, Taylor & Francis.

Axelsson, E., A. Ratnakumar, M. L. Arendt, K. Maqbool, M. T. Webster, M. Perloski, O. Liberg, J. M. Arnemo, A. Hedhammar, and K. Lindblad-Toh. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 495 (7441):360-364.

Bachlava, E., C. A. Taylor, S. Tang, J. E. Bowers, J. R. Mandel, J. M. Burke, and S. J. Knapp. 2012. SNP discovery and development of a high-density genotyping array for sunflower. *PLoS One* 7 (1):e29814.

Bai, Y., L. Dougherty, M. Li, G. Fazio, L. Cheng, and K. Xu. 2012. A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the *Ma* locus is associated with low fruit acidity in apple. *Molecular Genetics and Genomics* 287 (8):663-678.

Balding, D. J. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7 (10):781-791.

Barendse, W. 2011. Haplotype analysis improved evidence for candidate genes for intramuscular fat percentage from a genome wide association study of cattle. *PLoS One* 6 (12):e29601.

Barnaud, A., V. Laucou, P. This, T. Lacombe, and A. Doligez. 2010. Linkage disequilibrium in wild French grapevine, *Vitis vinifera* L. subsp. *silvestris*. *Heredity* 104 (5):431-437.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21 (2):263-265.

Bassil, N. V., T. M. Davis, H. Zhang, S. Ficklin, M. Mittmann, T. Webster, L. Mahoney, D. Wood, E. S. Alperin, and U. R. Rosyara. 2015. Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. *BMC Genomics* 16 (1):155.

Beaumont, M. A. and D. J. Balding. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* 13 (4):969-980.

Belkhir, K., P. Borsa, L. Chikhi, N. Raufaste, and F. Bonhomme. 1996-2004. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. *Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier (France)*.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*:289-300.

Bianco, L., A. Cestaro, D. J. Sargent, E. Banchi, S. Derdak, M. Di Guardo, S. Salvi, J. Jansen, R. Viola, I. Gut, F. Laurens, D. Chagne, R. Velasco, E. van de Weg, and M. Troggio. 2014. Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus* x *domestica* Borkh). *PLoS One* 9 (10):e110377.

Billotte, N., M. F. Jourjon, N. Marseillac, A. Berger, A. Flori, H. Asmady, B. Adon, R. Singh, B. Nouy, and F. Potier. 2010. QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics* 120 (8):1673-1687.

Bink, M. C. A. M., M. P. Boer, C. J. F. Ter Braak, J. Jansen, R. E. Voorrips, and W. E. Van de Weg. 2008. Bayesian analysis of complex traits in pedigreed plant populations. *Euphytica* 161 (1-2):85-96.

Bink, M. C. A. M., J. Jansen, M. Madduri, R. E. Voorrips, C. E. Durel, A. B. Kouassi, F. Laurens, F. Mathis, C. Gessler, and D. Gobbin. 2014. Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. *Theoretical and Applied Genetics* 127 (5):1073-1090.

Bink, M. C. A. M., P. Uimari, M. Sillanpää, L. Janss, and R. Jansen. 2002. Multiple QTL mapping in related plant populations via a pedigree-analysis approach. *Theoretical and Applied Genetics* 104 (5):751-762.

Blanc, G., A. Charcosset, B. Mangin, A. Gallais, and L. Moreau. 2006. Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. *Theoretical and Applied Genetics* 113 (2):206-224.

Blankenship, S. M. 1987. Night-temperature effects on rate of apple fruit maturation and fruit quality. *Scientia horticulturae* 33 (3):205-212.

Bolormaa, S., B. J. Hayes, K. Savin, R. Hawken, W. Barendse, P. F. Arthur, R. M. Herd, and M. E. Goddard. 2011. Genome-wide association studies for feedlot and growth traits in cattle. *Journal of Animal Science* 89 (6):1684-1697.

Boré, J. M. and J. Fleckinger. 1997. Pommiers à cidre (variétés de France).

Bouchet, S., B. Servin, P. Bertin, D. Madur, V. Combes, F. Dumas, D. Brunel, J. Laborde, A. Charcosset, and S. Nicolas. 2013. Adaptation of maize to temperate climates: mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the *Vgt2* (*ZCN8*) locus. *PLoS One* 8 (8):e71377.

Bowen, J. K., C. H. Mesarich, V. G. M. Bus, R. M. Beresford, K. M. Plummer, and M. D. Templeton. 2011. *Venturia inaequalis*: the causal agent of apple scab. *Molecular Plant Pathology* 12 (2):105-122.

Brachi, B., G. Morris, and J. Borevitz. 2011. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology* 12 (10):232-239.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23 (19):2633-2635.

Brent, K. J and D. W Hollomon. 1995. *Fungicide resistance in crop pathogens: How can it be managed?*: Brussels: Fungicide Resistance Action Committee.

Brisset, M. N., S. Cesbron, S. V. Thomson, and J. P. Paulin. 2000. Acibenzolar-S-methyl induces the accumulation of defense-related enzymes in apple and protects from fire blight. *European Journal of Plant Pathology* 106 (6):529-536.

Brooks, A. N. 1926. Studies of the epidemiology and control of fireblight of apple. *Phytopathology* 16:665-696.

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley, and D. B. Neale. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the United States of America* 101 (42):15255-15260.

Browning, S. R. and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* 81 (5):1084-1097.

Bus, V. G. M., F. N. D. Laurens, W. E. van de Weg, R. L. Rusholme, E. H. A. Rikkerink, S. E. Gardiner, H. Bassett, L. P. Kodde, and K. M. Plummer. 2005. The *Vh8* locus of a new gene-for-gene interaction between *Venturia inaequalis* and the wild apple *Malus sieversii* is closely linked to the *Vh2* locus in *Malus pumila* R12740-7A. *New Phytologist* 166 (3):1035-1049.

Bus, V. G., E. H. Rikkerink, V. Caffier, C. E. Durel, and K. M. Plummer. 2011. Revision of the nomenclature of the differential host-pathogen interactions of *Venturia inaequalis* and *Malus*. *Annual Review of Phytopathology* 49:391-413.

Caffier, V., A. Patocchi, P. Expert, M. N. Bellanger, C. E. Durel, M. Hilber-Bodmer, G. Broggini, R. Groenwold, and V. Bus. 2014. Virulence characterization of *Venturia inaequalis* reference isolates on the differential set of *Malus* hosts. *Plant Disease* 99:370-375.

Caldwell, K. S., J. Russell, P. Langridge, and W. Powell. 2006. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* 172 (1):557-567.

Calenge, F., D. Drouet, C. Denance, W. E. Van de Weg, M. N. Brisset, J. P. Paulin, and C. E. Durel. 2005. Identification of a major QTL together with several minor additive or epistatic QTLs for resistance to fire blight in apple in two related progenies. *Theoretical and Applied Genetics* 111 (1):128-135.

Calenge, F., A. Faure, M. Goerre, C. Gebhardt, W. E. Van de Weg, L. Parisi, and C. E. Durel. 2004. Quantitative Trait Loci (QTL) analysis reveals both broad-spectrum and isolate-specific QTL for scab resistance in an apple progeny challenged with eight isolates of Venturia inaequalis. *Phytopathology* 94 (4):370-379.

Campo del, G. , J. I. Santos, I. Berregi, and A. Munduate. 2005. Differentiation of Basque cider apple juices from different cultivars by means of chemometric techniques. *Food Control* 16 (6):549-555.

Camus-Kulandaivelu, L., L. M. Chevin, C. Tollon-Cordet, A. Charcosset, D. Manicacci, and M. I. Tenaillon. 2008. Patterns of molecular evolution associated with two selective sweeps in the *Tb1–Dwarf8* region in maize. *Genetics* 180 (2):1107-1121.

Cao, H., J. Glazebrook, J. D. Clarke, S. Volko, and X. Dong. 1997. The *Arabidopsis NPR1* gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell* 88 (1):57-63.

Casals, F., A. Hodgkinson, J. Hussin, Y. Idaghdour, V. Bruat, T. de Maillard, J. C. Grenier, E. Gbeha, F. F. Hamdan, and S. Girard. 2013. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genetics* 9 (9):e1003815.

Celton, J. M., S. Martinez, M. J. Jammes, A. Bechti, S. Salvi, J. M. Legave, and E. Costes. 2011. Deciphering the genetic determinism of bud phenology in apple progenies: a new insight into chilling and heat requirement effects on flowering dates and positional candidate genes. *The New phytologist* 192 (2):378-392.

Chagne, D., R. N. Crowhurst, M. Troggio, M. W. Davey, B. Gilmore, C. Lawley, S. Vanderzande, R. P. Hellens, S. Kumar, A. Cestaro, R. Velasco, D. Main, J. D. Rees, A. Iezzoni, T. Mockler, L. Wilhelm, E. Van de Weg, S. E. Gardiner, N. Bassil, and C. Peace. 2012. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS One* 7 (2):e31745.

Chagne, D., C. Krieger, M. Rassam, M. Sullivan, J. Fraser, C. André, M. Pindo, M. Troggio, S. E. Gardiner, R. A. Henry, A. C. Allan, T. K. McGhie, and W. A. Laing. 2012. QTL and candidate gene mapping for polyphenolic composition in apple fruit. *BMC Plant Biology* 12 (1):1-16.

Chapman, N. H., J. Bonnet, L. Grivet, J. Lynn, N. Graham, R. Smith, G. Sun, P. G. Walley, M. Poole, and M. Causse. 2012. High-resolution mapping of a fruit firmness-related quantitative trait locus in tomato reveals epistatic interactions associated with a complex combinatorial locus. *Plant Physiology* 159 (4):1644-1657.

Cheeseman, K., J. Ropars, P. Renault, J. Dupont, J. Gouzy, A. Branca, A. L. Abraham, M. Ceppi, E. Conseiller, R. Debuchy, F. Malagnac, A. Goarin, P. Silar, S. Lacoste, E. Sallet, A. Bensimon, T. Giraud, and Y. Brygoo. 2014. Multiple recent horizontal transfers of a large genomic region in cheese making fungi. *Nature Communications* 5:2876.

Chevalier, M., Y. Lespinasse, and S. Renaudin. 1991. A microscopic study of the different classes of symptoms coded by the *Vf* gene in apple for resistance to scab (*Venturia inaequalis*). *Plant Pathology* 40 (2):249-256.

Clark, Andrew G., Melissa J. Hubisz, Carlos D. Bustamante, Scott H. Williamson, and Rasmus Nielsen. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* 15 (11):1496-1502.

Collard, B. C. Y., M. Z. Z. Jahufer, J. B. Brouwer, and E. C. K. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* 142 (1-2):169-196.

Collard, B. C. Y. and D. J. Mackill. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1491):557-572.

Comadran, J., L. Ramsay, K. MacKenzie, P. Hayes, T. J. Close, G. Muehlbauer, N. Stein, and R. Waugh. 2011. Patterns of polymorphism and linkage disequilibrium in cultivated barley. *Theoretical and Applied Genetics* 122 (3):523-531.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21 (18):3674-3676.

Cook, D. E., T. G. Lee, X. Guo, S. Melito, K. Wang, A. M. Bayless, J. Wang, T. J. Hughes, D. K. Willis, and T. E. Clemente. 2012. Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* 338 (6111):1206-1209.

Cornille, A., T. Giraud, M. J. Smulders, I. Roldan-Ruiz, and P. Gladieux. 2014. The domestication and evolutionary ecology of apples. *Trends in Genetics* 30 (2):57-65.

Cornille, A., P. Gladieux, M. J. M. Smulders, I. Roldán-Ruiz, F. Laurens, B. Le Cam, A. Nersesyan, J. Clavel, M. Olonova, L. Feugey, I. Gabrielyan, X. G. Zhang, M. I. Tenaillon, and T. Giraud. 2012.

New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics* 8 (5):e1002703.

Cornuet, J. M. and G. Luikart. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144 (4):2001-2014.

Costa, F., C. P. Peace, S. Stella, S. Serra, S. Musacchi, M. Bazzani, S. Sansavini, and W. E. Van de Weg. 2010. QTL dynamics for fruit firmness and softening around an ethylene-dependent polygalacturonase gene in apple (*Malus* × *domestica* Borkh.). *Journal of Experimental Botany* 61 (11):3029-3039.

Courtois, B., A. Audebert, A. Dardou, S. Roques, T. Ghneim-Herrera, G. Droc, J. Frouin, L. Rouan, E. Gozé, and A. Kilian. 2013. Genome-wide association mapping of root traits in a japonica rice panel. *PLoS One* 8 (11):e78037.

Crow, T. J. 2011. The missing genes: what happened to the heritability of psychiatric disorders? *Molecular psychiatry* 16 (4):362-364.

Croxall, H. E., D. C. Gwynne, and J. E. E. Jenkins. 1952. The rapid assessment of apple scab on leaves. *Plant Pathology* 1 (2):39-41.

Dapena, E., M. Minarro, and M. D. Blazquez. 2005. Organic cider-apple production in Asturias (NW Spain). *IOBC-WPRS Bulletin* 28 (7):161.

de Bakker, P. I. W., R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. 2005. Efficiency and power in genetic association studies. *Nature Genetics* 37 (11):1217-1223.

Després, C., C. de Long, S. Glaze, E. Liu, and P. R. Fobert. 2000. The *Arabidopsis NPR1/NIM1* protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors. *The Plant Cell Online* 12 (2):279-290.

Dolejsi, E., B. Bodenstorfer, and F. Frommlet. 2014. Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian Information Criterion. *PLoS One* 9 (7):e103322.

Douglas, G. L. and T. R. Klaenhammer. 2010. Genomic evolution of domesticated microorganisms. *Annual Review of Food Science and Technology* 1:397-414.

Drouaud, J., C. Camilleri, P. Y. Bourguignon, A. Canaguier, A. Bérard, D. Vezon, S. Giancola, D. Brunel, V. Colot, B. Prum, H. Quesneville, and C. Mézard. 2006. Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". *Genome Research* 16 (1):106-114.

Dugé De Bernonville, T., M. Gaucher, V. Flors, S. Gaillard, J. P. Paulin, J. F. Dat, and M. N. Brisset. 2012. T3SS-dependent differential modulations of the jasmonic acid pathway in susceptible and resistant genotypes of *Malus* spp. challenged with *Erwinia amylovora*. *Plant Science* 188:1-9.

Dunemann, F., D. Ulrich, A. Boudichevskaia, C. Grafe, and W. E. Weber. 2009. QTL mapping of aroma compounds analysed by headspace solid-phase microextraction gas chromatography in the apple progeny 'Discovery' × 'Prima'. *Molecular Breeding* 23 (3):501-521.

Durel, C. E., F. Calenge, L. Parisi, W. E. van de Weg, L. P. Kodde, R. Liebhard, C. Gessler, M. Thiermann, F. Dunemann, and F. Gennari. 2004. An overview of the position and robustness of scab resistance QTLs and major genes by aligning of genetic maps in five apple progenies. *Acta Horticulturae* 663:135-140.

Durel, C. E., C. Denance, and M. N. Brisset. 2009. Two distinct major QTL for resistance to fire blight co-localize on linkage group 12 in apple genotypes 'Evereste' and *Malus floribunda* clone 821. *Genome* 52 (2):139-147.

Durel, C. E., L. Parisi, F. Laurens, W. E. Van de Weg, R. Liebhard, and M. F. Jourjon. 2003. Genetic dissection of partial resistance to race 6 of *Venturia inaequalis* in apple. *Genome* 46 (2):224-234.

Durrant, W. E. and X. Dong. 2004. Systemic acquired resistance. *Annual Review of Phytopathology* 42:185-209.

Ebel, R. C., E. L. Proebsting, and M. E. Patterson. 1993. Regulated deficit irrigation may alter apple maturity, quality, and storage life. *Horticultural Science* 28 (2):141-143.

Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra, S. C. González-Martínez, and D. B. Neale. 2010. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., *Pinaceae*). *Genetics* 185 (3):969-982.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6 (5):e19379.

Excoffier, L., M. Foll, and R. J. Petit. 2009. Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* 40 (1):481-501.

Fachal, L. and A. M. Dunning. 2015. From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Current Opinion in Genetics & Development* 30 (0):32-41.

Fan, W. and X. Dong. 2002. In vivo interaction between *NPR1* and transcription factor TGA2 leads to salicylic acid–mediated gene activation in *Arabidopsis*. *The Plant Cell Online* 14 (6):1377-1389.

Fan, Z., Z. ChaoHong, X. Yan, and W. YueJin. 2010. Cloning and prokaryotic expression of Zaosu pear nonexpressor of pathogenesis-related genes 1 gene (*NPR1*) in *Escherichia coli*. *Journal of Agricultural Biotechnology* 18 (1):18-23.

Fay, J. C. and C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155 (3):1405-1413.

Flint-Garcia, S. A., J. M. Thornsberry, and E. S. Buckler. 2003. Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54:357-374.

Foll, M. and O. Gaggiotti. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180 (2):977-993.

Forsline, P. L., H. S. Aldwinckle, E. E. Dickson, J. J. Luby, and S. C. Hokanson. 2003. Collection, maintenance, characterization, and utilization of wild apples of Central Asia. *Horticultural Reviews. Westport Then New York* 29:1-62.

Frankel, O. H. and A. H. D. Brown. 1984. Plant genetic resources today: a critical appraisal. *Crop Genetic Resources: Conservation & Evaluation*.

Frankel, R. and E. Galun. 1977. *Pollination mechanisms, reproduction and plant breeding*: Berlin, Heidelberg, New York: Springer-Verlag.

Fu, Y. X. and W. H. Li. 1993. Statistical tests of neutrality of mutations. *Genetics* 133 (3):693-709.

Gallavotti, A., Q. Zhao, J. Kyozuka, R. B. Meeley, M. K. Ritter, J. F. Doebley, M. E. Pe, and R. J. Schmidt. 2004. The role of *barren stalk1* in the architecture of maize. *Nature* 432 (7017):630-635.

Gao, X., L. C. Becker, D. M. Becker, J. D. Starmer, and M. A. Province. 2010. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic epidemiology* 34 (1):100-105.

Gaucher, M. 2012. Identification par approches comparatives de facteurs moléculaires associés à la résistance ou à la sensibilité du pommier (*Malus* x *domestica*) à *Erwinia amylovora*, agent du feu bactérien.

Gaucher, M., T. Dugéde Bernonville, S. Guyot, J. F. Dat, and M. N. Brisset. 2013. Same ammo, different weapons: Enzymatic extracts from two apple genotypes with contrasted susceptibilities to fire blight (*Erwinia amylovora*) differentially convert phloridzin and phloretin in vitro. *Plant Physiology and Biochemistry* 72:178-189.

Gessler, C., A. Patocchi, S. Sansavini, S. Tartarini, and L. Gianfranceschi. 2006. *Venturia inaequalis* resistance in apple. *Critical Reviews in Plant Sciences* 25 (6):473-503.

Gibson, G. 2012. Rare and common variants: twenty arguments. *Nature Reviews Genetics* 13 (2):135-145.

Glazebrook, J. 2005. Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annual Review of Phytopathology* 43:205-227.

Goddard, M. E. and B. J. Hayes. 2007. Genomic selection. *Journal of Animal Breeding and Genetics* 124 (6):323-330.

Guitton, B., J. J. Kelner, R. Velasco, S. E. Gardiner, D. Chagne, and E. Costes. 2012. Genetic control of biennial bearing in apple. *Journal of Experimental Botany* 63 (1):131-149.

Gupta, P. K., S. Rustgi, and P. L. Kulwal. 2005. Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Molecular Biology* 57 (4):461-485.

Gygax, M., L. Gianfranceschi, R. Liebhard, M. Kellerhals, C. Gessler, and A. Patocchi. 2004. Molecular markers linked to the apple scab resistance gene *Vbj* derived from *Malus baccata jackii*. *Theoretical and Applied Genetics* 109 (8):1702-1709.

Hammond-Kosack, K. E. and J. D. Jones. 1996. Resistance gene-dependent plant defense responses. *The Plant Cell* 8 (10):1773.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science* 92 (2):433-443.

Helyar, S. J., J. Hemmer-Hansen, D. Bekkevold, M. I. Taylor, R. Ogden, M. T. Limborg, A. Cariani, G. E. Maes, E. Diopere, G. R. Carvalho, and E. E. Nielsen. 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* 11:123-136.

Heuertz, M., E. De Paoli, T. Källman, H. Larsson, I. Jurman, M. Morgante, M. Lascoux, and N. Gyllenstrand. 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* 174 (4):2095-2105.

Hill, W. G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38 (6):226-231.

Holb, I. J. 2006. Effect of six sanitation treatments on leaf litter density, ascospore production of *Venturia inaequalis* and scab incidence in integrated and organic apple orchards. *European Journal of Plant Pathology* 115 (3):293-307.

Holland, J. B. 2007. Genetic architecture of complex traits in plants. *Current Opinion in Plant Biology* 10 (2):156-161.

Huang, X. and B. Han. 2014. Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology* 65:531-551.

Hyten, D. L., I. Y. Choi, Q. Song, R. C. Shoemaker, R. L. Nelson, J. M. Costa, J. E. Specht, and P. B. Cregan. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175 (4):1937-1944.

Imamura, M. and S. Maeda. 2011. Genetics of type 2 diabetes: the GWAS era and future perspectives. *Endocrine Journal* 58 (9):723-739.

Ingvarsson, P. K. 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* 169 (2):945-953.

Ingvarsson, P. K. and N. R. Street. 2011. Association genetics of complex traits in plants. *The New phytologist* 189 (4):909-922.

Ingvarsson, Pär K and Nathaniel R Street. 2011. Association genetics of complex traits in plants. *New Phytologist* 189 (4):909-922.

Iwata, H., T. Hayashi, S. Terakami, N. Takada, Y. Sawamura, and T. Yamamoto. 2013. Potential assessment of genome-wide association study and genomic selection in Japanese pear *Pyrus pyrifolia*. *Breeding Science* 63 (1):125-140.

Jaeger, S. R., Z. Andani, I. N. Wakeling, and H. J. H. MacFie. 1998. Consumer preferences for fresh and aged apples: a cross-cultural comparison. *Food Quality and Preference* 9 (5):355-366.

Janick, J. 2005. The origins of fruits, fruit growing, and fruit breeding. In *Plant Breeding Reviews*: John Wiley & Sons, Inc.

Jonkers, H. 1979. Biennial bearing in apple and pear: a literature survey. *Scientia horticulturae* 11 (4):303-317.

Jugdé, H., D. Nguy, I. Moller, J. M. Cooney, and R. G. Atkinson. 2008. Isolation and characterization of a novel glycosyltransferase that converts phloretin to phlorizin, a potent antioxidant in apple. *FEBS journal* 275 (15):3804-3814.

Juniper, B. E. and D. J. Mabberley. 2006. *The story of the apple*. Portland, Or.: Timber Press.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 42 (4):348-354.

Kang, Y. U. N., M. Sakiroglu, N. Krom, J. Stanton-Geddes, M. Wang, Y. C. Lee, N. D. Young, and M. Udvardi. 2015. Genome-wide association of drought-related and biomass traits with HapMap SNPs in *Medicago truncatula*. *Plant, Cell & Environment*.

Keitt, G. W. and D. H. Palmiter. 1937. Hetrotallism in *Venturia inaequalis*. *Science* 85:498.

Kenis, K., J. Keulemans, and M. W. Davey. 2008. Identification and stability of QTLs for fruit quality traits in apple. *Tree Genetics & Genomes* 4 (4):647-661.

Khan, M. A., B. Duffy, C. Gessler, and A. Patocchi. 2006. QTL mapping of fire blight resistance in apple. *Molecular Breeding* 17 (4):299-306.

Khan, M. A., C. E. Durel, B. Duffy, D. Drouet, M. Kellerhals, C. Gessler, and A. Patocchi. 2007. Development of molecular markers linked to the 'Fiesta' linkage group 7 major QTL for fire blight resistance and their application for marker-assisted selection. *Genome* 50 (6):568-577.

Khan, S. A., P. Y. Chibon, R. C. H. de Vos, B. A. Schipper, E. Walraven, J. Beekwilder, T. van Dijk, R. Finkers, R. G. F. Visser, and E. W. van de Weg. 2012. Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16. *Journal of Experimental Botany* 63 (8):2895-2908.

Khan, S. A., J. Schaart, J. Beekwilder, A. Allan, Y. Tikunov, E. Jacobsen, and H. Schouten. 2012. The mQTL hotspot on linkage group 16 for phenolic compounds in apple fruits is probably the result of a leucoanthocyanidin reductase gene at that locus. *BMC Research Notes* 5 (1):618.

Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark, S. Ossowski, J. R. Ecker, D. Weigel, and M. Nordborg. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* 39 (9):1151-1155.

Kim, Y. and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167 (3):1513-1524.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* 217 (5129):624-626.

Kollar, A. 1996. Evidence for loss of ontogenetic resistance of apple leaves against *Venturia inaequalis*. *European Journal of Plant Pathology* 102 (8):773-778.

Köller, W. 1994. Chemical control of apple scab-status quo and future. *Norwegian Journal of Agricultural Sciences (Norway)* 17:149-170.

Köller, W., D. M. Parker, W. W. Turechek, C. Avila-Adame, and K. Cronshaw. 2004. A two-phase resistance response of *Venturia inaequalis* populations to the QoI fungicides kresoxim-methyl and trifloxystrobin. *Plant Disease* 88 (5):537-544.

Korte, A. and A. Farlow. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* 9 (1):29.

Kraakman, A. T. W., R. E. Niks, P. M. M. M. Van den Berg, P. Stam, and F. A. Van Eeuwijk. 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* 168 (1):435-446.

Kranis, A., A. Gheyas, C. Boschiero, F. Turner, L. Yu, S. Smith, R. Talbot, A. Pirani, F. Brew, P. Kaiser, P. Hocking, M. Fife, N. Salmon, J. Fulton, T. Strom, G. Haberer, S. Weigend, R. Preisinger, M. Gholami, S. Qanbari, H. Simianer, K. Watson, J. Woolliams, and D. Burt. 2013. Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* 14 (1):59.

Kumar, S., D. Chagné, M. C. A. M. Bink, R. K. Volz, C. Whitworth, and C. Carlisle. 2012. Genomic selection for fruit quality traits in apple (*Malus*×*domestica* Borkh.). *PLoS One* 7 (5):e36674.

Lachance, J. and S. A. Tishkoff. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 35 (9):780-786.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution*:314-334.

Larkin, D. M., H. D. Daetwyler, A. G. Hernandez, C. L. Wright, L. A. Hetrick, L. Boucek, S. L. Bachman, M. R. Band, T. V. Akraiko, M. Cohen-Zinder, J. Thimmapuram, I. M. Macleod, T. T. Harkins, J. E. McCague, M. E. Goddard, B. J. Hayes, and H. A. Lewin. 2012. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proceedings of the National Academy of Sciences* 109 (20):7693-7698.

Larson, G. and J. Burger. 2013. A population genetics view of animal domestication. *Trends in Genetics* 29 (4):197-205.

Lassois, L., C. Denance, E. Ravon, A. Guyader, R. Guisnel, L. Hibrand Saint-Oyan, C. Poncet, P. Lasserre-Zuber, L. Feugey, and C. E. Durel. 2015. Genetic diversity, population structure, parentage analysis and construction of core collections in the French apple germplasm based on SSR markers. submitted.

Lê, S., J. Josse, and F. Husson. 2008. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software* 25 (1):1-18.

Lê Van, A., V. Caffier, P. Lasserre-Zuber, A. Chauveau, D. Brunel, B. Le Cam, and C. E. Durel. 2013. Differential selection pressures exerted by host resistance quantitative trait loci on a pathogen population: a case study in an apple × *Venturia inaequalis* pathosystem. *The New phytologist* 197 (3):899-908.

Lea, A. G. H. and J. R. Piggott. 2003. *Fermented beverage production*. 2nd ed. New York: Kluwer Academic/Plenum Publishers.

Lee, Y. G., N. Jeong, J. H. Kim, K. Lee, K. H. Kim, A. Pirani, B. K. Ha, S. T. Kang, B. S. Park, and J. K. Moon. 2014. Development, validation, and genetic analysis of a large soybean SNP genotyping array. *The Plant Journal* 81 (4):625-636.

Lespinasse, Y. and J.P. Paulin. 1990. Apple breeding programme for fire blight resistance : strategy used and first results *Acta Horticulturae (ISHS)* 273:285-296.

Lespinasse, Y., F. Rousselle-Bourgeois, and P. Rousselle. 1992. Breeding apple tree: aims and methods. Paper read at Proceedings of the Joint Conference of the EAPR Breeding & Varietal Assessment Section and the EUCARPIA Potato Section, Landerneau, France, 12-17 January 1992.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49 (1):49.

Li, M. X., J. M. Y. Yeung, S. S. Cherny, and P. C. Sham. 2012. Evaluating the effective numbers of independent tests and significant *p*-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics* 131 (5):747-756.

Li, X. W., X. Q. Meng, H. J. Jia, M. L. Yu, R. J. Ma, L. R. Wang, K. Cao, Z. J. Shen, L. Niu, J. B. Tian, M. J. Chen, M. Xie, P. Arus, Z. S. Gao, and M. J. Aranzana. 2013. Peach genetic resources: diversity, population structure and linkage disequilibrium. *BMC Genetics* 14 (1):84.

Liebhard, R., M. Kellerhals, W. Pfammatter, M. Jertmini, and C. Gessler. 2003. Mapping quantitative physiological traits in apple (*Malus* × *domestica* Borkh.). *Plant Molecular Biology* 52 (3):511-526.

Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li, P. J. Bradbury, M. A. Gore, E. S. Buckler, and Z. Zhang. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28 (18):2397-2399.

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman. 2011. FaST linear mixed models for genome-wide association studies. *Nature methods* 8 (10):833-835.

Loh, P. R., G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjalmsson, H. K. Finucane, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, and N. Patterson. 2015. Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics* 47:284-290.

Longhi, S., M. T. Hamblin, L. Trainotti, C. P. Peace, R. Velasco, and F. Costa. 2013. A candidate gene based approach validates *Md-PG1* as the main responsible for a QTL impacting fruit texture in apple (*Malus* x *domestica* Borkh). *BMC Plant Biology* 13 (1):37.

Longhi, S., M. Moretto, R. Viola, R. Velasco, and F. Costa. 2012. Comprehensive QTL mapping survey dissects the complex fruit texture physiology in apple (*Malus* x *domestica* Borkh.). *Journal of Experimental Botany* 63 (3):1107-1121.

Lorenz, A. J., M. T. Hamblin, and J. L. Jannink. 2010. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS One* 5 (11):e14079.

MacHardy, W. E. 1996. *Apple scab: biology, epidemiology, and management*: The American Phytopathological Society Press. St. Paul.

MacHardy, W. E. and D. M. Gadoury. 1986. Patterns of ascospore discharge by *Venturia inaequalis*. *Phytopathology (USA)* 76:985-990.

MacHardy, W. E., D. M. Gadoury, and C. Gessler. 2001. Parasitic and biological fitness of *Venturia inaequalis*: relationship to disease management strategies. *Plant Disease* 85 (10):1036-1051.

Mackay, Trudy FC, Eric A Stone, and Julien F Ayroles. 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* 10 (8):565-577.

Maenhout, S., B. De Baets, and G. Haesaert. 2009. CoCoa: a software tool for estimating the coefficient of coancestry from multilocus genotype data. *Bioinformatics* 25 (20):2753-2754.

Maher, B. 2008. Personal genomes: The case of the missing heritability. *Nature News* 456 (7218):18-21.

Maliepaard, C., F. H. Alston, G. van Arkel, L. M. Brown, E. Chevreau, F. Dunemann, K. M. Evans, S. Gardiner, P. Guilford, A. W. van Heusden, J. Janse, F. Laurens, J. R. Lynn, A. G. Manganaris, A. P. M. den Nijs, N. Periam, E. Rikkerink, P. Roche, C. Ryder, S. Sansavini, H. Schmidt, S. Tartarini, J. J. Verhaegh, M. Vrielink-van Ginkel, and G. J. King. 1998. Aligning male and female linkage maps of apple (*Malus pumila* Mill.) using multi-allelic markers. *Theoretical and Applied Genetics* 97 (1-2):60-73.

Malnoy, M., Q. Jin, E. E. Borejsza-Wysocka, S. Y. He, and H. S. Aldwinckle. 2007. Overexpression of the apple *MpNPR1* gene confers increased disease resistance in *Malus* × *domestica*. *Molecular Plant-Microbe Interactions* 20 (12):1568-1580.

Malnoy, M., S. Martens, J. L. Norelli, M. A. Barny, G. W. Sundin, T. H. M. Smits, and B. Duffy. 2012. Fire blight: applied genomic insights of the pathogen and host. *Annual Review of Phytopathology* 50:475-494.

Mangin, B., A. Siberchicot, S. Nicolas, A. Doligez, P. This, and C. Cierco-Ayrolles. 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* 108 (3):285-291.

Manktelow, D. W. L., R. M. Beresford, T. A. Batchelor, and J. T. S. Walker. 1995. Use patterns and economics of fungicides for disease control in New Zealand apples. Paper read at International Conference on Integrated Fruit Production 422.

Manolio, T. A. 2009. Cohort studies and the genetics of complex disease. *Nature Genetics* 41 (1):5-6.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, and A. Chakravarti. 2009. Finding the missing heritability of complex diseases. *Nature* 461 (7265):747-753.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* 9 (5):356-369.

McTavish, E. J., J. E. Decker, R. D. Schnabel, J. F. Taylor, and D. M. Hillis. 2013. New World cattle show ancestry from multiple independent domestication events. *Proceedings of the National Academy of Sciences of the United States of America* 110 (15):1398-1406.

Meyer, R. S., A. E. DuVal, and H. R. Jensen. 2012. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *The New phytologist* 196 (1):29-48.

Micheletti, D., F. Costa, P. Baldi, M. Troggio, M. Pindo, M. Komjanc, M. Malnoy, A. Zharkikh, P. Magnago, R. Velasco, and S. Salvi. 2008. Linkage disequilibrium analysis to enable more efficient gene and QTL mapping in apple. *RGC4*.

Mikol Segonne, S. , M. Bruneau, J. M. Celton, S. Le Gall, M. Francin-Allami, M. Juchaux, F. Laurens, M. Orsel, and J. P. Renou. 2014. Multiscale investigation of mealiness in apple: an atypical role for a pectin methylesterase during fruit maturation. *BMC Plant Biology* 14 (1):1593.

Miller, P. W. 1929. Studies of fire blight of apple in Wisconsin. *Journal of Agriculture Research* 39 (8):579-621.

Morgan, J., A. Richards, and E. Dowle. 2002. *The new book of apples: the definitive guide to apples, including over 2000 varieties*: Ebury.

Mratinić, E. and M. Fotirić-Akšić. 2011. Evaluation of phenotypic diversity of apple (*Malus sp.*) germplasm through the principle component analysis. *Genetika* 43 (2):331-340.

Mudgett, M. B. 2005. New insights to the function of phytopathogenic bacterial type III effectors in plants. *Annual Review of Plant Biology* 56:509-531.

Muranty, H. 1996. Power of tests for quantitative trait loci detection using full-sib families in different schemes. *Heredity* 76 (2):156-165.

Myles, S. 2013. Improving fruit and wine: what does genomics have to offer? *Trends in Genetics* 29 (4):190-196.

Myles, S., J. Peiffer, P. J. Brown, E. S. Ersoz, Z. Zhang, D. E. Costich, and E. S. Buckler. 2009. Association mapping: critical considerations shift from genotyping to experimental design. *The Plant Cell Online* 21 (8):2194-2202.

Nava, G., A. R. Dechen, and G. R. Nachtigall. 2007. Nitrogen and potassium fertilization affect apple fruit quality in southern Brazil. *Communications in soil science and plant analysis* 39 (1-2):96-107.

Nielsen, R. 2004. Population genetic analysis of ascertained SNP data. *Hum Genomics* 1 (3):218-224.

———. 2005. Molecular signatures of selection. *Annual Review of Genetics* 39 (1):197-218.

Nielsen, R. and J. Signorovitch. 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology* 63 (3):245-255.

Noé, L. and G. Kucherov. 2005. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Research* 33 (Suppl 2):540-W543.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* 154 (2):923-929.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory, J. Hagenblad, M. Kreitman, J. N. Maloof, T. Noyes, P. J. Oefner, E. A. Stahl, and D. Weigel. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* 30 (2):190-193.

Nordborg, M. and S. Tavare. 2002. Linkage disequilibrium: what history has to tell us. *Trends in Genetics* 18 (2):83-90.

Oh, C. S. and S. V. Beer. 2005. Molecular genetics of *Erwinia amylovora* involved in the development of fire blight. *FEMS microbiology letters* 253 (2):185-192.

Oraguzie, N., P. Alspach, R. Volz, C. Whitworth, C. Ranatunga, R. Weskett, and R. Harker. 2009. Postharvest assessment of fruit quality parameters in apple using both instruments and an expert panel. *Postharvest Biology and Technology* 52 (3):279-287.

Oraguzie, N. C., A. J. Currie, J. Vollmann, H. Grausgruber, and P. Ruckenbauer. 2004. Apple breeding: exploitation of genetic variation via recurrent selection. Paper read at Genetic variation for plant breeding. Proceedings of the 17th EUCARPIA General Congress, Tulln, Austria, 8-11 September 2004.

Palaisa, K. A., M. Morgante, M. Williams, and A. Rafalski. 2003. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *The Plant Cell* 15 (8):1795-1806.

Paternoster, T., G. Défago, B. Duffy, C. Gessler, and I. Pertot. 2011. Selection of a biocontrol agent based on a potential mechanism of action: degradation of nicotinic acid, a growth factor essential for *Erwinia amylovora*. *International Microbiology* 13 (4):195-206.

Paulin, J. P., G. Lachaud, R. Chartier, and J.M. Bore. 1988. Sensibilité au feu bactérien de variétés de pommiers à cidre. Résultats de 3 années d'expérimentation.35.

Pe'er, I., Y. R. Chretien, P. I. de Bakker, J. C. Barrett, M. J. Daly, and D. M. Altshuler. 2006. Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *American Journal of Human Genetics* 78 (4):588-603.

Perazzolli, M., G. Malacarne, A. Baldo, L. Righetti, A. Bailey, P. Fontana, R. Velasco, and M. Malnoy. 2014. Characterization of resistance gene analogues (RGAs) in apple (*Malus* × *domestica* Borkh.) and their evolutionary history of the *Rosaceae* family. *PLoS One* 9 (2):e83844.

Pereira-Lorenzo, S., A. M. Ramos-Cabrer, and M. Fischer. 2009. Breeding apple (*Malus* x *domestica* Borkh). In *Breeding Plantation Tree Crops: Temperate Species*: Springer.

Perrier, X, A Flori, and F Bonnot. 2003. Data analysis methods. *Genetic diversity of cultivated tropical plants*:43-76.

Perrier, X. and J. P. Jacquemoud-Collet. 2006. DARwin software.

Pfaff, C. L., E. J. Parra, C. Bonilla, K. Hiester, P. M. McKeigue, M. I. Kamboh, R. G. Hutchinson, R. E. Ferrell, E. Boerwinkle, and M. D. Shriver. 2001. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *The American Journal of Human Genetics* 68 (1):198-207.

Phillips, R. L. and I. K. Vasil. 2001. *DNA-based markers in plants*. Vol. 6: Springer Science & Business Media.

Pinheiro, J. C. and D. M. Bates. 2000. *Mixed-effects models in S and S-PLUS*: Springer Science & Business Media.

Ponting, R. C., M. C. Drayton, N. O. I. Cogan, M. P. Dobrowolski, G. C. Spangenberg, K. F. Smith, and J. W. Forster. 2007. SNP discovery, validation, haplotype structure and linkage disequilibrium in full-length herbage nutritive quality genes of perennial ryegrass (*Lolium perenne* L.). *Molecular Genetics and Genomics* 278 (5):585-597.

Prada, D. 2009. Molecular population genetics and agronomic alleles in seed banks: searching for a needle in a haystack? *Journal of Experimental Botany* 60 (9):2541-2552.

Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. *Nature reviews. Genetics* 11 (7):459-463.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155 (2):945-959.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81 (3):559-575.

Rafalski, J. A. 2010. Association genetics in crop improvement. *Current Opinion in Plant Biology* 13 (2):174-180.

Ridge, P. G., S. Mukherjee, P. K. Crane, and J. S. K. Kauwe. 2013. Alzheimer's disease: analyzing the missing heritability. *PLoS One* 8 (11):e79771.

Rincent, R., L. Moreau, H. Monod, E. Kuhn, A. E. Melchinger, R. A. Malvar, J. Moreno-Gonzalez, S. Nicolas, D. Madur, V. Combes, F. Dumas, T. Altmann, D. Brunel, M. Ouzunova, P. Flament, P. Dubreuil, A. Charcosset, and T. Mary-Huard. 2014. Recovering power in association mapping panels with variable levels of linkage disequilibrium. *Genetics* 197 (1):375-387.

Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science-AAAS-Weekly Paper Edition* 273 (5281):1516-1517.

Rivas, M. A., M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, and N. Burtt. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics* 43 (11):1066-1073.

Rothammer, S., D. Seichter, M. Forster, and I. Medugorac. 2013. A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. *BMC Genomics* 14:908.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419 (6909):832-837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, and E. S. Lander. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449 (7164):913-918.

Salvi, S., G. Sponza, M. Morgante, D. Tomes, X. Niu, K. A. Fengler, R. Meeley, E. V. Ananiev, S. Svitashev, and E. Bruggemann. 2007. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proceedings of the National Academy of Sciences* 104 (27):11376-11381.

Salvi, S. and R. Tuberosa. 2005. To clone or not to clone plant QTLs: present and future challenges. *Trends in Plant Science* 10 (6):297-304.

Sanoner, P., S. Guyot, N. Marnet, D. Molle, and J. P. Drilleau. 1999. Polyphenol profiles of French cider apple varieties (*Malus domestica* sp.). *Journal of Agricultural and Food Chemistry* 47 (12):4847-4853.

Schaffner, S. and P. Sabeti. 2008. Evolutionary adaptation in the human lineage. *Nature Education* 1 (1):14.

Scheet, P. and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics* 78 (4):629-644.

Scheet, P. and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* 78:629-644.

Segura, V., C. Cilas, and E. Costes. 2008. Dissecting apple tree architecture into genetic, ontogenetic and environmental effects: mixed linear modelling of repeated spatial and temporal measures. *The New phytologist* 178 (2):302-314.

Segura, V., B. J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, and M. Nordborg. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44 (7):825-830.

Slatkin, M. 2009. Epigenetic inheritance and the missing heritability problem. *Genetics* 182 (3):845-850.

Smereka, K. J., W. E. MacHardy, and A. P. Kausch. 1987. Cellular differentiation in *Venturia inaequalis* ascospores during germination and penetration of apple leaves. *Canadian Journal of Botany* 65 (12):2549-2561.

Smith, J. M. and J. Haigh. 2007. The hitch-hiking effect of a favourable gene. *Genet Res* 89 (5-6):391-403.

Smith, R. B., E. C. Lougheed, E. W. Franklin, and I. McMillan. 1979. The starch iodine test for determining stage of maturation in apples. *Canadian journal of plant science* 59 (3):725-735.

Soller, M. 1994. Marker assisted selection - an overview. *Animal Biotechnology* 5 (2):193-207.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding. 2012. Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics* 91 (6):1011-1021.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 52 (3):506.

Stephens, M. and P. Donnelly. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* 73 (5):1162-1169.

Stich, B., A. E. Melchinger, M. Frisch, H. P. Maurer, M. Heckenberger, and J. C. Reif. 2005. Linkage disequilibrium in European elite maize germplasm investigated with SSRs. *Theoretical and Applied Genetics* 111 (4):723-730.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123 (3):585-595.

Thomson, S. V. 2000. Epidemiology of fire blight 2. *Fire Blight: the disease and its causative agent, Erwinia amylovora*:9.

Thornton, K. 2003. libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19 (17):2325-2327.

Trynka, G., K. A. Hunt, N. A. Bockett, J. Romanos, V. Mistry, A. Szperl, S. F. Bakker, M. T. Bardella, L. Bhaw-Rosun, and G. Castillejo. 2011. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics* 43 (12):1193-1201.

Turner, S. D. 2014. *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots*.

Tustin, D. S., P. M. Hirst, and I. J. Warrington. 1988. Influence of orientation and position of fruiting laterals on canopy light penetration, yield, and fruit quality of 'Granny Smith' apple. *Journal of the American Society for Horticultural Science (USA)*.

Unterseer, S., E. Bauer, G. Haberer, M. Seidel, C. Knaak, M. Ouzunova, T. Meitinger, T. M. Strom, R. Fries, and H. Pausch. 2014. A powerful tool for genome analysis in maize: development and evaluation of the high density 600k SNP genotyping array. *BMC Genomics* 15 (1):823.

Vaillancourt, L. and J. R. Hartman. 2000. Apple scab. *The Plant Health Instructor*.

van Berloo, R., A. Zhu, R. Ursem, H. Verbakel, G. Gort, and F. A. van Eeuwijk. 2008. Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. *Theoretical and Applied Genetics* 117 (1):89-101.

Velasco, R., A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana, S. K. Bhatnagar, M. Troggio, D. Pruss, S. Salvi, M. Pindo, P. Baldi, S. Castelletti, M. Cavaiuolo, G. Coppola, F. Costa, V. Cova, A. Dal Ri, V. Goremykin, M. Komjanc, S. Longhi, P. Magnago, G. Malacarne, M. Malnoy, D. Micheletti, M. Moretto, M. Perazzolli, A. Si-Ammour, S. Vezzulli, E. Zini, G. Eldredge, L. M. Fitzgerald, N. Gutin, J. Lanchbury, T. Macalma, J. T. Mitchell, J. Reid, B. Wardell, C. Kodira, Z. Chen, B. Desany, F. Niazi, M. Palmer, T. Koepke, D. Jiwan, S. Schaeffer, V. Krishnan, C. Wu, V. T. Chu, S. T. King, J. Vick, Q. Tao, A. Mraz, A. Stormo, K. Stormo, R. Bogden, D. Ederle, A. Stella, A. Vecchietti, M. M. Kater, S. Masiero, P. Lasserre, Y. Lespinasse, A. C. Allan, V. G. Bus, D. Chagne, R. N. Crowhurst, A. P. Gleave, E. Lavezzo, J. A. Fawcett, S. Proost, P. Rouze, L. Sterck, S. Toppo, B. Lazzari, R. P. Hellens, C. E. Durel, A. Gutin, R. E. Bumgarner, S. E. Gardiner, M. Skolnick, M. Egholm, Y. Van de Peer, F. Salamini, and R. Viola. 2010. The genome of the domesticated apple (*Malus* x *domestica* Borkh.). *Nature Genetics* 42 (10):833-839.

Verde, I., A. G. Abbott, S. Scalabrin, S. Jung, S. Shu, F. Marroni, T. Zhebentyayeva, M. T. Dettori, J. Grimwood, F. Cattonaro, A. Zuccolo, L. Rossini, J. Jenkins, E. Vendramin, L. A. Meisel, V. Decroocq, B. Sosinski, S. Prochnik, T. Mitros, A. Policriti, G. Cipriani, L. Dondini, S. Ficklin, D. M. Goodstein, P. Xuan, C. Del Fabbro, V. Aramini, D. Copetti, S. Gonzalez, D. S. Horner, R. Falchi, S. Lucas, E. Mica, J. Maldonado, B. Lazzari, D. Bielenberg, R. Pirona, M. Miculan, A. Barakat, R. Testolin, A. Stella, S. Tartarini, P. Tonutti, P. Arus, A. Orellana, C. Wells, D. Main, G. Vizzotto, H. Silva, F. Salamini, J. Schmutz, M. Morgante, and D. S. Rokhsar. 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., N. Bassil, S. Scalabrin, B. Gilmore, C. T. Lawley, K. Gasic, D. Micheletti, U. R. Rosyara, F. Cattonaro, and E. Vendramin. 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.

Verdu, C. F., S. Guyot, N. Childebrand, M. Bahut, J. M. Celton, S. Gaillard, P. Lasserre-Zuber, M. Troggio, D. Guilet, and F. Laurens. 2014. QTL analysis and candidate gene mapping for the polyphenol content in cider apple. *PLoS One* 9 (10):e107103.

Verma, M. 2012. Epigenome-wide association studies (EWAS) in cancer. *Current genomics* 13 (4):308.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang. 2012. Five years of GWAS discovery. *The American Journal of Human Genetics* 90 (1):7-24.

Vogt, I., T. Wöhner, K. Richter, H. Flachowsky, G. W. Sundin, A. Wensing, E. A. Savory, K. Geider, B. Day, M. V. Hanke, and A. Peil. 2013. Gene-for-gene relationship in the host–pathogen system *Malus × robusta* 5 – *Erwinia amylovora*. *New Phytologist* 197 (4):1262-1275.

Volk, G. M., C. T. Chao, J. Norelli, S. K. Brown, G. Fazio, C. Peace, J. McFerson, G. Y. Zhong, and P. Bretting. 2014. The vulnerability of US apple (*Malus*) genetic resources. *Genetic Resources and Crop Evolution*:1-30.

Walsh, B. 2008. Using molecular markers for detecting domestication, improvement, and adaptation genes. *Euphytica* 161 (1-2):1-17.

Wang, H., T. Nussbaum-Wagler, B. Li, Q. Zhao, Y. Vigouroux, M. Faller, K. Bomblies, L. Lukens, and J. F. Doebley. 2005. The origin of the naked grains of maize. *Nature* 436 (7051):714-719.

Wang, M., J. Yan, J. Zhao, W. Song, X. Zhang, Y. Xiao, and Y. Zheng. 2012. Genome-wide association study (GWAS) of resistance to head smut in maize. *Plant Science* 196 (0):125-131.

Wang, R. L., A. Stec, J. Hey, L. Lukens, and J. F. Doebley. 1999. The limits of selection during maize domestication. *Nature* 398 (6724):236-239.

Wang, S., D. Wong, K. Forrest, A. Allen, S. Chao, B. E. Huang, M. Maccaferri, S. Salvi, S. G. Milner, and L. Cattivelli. 2014. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant biotechnology journal* 12 (6):787-796.

Wang, Z., Q. Chen, Y. Yang, H. Yang, P. He, Z. Zhang, Z. Chen, R. Liao, Y. Tu, and X. Zhang. 2014. A genome-wide scan for selection signatures in Yorkshire and Landrace pigs based on sequencing data. *Animal genetics* 45 (6):808-816.

Whiting, G. C. and R. A. Coggins. 1975. Estimation of the monomeric phenolics of ciders. *Journal of the Science of Food and Agriculture* 26 (12):1833-1838.

Whitt, S. R., L. M. Wilson, M. I. Tenaillon, B. S. Gaut, and E. S. Buckler. 2002. Genetic diversity and selection in the maize starch pathway. *Proceedings of the National Academy of Sciences* 99 (20):12959-12962.

Wilcox, W. F. 1994. Fire blight. *Department of Plant Pathology, NYS Agricultural Experiment Station at Geneva, Cornell University*.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley, M. D. McMullen, and B. S. Gaut. 2005. The effects of artificial selection on the maize genome. *Science* 308 (5726):1310-1314.

Wright, S. I. and B. S. Gaut. 2005. Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution* 22 (3):506-519.

Xu, X. M. and J. Robinson. 2005. Modelling the effects of wetness duration and fruit maturity on infection of apple fruits of Cox's Orange Pippin and two clones of Gala by *Venturia inaequalis*. *Plant Pathology* 54 (3):347-356.

Yamasaki, M., M. I. Tenaillon, I. V. Bi, S. G. Schroeder, H. Sanchez-Villeda, J. F. Doebley, B. S. Gaut, and M. D. McMullen. 2005. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *The Plant Cell* 17 (11):2859-2872.

Yan, J., T. Shah, M. L. Warburton, E. S. Buckler, M. D. McMullen, and J. Crouch. 2009. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One* 4 (12):e8451.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, and G. W. Montgomery. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42 (7):565-569.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* 88 (1):76-82.

Yoshioka, Y. and N. Fukino. 2010. Image-based phenotyping: use of colour signature in evaluation of melon fruit colour. *Euphytica* 171 (3):409-416.

Ytournel, F. 2008. Déséquilibre de liaison et cartographie de QTL en population sélectionnée, AgroParisTech.

Yu, J. and E. S. Buckler. 2006. Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology* 17 (2):155-160.

Yuan, J. H., A. Cornille, T. Giraud, F. Y. Cheng, and Y. H. Hu. 2014. Independent domestications of cultivated tree peonies from different wild peony species. *Molecular Ecology* 23 (1):82-95.

Zhang, L., A. S. Peek, D. Dunams, and B. S. Gaut. 2002. Population genetics of duplicated disease-defense genes, *hm1* and *hm2*, in maize (*Zea mays* ssp. *mays* L.) and its wild ancestor (*Zea mays* ssp. *parviglumis*). *Genetics* 162 (2):851-860.

Zhou, X., P. Carbonetto, and M. Stephens. 2013. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics* 9 (2):e1003264.

Zhou, X. and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44 (7):821-824.

———. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods* 11 (4):407-409.

Zini, E., F. Biasioli, F. Gasperi, D. Mott, E. Aprea, T. D. Märk, A. Patocchi, C. Gessler, and M. Komjanc. 2005. QTL mapping of volatile compounds in ripe apples detected by proton transfer reaction-mass spectrometry. *Euphytica* 145 (3):269-279.

Zuk, O., E. Hechter, S. R. Sunyaev, and E. S. Lander. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* 109 (4):1193-1198.

Annexes

Annexe 1 : Liste des gènes dans lesquels des fragments d'environ 500 pb ont été reséquencés dans la CC278 et nombre de SNPs par fragment

Nom du gène	Longueur fragment	Début	LG	SNPs
MDP0000133269	500	24754887	1	26
MDP0000166405_A	499	24455726	1	26
MDP0000166405_B	500	24456538	1	21
MDP0000179513_A	498	23802993	1	36
MDP0000179513_B	498	23805594	1	43
MDP0000232836	487	24003554	1	37
MDP0000238940_A	491	24168548	1	3
MDP0000238940_B	490	24169392	1	31
MDP0000321036	500	24025747	1	4
MDP0000322294	477	23743007	1	24
MDP0000662922_A	476	24334817	1	73
MDP0000662922_B	494	24335921	1	43
MDP0000855736_A	500	24267395	1	17
MDP0000855736_B	481	24268238	1	12
MDP0000930114_A	496	24202361	1	44
MDP0000930114_B	476	24203952	1	22
MDP0000148339	499	2843051	3	41
MDP0000160327_A	497	3320373	3	15
MDP0000160327_B	496	3319462	3	26
MDP0000162814_A	494	2152571	3	26
MDP0000162814_B	500	2151765	3	44
MDP0000206447	482	2434162	3	10
MDP0000208326_A	498	2163824	3	4
MDP0000208326_B	493	2167848	3	38
MDP0000252859_A	483	3111832	3	23
MDP0000252859_B	471	3112880	3	11
MDP0000300483_A	478	3496511	3	6
MDP0000300483_B	489	3501649	3	11
MDP0000305238_A	484	2926642	3	73
MDP0000305238_B	484	2928968	3	34
MDP0000949487_A	500	3643639	3	5
MDP0000949487_B	485	3647684	3	45
MDP0000129923_A	486	23028178	7	9
MDP0000129923_B	496	23030220	7	8
MDP0000189315_A	496	22868871	7	49
MDP0000189315_B	487	22869665	7	36
MDP0000275088_A	479	23275151	7	15

MDP0000275088_B	479	23275999	7	16
MDP0000383616_A	500	23483188	7	11
MDP0000383616_B	496	23484017	7	6
MDP0000561265_A	500	22843663	7	25
MDP0000561265_B	484	22842859	7	8
MDP0000561267	499	22835879	7	11
MDP0000791555_A	486	23230461	7	17
MDP0000791555_B	486	23229609	7	8
MDP0000937158_A	489	23159733	7	17
MDP0000937158_B	499	23160629	7	8
MDP0000131641	494	25324065	10	43
MDP0000204278_A	478	24731393	10	49
MDP0000204278_B	498	24733145	10	6
MDP0000262892_A	495	24882025	10	3
MDP0000262892_B	492	24881185	10	10
MDP0000290079	492	25361691	10	22
MDP0000467552_A	487	24634869	10	3
MDP0000467552_B	484	24635775	10	6
MDP0000593517_A	487	25351382	10	8
MDP0000593517_B	495	25352272	10	11
MDP0000868782_A	498	24430105	10	36
MDP0000868782_B	494	24429140	10	34
MDP0000868785_A	497	24418698	10	27
MDP0000868785_B	473	24417302	10	36
MDP0000945841_A	496	24510711	10	11
MDP0000945841_B	478	24509875	10	7
MDP0000121185_A	496	3517647	11	30
MDP0000121185_B	488	3516839	11	26
MDP0000148169_A	500	5866072	11	26
MDP0000148169_B	497	5865238	11	15
MDP0000172931_A	472	4194124	11	22
MDP0000172931_B	498	4194918	11	16
MDP0000255648_A	482	6382214	11	7
MDP0000255648_B	500	6380179	11	39
MDP0000275915_A	495	5828490	11	54
MDP0000275915_B	500	5827689	11	9
MDP0000282879_A	482	3617348	11	15
MDP0000282879_B	477	3618415	11	14
MDP0000307848_A	495	3067091	11	1
MDP0000307848_B	498	3067718	11	16
MDP0000509183_A	500	5424428	11	6

MDP0000509183_B	497	5428310	11	4
MDP0000196196_A	494	7203417	17	36
MDP0000196196_B	491	7206838	17	6
MDP0000218438_A	485	6854414	17	53
MDP0000218438_B	487	6853328	17	17
MDP0000224380_A	500	7024418	17	13
MDP0000224380_B	499	7025146	17	15
MDP0000233031_A	489	6883031	17	8
MDP0000233031_B	464	6884709	17	8
MDP0000244520_A	495	6833712	17	6
MDP0000244520_B	475	6834562	17	12
MDP0000269326_A	494	7609268	17	17
MDP0000269326_B	499	7608461	17	17
MDP0000288817_A	470	6965961	17	6
MDP0000288817_B	500	6976258	17	9
MDP0000304854_A	498	7778310	17	11
MDP0000304854_B	473	7780205	17	6
MDP0000547261	490	7281471	17	68

Annexe 2 : Script perl utilisé pour extraire les haplotypes de SNPs des fichiers d'alignement en sortie de CLC Genomics

```
#!/usr/bin/perl
#use strict;

$pos_file = 'SNP_positions.txt'; #nom du fichier avec les positions de SNP
séparés par des tab
                                # dans la premiere colonne noms des marqueurs
                                # dans la deuxieme colonne longueur de la
reference
                                # dans la troisieme colonne nb de SNP
                                # dans les colonnes suivantes les positions des
SNP

@SNP_matrix = &matrix_read_file($pos_file);
#print "$SNP[8][0]\n";  #$SNP[ligne][colonne]
                        #colonne 0: noms marqueurs
                        #colonne 1: longueur de la séquence de reference
                        #colonne 2: nombre de SNP
$nb_marqueur = @SNP_matrix;
#print "$nb_marqueur\n";
$l=0; #ligne  les differents marqueurs

for ($l=0; $l<$nb_marqueur; $l+=1){    #$l pour les differents marqueurs
        $nommarqueur='';
        $ref_length='';
        $nommarqueur= $SNP_matrix[$l][0];
        $ref_length=$SNP_matrix[$l][1];
        $nb_haplo_marqueur= 0;
        #creation d'un tableau avec les positions des SNP pour le marqueur $l
        my @SNP='';
        my $nb_SNP='';
        my $SNPlist='';
        $nb_SNP=$SNP_matrix[$l][2];
        my $i=0;
        for ($i = 3; $i<=$nb_SNP+3; $i +=1){
                push (@SNP,$SNP_matrix[$l][$i]);
        }
        shift @SNP;
        #print "@SNP\n";

        #creation de la liste des positions des SNP pour le marqueur $l
        my $p=0;
        for ($p=0; $p<$nb_SNP; $p+=1){
                $SNPlist.="$SNP[$p]\t";
        }

        #pour le marqueur $l ouverture des fichiers .sam des differents geno
        my @sam_files='';
        @sam_files = <*.sam>;
        $nb_geno = @sam_files;
        #print "$nb_geno\n";
        my $j=0;
```

```
        for ($j = 0; $j<$nb_geno; $j +=1){       #$j pour les different geno
                $sam_file = "$sam_files[$j]";
                my $geno='';
                $geno=$sam_file;
                $geno =~ s/.sam//;
                #print "$nom_marqueur $geno\n";
                #print ">>alignout_$nommarqueur$geno.txt";
                $align_out = ">>alignout_$nommarqueur$geno.txt"; #nom du
fichier de sortie avec alignement des reads
                $align_out2 = "alignout_$nommarqueur$geno.txt"; #nom du fichier
de sortie avec alignement des reads
                $align = "alignout_$nommarqueur$geno.merge.txt"; #nouveau nom
pour pouvoir le lire par la suite TOTO
                $SNP_out = ">>SNP_reads_$nommarqueur$geno.txt"; #alignement des
reads avec uniquement les SNP
                $haplo_out = ">>haplotypes_$nommarqueur$geno.txt";  #haplotype
avec leur nombre non triés
                $real_haplo = ">>real_haplo_$nommarqueur.txt"; #haplotype
frequence reads >0.1 pour tous les geno
                $real = "real_haplo_$nommarqueur.txt"; #pour pouvoir lire le
fichier
                $haplo_fin = ">>haplo_fin_$nommarqueur.txt"; #fichier formate
pour Phylip

                #pour chaque marqueur et chaque geno, lecture des .sam et
creation d'un fichier avec alignement des reads
                open(OUT, $align_out) or die "Couldn't open \n";
                open(SAM,$sam_file) or die "The file $sam_file could" .
                " not be found.\n";
                my @line = '';
                while (<SAM>){
                        next if ($_ =~ /^\@/);
                        next if ($_ =~ /^read_\d/);
                        @line = split (/\t/,$_);
                        $marqueur = $line[2];
                        $start = $line[3];
                        $cigar = $line[5];
                        $qstr = $line[9];
                        if ($marqueur eq $nommarqueur){
                                $qaln = &parse_cigar($qstr, $cigar);
                                print OUT "$qaln\n";
                        }
                }

                close SAM;
                close OUT;

#my ($filename) = "alignout_MDP0000121185_AA008DNZA5.txt";
my $filename = "alignout_$nommarqueur$geno.txt";
merge_file($filename,0);

#sub-routine pour fusionner les lignes
sub merge_file {
        my ($filename, $debug) = @_;

        my $i = 0;
```

```
        my $forward = false;
        my $reverse = false;
        my $forward_str = "";
        my $forward_line = -1;
        my $forward_pos = -1;
        my $forward_end_pos = -1;
        my $reverse_str = "";
        my $reverse_line = -1;
        my $reverse_pos = -1;
        my $reverse_end_pos = -1;
        my $where = -1;
        my $where_end = -1;
        my $line_nb = 0;

        open (F, $align_out2) || die "Could not open $align_out2: $!";
        @file = split(/\./,$align_out2);
        $file_merge = ">" . $file[0] . ".merge.txt";
        open (G, $file_merge) || die "Could not open $file_merge: $!";
        if ($debug == 1) {
                print G "BEGIN DEBUG ====================================\n";
        }
        while ($line = <F>) {
                chomp($line);
                next if $line =~ /^\@/; # skip blank lines
                $len = length($line);
                $line_nb += 1;

                if ($debug == 1) {
                        print G "from line $line_nb len $len\n$line\n";
                }

                @splitted = split(/N/,$line);
                $where = -1;
                $where_end = -1;
                foreach $item (@splitted) {
                        if ( $item ne "" ) {
                                if ($debug == 1) {
                                        print G "extract search\n$item\n";
                                }
                                if ( $where == -1 ) {
                                        $where = index($line,$item);
                                        if ( $where != -1 ) {
                                                $where_end =
index($line,"N",$where);
                                        }
                                } else {
                                        $find_end = index($line,$item);
                                        if ( $find_end > $where_end ) {
                                                if ( $find_end != -1 ) {
                                                        $where_end =
index($line,"N",$find_end);
                                                }
                                        } else {
                                                $find_end = rindex($line,$item);
                                                if ( $find_end > $where_end ) {
                                                        if ( $find_end != -1 ) {
```

```
                                                                        $where_end =
index($line,"N",$find_end);
                                                                }
                                                        }
                                                }
                                        }
                                }
                        }
                        if ( $where != -1 ) {
                                if ( $where < $len/2 && ($where_end != -1 && $where_end
< 1.1*$len/2)) {
                                        if ($debug == 1) {
                                                print G "forward line found (begin
$where end $where_end)\n";
                                        }
                                        if ( $forward_str ne "" ) {
                                                if ($debug == 1) {
                                                        print G "already found forward
line $forward_line, replace with line $line_nb\n";
                                                } else {
                                                        print "already found forward
line $forward_line, replace with line $line_nb\n";
                                                }
                                        }
                                        $forward_str = $line;
                                        $forward_pos = $where-1;
                                        $forward_line = $line_nb;
                                        if ( $forward_pos < 0 ) {
                                                $forward_pos = 0;
                                        }
                                        $forward_end_pos = $where_end;
                                } else {
                                        if ( $forward_str eq "" ) {
                                                if ($debug == 1) {
                                                        print G "no forward, discard
reverse line $line_nb\n";
                                                } else {
                                                        print "no forward, discard
reverse line $line_nb\n";
                                                }
                                        } else {
                                                if ($debug == 1) {
                                                        print G "reverse line found
(begin $where end $where_end)\n";
                                                }
                                                $reverse_str = $line;
                                                $reverse_pos = $where-1;
                                                $reverse_end_pos = $where_end;
                                                $reverse_line = $line_nb;
                                                if ( $reverse_end_pos == -1 ) {
                                                        $reverse_end_pos = $len-1;
                                                }
                                        }
                                }
                        }
```

```
                if ( $forward_str ne "" && $reverse_str ne "" )
                {
                        $merge_line = substr($forward_str,0,$forward_end_pos) .
substr($reverse_str,$forward_end_pos);
                        if ($debug == 1) {
                                print G "merge lines\n";
                                print G "forward line
$forward_line:\n$forward_str\nreverse line $reverse_line:\n$reverse_str\npos
merge $forward_end_pos\n";
                        }
                        my $Ncpt = 0;
                        for (my $i = 0 ; $i < length($merge_line) ; $i++) {
                                if ($merge_line[$i] eq 'N') {
                                        $Ncpt ++;
                                }
                        }
                        if ($Ncpt < 50) {
                                print G "$merge_line\n";
                                #print "Ligne OK\n";
                        } else {
                                print "Discard line N >= 50 [" . $Ncpt . "]\n";
                        }
                        $forward_str = "";
                        $reverse_str = "";
                }

        }

        if ($debug == 1) {
                print G "EOF ========================================EOF\n";
        }

        close(F);
        close(G);
}

                #pour chaque marqueur et chaque geno, lecture des fichiers
alignements et creation des fichiers avec uniquement les SNP
                @tabSNP = '';
                $readnb = 0;
                open(TOTO,$align) or die "The file $align could" .
                " not be found.\n";
                while (<TOTO>){
                        my $line=$_;
                        my $k = 0;
                        for ($k = 0; $k<$nb_SNP; $k +=1) {
                                my $temp = substr $line,($SNP[$k]-1),1;
                                $tabSNP[$readnb] .= "$temp,";
                        }
                        $readnb = $readnb + 1;
                }
                close TOTO;
                @tabSNP = sort(@tabSNP);

                open(SNPOUT,$SNP_out) or die "The file $SNP_out could" .
                " not be found.\n";
```

```
                my $m=0;
                for ($m = 0; $m<=$readnb; $m +=1) {
                        print SNPOUT "$tabSNP[$m]\n";
                }
                close SNPOUT;

                #pour chaque marqueur et chaque geno, comptage des haplotypes
                my @haplo = '';
                my @count = '';
                my $cc = 1;
                my $tempo = '';
                my $n ='';
                for ($n =1; $n<=$readnb; $n +=1) {
                        $tempo = $tabSNP[$n-1];
                        if ($tempo eq $tabSNP[$n]){
                                $cc +=1;
                        }
                        else {
                        push (@haplo,$tempo);
                        push (@count,$cc);
                        $cc = 1;
                        }
                }
                push (@haplo,$tempo);
                push (@count,$cc);
                shift @haplo;
                shift @count;
                $nbhaplo=0;
                $nbhaplo = @haplo;
                #print "nb haplo $nommarqueur$geno: $nbhaplo\n";

                #pour chaque marqueur et chaque geno, ecriture des haplotypes
et leur nombre
                my $o=0;
                @haplotype = '';
                for ($o =0; $o<$nbhaplo; $o +=1) {
                        $haplotype[$o] = "$count[$o]\t$haplo[$o]\n";
                }

                open(HAPLO,$haplo_out) or die "The file $haplo_out could" .
                " not be found.\n";
                print HAPLO "$nommarqueur $geno\n";
                print HAPLO "SNP positions: $SNPlist\n";
                print HAPLO "Number of haplotypes: $nbhaplo\n";
                print HAPLO @haplotype;
                close HAPLO;

                #obtention des haplotypes dont la frequence dans les 6 reads
les plus fréquents > 0.1
                #création d'un hachage pour faire le lien entre haplotype (clé)
et
                # le nombre de fois qu'il a été vu (valeur)
                %hash=();
                my $r=0;
                        for ($r=0; $r<$nbhaplo; $r +=1){
                        $hash{$haplo[$r]} = $count[$r];
```

```
                        }
                @sortcount='';
                @sortcount = sort {$b <=> $a }@count;
                $sum='';
                $sum = $sortcount[0]+ $sortcount[1] + $sortcount[2] +
$sortcount[3]
                + $sortcount[4] + $sortcount[5];

                @real_haplo = '';
                my $s=0;
                for ($s=0; $s<=5; $s +=1){
                        if ($sortcount[$s]/$sum > 0.1){
                                push (@real_haplo, $sortcount[$s]);
                                $nb_haplo_marqueur +=1;
                        }
                        else {
                                next;
                        }
                }
                shift @real_haplo;

                $nb_real_hapl = 0;
                $nb_real_hapl = @real_haplo;

                @real_haplotype = '';
                my $t=0;
                for ($t=0; $t<$nb_real_hapl; $t +=1){
                        while (($cle, $valeur) = each %hash){
                                if ($valeur eq $real_haplo[$t]){
                                        push (@real_haplotype, "$cle");
                                }
                        }
                }
                shift @real_haplotype;
                open(REAL_HAPLO,$real_haplo) or die "The file $real_haplo
could" .
                " not be found.\n";
                my$u=0;
                for ($u=0; $u<$nb_real_hapl; $u +=1){
                        my $tempor='';
                        $tempor = $real_haplotype[$u];
                        $tempor =~ s/,//g;
                        print REAL_HAPLO "$geno$u$tempor\n";
                }
                close REAL_HAPLO;
                }
        open(HAPLOFIN,$haplo_fin) or die "The file $haplo_fin could" .
        " not be found.\n";
        print HAPLOFIN  "$nb_haplo_marqueur          $nb_SNP\n";
        open(TATA,$real) or die "The file $real could" .
        " not be found.\n";
        while (<TATA>){
                        my $lin=$_;
                        print HAPLOFIN "$lin";
        }
        close TATA;
```

```
        close HAPLOFIN;

}

print "The job is done!";

##############################################################################
#sub-routine pour mettre un tableau dans une matrice
sub matrix_read_file {
    my ($filename) = @_;
    my @matrix ='';
    open (F, $filename) || die "Could not open $filename: $!";
    while ($line = <F>) {
        chomp($line);
        next if $line =~ /^\@/; # skip blank lines
        my (@row) = split (/\s+/, $line);
        push (@matrix, \@row); # insert the row-array into
                                            # the outer matrix array
    }
    shift @matrix;
    close(F);
    return @matrix;
}

#sub-routine pour interpreter le cigar
sub parse_cigar {
        my ($q, $cigar) = @_;
        my $c = $cigar;
        my $out = "N"x($start-1);
        while ($c){
                if ($c =~ m{^([0-9]+)([MX+])(.*)$}){
                        $out .= substr $q, 0, $1;
                        $q = substr $q, $1, length($q)-$1;
                        $c = $3;
                    }
                if ($c =~ m{^([0-9]+)([HS])(.*)$}){
                        #$out .= substr $q, 0, $1;
                        $q = substr $q, $1, length($q)-$1;
                        $c = $3;
                    }
                if ($c =~ m{^([0-9]+)([I])(.*)$}){
                        #$out .= "";
                        $q = substr $q, $1, length($q)-$1;
                        $c = $3;
                    }
                 if ($c =~ m{^([0-9]+)([DN])(.*)$}){
                        $out .= '-'x$1;
                        #$q = substr $q, $1, length($q)-$1;
                        $c = $3;
                    }
         }
         $lout = length $out;
         $out .= 'N'x($ref_length-$lout);
         return $out;
}
```

Annexe 3 : Names of the 96 cultivars chosen from the INRA Angers collections of old cider and dessert apple varieties

Name of the variety	Type
AMADOU	Dessert
API	Dessert
BARBE	Dessert
BEACON	Dessert
BEAUTY OF BATH	Dessert
BELLE DE MAGNI	Dessert
BELLE FILLE DE L'INDRE	Dessert
BELLE FILLE DE ST FEYRE	Dessert
BELLE FLEUR KRASNYI	Dessert
BORDES	Dessert
BOROWITSKY	Dessert
CACHAO SAGARRA	Dessert
CALVILLE DE MLEIEV	Dessert
CALVILLE DU ROI	Dessert
COLATE	Dessert
DE BONDE	Dessert
DE NOEL	Dessert
DIRECTEUR LESAGE	Dessert
ELLISON'S ORANGE	Dessert
ENTZEA SAGARRA	Dessert
FEVRETTE	Dessert
FLEURITARD ROUGE	Dessert
GELADE	Dessert
GEWURTZLUIKEN	Dessert
GIREUSE	Dessert
GRAIN D'OR DES CHARENTES	Dessert
GRENADIER	Dessert
GROSSE SAINT CLEMENT	Dessert
MERVEILLE DE VITRY	Dessert
PETIT API	Dessert
PETIT MUSEAU DE LIEVRE	Dessert
PETITE MADELEINE	Dessert
PIGEON DE JERUSALEM	Dessert
POMME BLANCHE D'ETE	Dessert
POMME D'ETE ROUGEUR DE PECHE	Dessert
POMME FER	Dessert
POMME ORANGE	Dessert
POMME ST JACQUES	Dessert
REINETTE DE PLUVIGNE	Dessert
REINETTE ETOILEE	Dessert

REINETTE SANGUINE DU RHIN	Dessert
ROMARIN	Dessert
ROUMENTINE	Dessert
SAINT MICHEL	Dessert
TAFFETAS BLANC	Dessert
TETON DE DEMOISELLE	Dessert
TRANSPARENTE BLANCHE	Dessert
TRELAGE	Dessert
ABONDANCE	Cider
AMERE NOUVELLE	Cider
AMERE ST JACQUES	Cider
AVROLLES GROSSE	Cider
BINET BLANC	Cider
BLANC JAUNET	Cider
BLANC MOLLET	Cider
BLANCHET	Cider
CARISI	Cider
CHUERO RU	Cider
CROLLON	Cider
CUL D'OISAN	Cider
DAMELOT	Cider
DOUCE COET LIGNE	Cider
DOUCE ROUSSE	Cider
DOUX CORIER	Cider
DOUX LOZON	Cider
Dx V CARROUGES	Cider
GALOPIN	Cider
GENERAL	Cider
GRISE DIEPPOIS	Cider
GUILLEVIC	Cider
HERBAGE SEC	Cider
MANERBE	Cider
MARECHAL	Cider
MAUGERE	Cider
MERISIER	Cider
MONTE EN HAUT	Cider
MOULIN A VENT EURE	Cider
MUSCADET DE DIEPPE	Cider
NOE BINAY	Cider
P.G.R. JANZE	Cider
PERICO	Cider
PETIT DOUX DE BRETAGNE	Cider
PETIT FREQUIN ROUGE	Cider
PETIT JAUNE	Cider

PETIT MARIN ONFROY	Cider
PETITE SORTE	Cider
PIERRE ALLAIRE	Cider
POMME DE MOET	Cider
PORTIER	Cider
RADOR	Cider
REINETTE MARBREE DE LUZOIR	Cider
RENAO	Cider
RENAO PETIT	Cider
RENE MARTIN	Cider
ROUGET DE DOL	Cider
SAINT BAZYL	Cider

Annexe 4 : Results from the Blast2GO software on the genes identified around the significant SNPs

Trait	Seq. Name	Seq. Description	#Hits	min. eValue	MS[a]	GOs[b]	Enzyme Codes
FST	MDP0000076511	endoplasmic reticulum-golgi intermediate compartment protein 3-like	2.00E+01	0.00%	91.30%	C:endoplasmic reticulum	-
	MDP0000084203	vacuolar-processing enzyme-like	20	0.00%	94.35%	P:biological_process; F:peptidase activity	EC:3.4.22; EC:3.4
	MDP0000126636	probable protein phosphatase 2c 72	20	0.00%	80.20%	P:biological_process; F:molecular_function	-
	MDP0000131883	---NA---	0			-	
	MDP0000133911	udp-glycosyltransferase 85a2-like	20	0.00%	84.85%	F:transferase activity, transferring hexosyl groups; P:metabolic process	-
	MDP0000136026	nedd8 ultimate buster 1	2.00E+01	1.18E-180	92.10%	F:protein binding	-
	MDP0000139577	PREDICTED: uncharacterized protein LOC103426009	20	5.75E-162	77.05%	-	
	MDP0000149647	nodulation receptor kinase-like	2.00E+01	6.48E-92	94.25%	P:signal transduction; F:ion binding; P:cellular protein modification process; P:cellular amino acid metabolic process; C:cellular_component; F:signal transducer activity; F:kinase activity	EC:2.7.11.25; EC:2.7.11
	MDP0000150721	udp-glycosyltransferase 85a2-like	20	0.00%	85.10%	P:metabolic process; F:transferase activity, transferring hexosyl groups	-
	MDP0000150726	e3 ubiquitin-protein ligase march10	2.00E+01	0.00%	75.15%	F:ion binding	-
	MDP0000150727	vacuolar protein sorting-associated protein 53 homolog	20	0.00%	93.60%	-	
	MDP0000150729	chlorophyll a-b binding protein chloroplastic	2.00E+01	0.00%	93.00%	P:generation of precursor metabolites and energy; P:cellular protein modification process; P:photosynthesis; C:protein complex; C:cellular_component; C:plastid; F:molecular_function; C:thylakoid	-
	MDP0000159850	clathrin interactor epsin 1-like	20	0.00%	84.85%	C:intracellular; F:molecular_function	-
	MDP0000168905	nodulation receptor kinase-like	2.00E+01	3.28E-55	56.65%	F:molecular_function	EC:2.7.11
	MDP0000172806	gibberellin 3-beta-dioxygenase 1-like	17	8.32E-08	69.94%	P:oxidation-reduction process; F:oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors; F:oxidoreductase activity; F:gibberellin 3-beta-dioxygenase activity; F:iron ion binding; P:response to gibberellin; P:response to red light; P:response to red or far red light; F:transcription factor binding	
	MDP0000173095	mtd1 family protein	2.00E+01	6.72E-144	73.35%	-	
	MDP0000175487	cmp-sialic acid transporter 5-like	20	3.18E-143	93.20%	P:transmembrane transport; F:transmembrane transporter activity; C:cytoplasm; P:anatomical structure formation involved in morphogenesis; C:Golgi apparatus; C:cellular_component; P:transport	-
	MDP0000179065	oxidation resistance protein 1-like	20	0.00%	81.10%	-	
	MDP0000181379	tubulin-folding cofactor c-like	2.00E+01	5.99E-19	70.05%	P:biological_process	-
	MDP0000191944	calcium uptake protein mitochondrial-like	2.00E+01	4.32E-41	54.60%	F:calcium ion binding	-

MDP0000191948	protein unc-45-a-like protein	2.00E+01	4.84E-85	91.80%	F:protein binding	-
MDP0000192297	myb family transcription factor-related protein	2.00E+01	1.89E-107	58.85%	C:nucleus; F:DNA binding; P:sulfur compound metabolic process; F:nucleic acid binding transcription factor activity; P:anatomical structure development; P:cellular amino acid metabolic process; P:cellular nitrogen compound metabolic process; P:biosynthetic process; F:molecular_function; P:response to stress	-
MDP0000197740	f-box kelch-repeat protein at3g06240-like	20	1.07E-37	46.45%	-	
MDP0000200785	ras-related protein rabh1e	2.00E+01	9.63E-149	98.15%	P:biological_process; P:signal transduction; F:ion binding; P:transport; F:molecular_function	EC:3.6.1; EC:3.6.1.15
MDP0000201472	l-lactate dehydrogenase a	20	1.77E-12	80.65%	P:small molecule metabolic process; P:carbohydrate metabolic process; P:biological_process; P:generation of precursor metabolites and energy; C:cytoplasm; F:oxidoreductase activity; P:catabolic process	EC:1.1.1.27
MDP0000201500	transmembrane 9 superfamily member 4	20	0.00%	91.55%	C:endosome; C:cytoplasm; C:Golgi apparatus; C:vacuole; C:cellular_component	-
MDP0000208497	plant t5j17-70	2.00E+01	2.41E-108	75.95%	-	
MDP0000213515	btb poz domain-containing protein dot3	2.00E+01	0.00%	83.75%	P:biological_process; P:anatomical structure development	-
MDP0000221796	cmp-sialic acid transporter 5-like	2.00E+01	4.47E-140	93.65%	P:transmembrane transport; F:transmembrane transporter activity; C:cytoplasm; P:anatomical structure formation involved in morphogenesis; C:Golgi apparatus; C:cellular_component; P:transport	-
MDP0000222448	---NA---	0			-	
MDP0000222909	asparagine synthetase	20	4.25E-76	92.60%	F:ion binding; P:cellular amino acid metabolic process; P:cellular nitrogen compound metabolic process; P:biosynthetic process; F:ligase activity	EC:6.3.5.4
MDP0000225641	glycosyl hydrolase family 10 protein carbohydrate-binding domain-containing protein isoform 1	20	0.00%	87.15%	P:carbohydrate metabolic process; P:catabolic process; P:cell wall organization or biogenesis; F:hydrolase activity, acting on glycosyl bonds	-
MDP0000232773	portal 56	20	0.00%	79.25%	-	
MDP0000232774	flowering time control protein fpa	20	0.00%	52.10%	F:nucleic acid binding; F:nucleotide binding	
MDP0000235629	nodulation receptor kinase-like	20	0.00%	71.00%	P:biological_process; P:cellular amino acid metabolic process; C:cellular_component; F:kinase activity; F:molecular_function	EC:2.7.11
MDP0000240075	nodulation receptor kinase-like	2.00E+01	1.60E-73	94.30%	P:signal transduction; F:ion binding; P:cellular protein modification process; P:cellular amino acid metabolic process; C:cellular_component; F:signal transducer activity; F:kinase activity	EC:2.7.11.25; EC:2.7.11
MDP0000241162	vacuolar-processing enzyme-like	20	0.00%	86.75%	P:biological_process; F:peptidase activity	EC:3.4.22; EC:3.4
MDP0000241165	PREDICTED: uncharacterized protein LOC103440997	20	2.45E-180	88.25%	C:nucleus; C:protein complex; P:DNA metabolic process; P:response to stress	-
MDP0000241166	serine threonine-protein kinase at5g01020	2.00E+01	1.35E-96	82.60%	F:ion binding; P:cellular protein modification process; P:cellular amino acid metabolic process; F:kinase activity	EC:2.7.10; EC:2.7.11; EC:2.7.10.2

MDP0000243737	cyclin-dependent kinase c-2-like	20	0.00%	77.75%	P:biological_process; F:kinase activity; F:molecular_function	EC:2.7.11
MDP0000243738	probable mitochondrial adenine nucleotide transporter btl3	20	0.00%	83.95%	P:transmembrane transport; C:cellular_component; C:plastid	-
MDP0000244376	ras-related protein rabh1e-like	2.00E+01	1.71E-144	98.05%	P:biological_process; P:signal transduction; F:ion binding; P:transport; F:molecular_function	EC:3.6.1; EC:3.6.1.15
MDP0000256619	glucose-1-phosphate adenylyltransferase small chloroplastic amyloplastic	2.00E+01	0.00%	95.55%	P:carbohydrate metabolic process; F:nucleotidyltransferase activity; P:generation of precursor metabolites and energy; C:cytoplasm; P:anatomical structure development; C:protein complex; P:biosynthetic process; C:extracellular region; C:plastid	EC:2.7.7.27
MDP0000258981	nodulation receptor kinase-like	2.00E+01	1.99E-70	69.90%	P:biological_process; F:kinase activity	-
MDP0000266156	transcription factor rax2-like	2.00E+01	2.64E-180	67.15%	F:chromatin binding; F:DNA binding; C:chromatin	-
MDP0000268708	exonuclease 3 -5 domain-containing protein 1-like	2.00E+01	1.73E-121	90.00%	F:RNA binding; F:nuclease activity; P:cellular nitrogen compound metabolic process	EC:3.1
MDP0000275432	glucan endo- -beta- basic isoform-like	7	1.03E-52	66.00%	F:molecular_function	-
MDP0000280307	homeobox protein knotted-1-like 2	20	0.00%	87.80%	C:nucleus; F:DNA binding; F:nucleic acid binding transcription factor activity; P:cellular nitrogen compound metabolic process; P:biosynthetic process; C:protein complex; C:intracellular	-
MDP0000283141	phosphoglucan phosphatase amyloplastic	2.00E+01	7.17E-156	72.30%	P:carbohydrate metabolic process; P:biological_process; F:phosphatase activity; C:plastid	EC:3.1.3.16; EC:3.1; EC:3.1.3.41; EC:3.1.3
MDP0000290409	calmodulin-binding transcription activator 3	2.00E+01	0.00%	80.00%	C:nucleus; F:DNA binding	-
MDP0000290414	nedd8 ultimate buster 1	20	0.00%	84.70%	F:protein binding	-
MDP0000290415	protein os-9-like	20	0.00%	88.05%	P:catabolic process; C:endoplasmic reticulum; P:response to stress	-
MDP0000299712	03-mai	20	0	81.00%	F:RNA binding; F:nuclease activity; P:cellular nitrogen compound metabolic process	EC:3.1
MDP0000306669	cmp-sialic acid transporter 5-like	20	7.63E-178	94.05%	P:transmembrane transport; F:transmembrane transporter activity; C:cytoplasm; P:anatomical structure formation involved in morphogenesis; C:Golgi apparatus; C:cellular_component; P:transport	-
MDP0000306670	asparagine synthetase	2.00E+01	6.28E-80	86.60%	P:biological_process; P:sulfur compound metabolic process; F:ion binding; C:cytosol; P:cellular amino acid metabolic process; P:biosynthetic process; P:cellular nitrogen compound metabolic process; C:cellular_component; F:ligase activity; P:response to stress	EC:6.3.5.4; EC:6.3.1; EC:6.3.1.1
MDP0000320591	PREDICTED: uncharacterized protein LOC103952713 isoform X1	20	4.19E-27	71.20%	-	
MDP0000324950	---NA---	0			-	
MDP0000332596	serine mitochondrial	20	0.00%	95.20%	P:biological_process; F:ion binding; P:cellular amino acid metabolic process; F:methyltransferase activity; P:biosynthetic	EC:2.1.2.1

					process; C:ribosome; C:extracellular region; C:plasma membrane; C:plastid; F:molecular_function; P:response to stress; C:thylakoid; C:nucleus; F:RNA binding; C:mitochondrion; P:cofactor metabolic process; C:cytosol		
MDP0000332597	serine mitochondrial	20	0.00%	95.20%	P:biological_process; F:ion binding; P:cellular amino acid metabolic process; F:methyltransferase activity; P:biosynthetic process; C:ribosome; C:extracellular region; C:plasma membrane; C:plastid; F:molecular_function; P:response to stress; C:thylakoid; C:nucleus; F:RNA binding; C:mitochondrion; P:cofactor metabolic process; C:cytosol	EC:2.1.2.1	
MDP0000336114	---NA---	0			-		
MDP0000340361	---NA---	0			-		
MDP0000369409	---NA---	0			-		
MDP0000392459	---NA---	0			-		
MDP0000404395	transmembrane 9 superfamily member 4	20	0.00%	91.35%	C:endosome; C:cytoplasm; C:Golgi apparatus; C:vacuole; C:cellular_component	-	
MDP0000404396	myb family transcription factor apl-like isoform x2	4.00E+00	1.58E-39	60.75%	F:DNA binding		
MDP0000500222	f-box kelch-repeat protein at3g06240-like	20	0.00%	65.80%	F:protein binding	-	
MDP0000503661	---NA---	0			-		
MDP0000529897	asparagine synthetase	20	7.60E-60	85.70%	F:ion binding; C:cytosol; P:cellular amino acid metabolic process; P:cellular nitrogen compound metabolic process; P:biosynthetic process; F:ligase activity	EC:6.3.5.4; EC:6.3.1.1	EC:6.3.1;
MDP0000543680	formin-like protein 14	2.00E+01	2.61E-80	75.70%	-		
MDP0000579882	homeobox-leucine zipper protein hat22-like	20	0.00%	71.40%	C:nucleus; F:DNA binding; F:nucleic acid binding transcription factor activity; P:cellular nitrogen compound metabolic process; P:biosynthetic process; C:protein complex; C:intracellular	-	
MDP0000607885	ras-related protein rabh1e-like	2.00E+01	4.22E-35	96.65%	P:signal transduction; C:mitochondrion; F:ion binding; P:transport	-	
MDP0000613004	protein ralf-like 24	2.00E+01	5.22E-98	78.05%	-		
MDP0000613011	heat stress transcription factor a-5-like	20	0.00%	84.45%	C:nucleus; F:DNA binding; F:nucleic acid binding transcription factor activity; P:cellular nitrogen compound metabolic process; P:biosynthetic process; P:response to stress	-	
MDP0000682858	(-)-isopiperitenol (-)-carveol mitochondrial-like	20	9.90E-55	79.10%	P:biological_process; F:oxidoreductase activity	-	
MDP0000704869	asparagine synthetase	2.00E+01	9.31E-145	87.40%	F:ion binding; C:cytosol; P:cellular amino acid metabolic process; P:cellular nitrogen compound metabolic process; P:biosynthetic process; F:ligase activity	EC:6.3.5.4; EC:6.3.1.1	EC:6.3.1;
MDP0000716576	vacuolar protein sorting-associated protein 32 homolog 2-like	2.00E+01	1.15E-82	93.45%	P:transport; C:plasma membrane	-	

	MDP0000721663	PREDICTED: uncharacterized protein LOC103410688 isoform X1	20	6.80E-153	80.85%	-	
	MDP0000722046	glucose-1-phosphate adenylyltransferase small chloroplastic amyloplastic	2.00E+01	0.00%	93.25%	P:carbohydrate metabolic process; F:nucleotidyltransferase activity; P:generation of precursor metabolites and energy; C:cytoplasm; P:anatomical structure development; C:protein complex; P:biosynthetic process; C:extracellular region; C:plastid	EC:2.7.7.27
	MDP0000743523	glutamate receptor -like	10	7.74E-33	86.50%	-	
	MDP0000752179	transmembrane 9 superfamily member 4	20	0.00%	90.05%	C:endosome; C:cytoplasm; C:Golgi apparatus; C:vacuole; C:cellular_component	-
	MDP0000763250	asparagine synthetase	2.00E+01	1.23E-53	83.40%	F:DNA binding; P:biological_process; P:sulfur compound metabolic process; F:ion binding; C:cytosol; P:cellular amino acid metabolic process; P:cellular nitrogen compound metabolic process; P:biosynthetic process; C:cellular_component; F:ligase activity; P:response to stress	EC:6.3.5.4; EC:6.3.1; EC:6.3.1.1
	MDP0000841002	zinc finger ccch domain-containing protein 20-like	20	0.00%	74.70%	F:ion binding	-
	MDP0000844309	transmembrane 9 superfamily member 4	20	0.00%	95.75%	C:endosome; C:cytoplasm; C:Golgi apparatus; C:vacuole; C:cellular_component	-
	MDP0000884993	glucose-1-phosphate adenylyltransferase small chloroplastic amyloplastic	20	0.00%	95.25%	P:carbohydrate metabolic process; F:nucleotidyltransferase activity; P:generation of precursor metabolites and energy; C:cytoplasm; P:anatomical structure development; C:protein complex; P:biosynthetic process; C:extracellular region; C:plastid	EC:2.7.7.27
	MDP0000938906	---NA---	0			-	
Variety Type	MDP0000122847	9-cis-epoxycarotenoid dioxygenase	2.00E+01	5.72E-34	90.70%	C:chloroplast thylakoid membrane; F:9-cis-epoxycarotenoid dioxygenase activity; P:response to water deprivation; P:seed dormancy process; P:oxidation-reduction process	EC:1.13.11; EC:1.13.11.51
	MDP0000123712	PREDICTED: uncharacterized protein LOC104900177	2000.00%	2.73E-06	49.30%	-	
	MDP0000123768	abc transporter b family member 11-like	2000.00%	0.00%	87.45%	C:integral component of membrane; F:ATP binding; F:ATPase activity, coupled to transmembrane movement of substances; P:ATP catabolic process; P:transmembrane transport	EC:3.6.1; EC:3.6.1.3; EC:3.6.1.15
	MDP0000130493	translation factor guf1 mitochondrial-like	20	6.03E-158	72.45%	-	
	MDP0000131617		1	0	100%	F:nucleotide binding; F:DNA-directed DNA polymerase activity; F:nucleoside binding; F:DNA binding; P:DNA replication initiation	EC:2.7.7.7
	MDP0000132257	protein ralf-like 34	20	2.34E-64	81.15%	-	
	MDP0000135062	PREDICTED: uncharacterized protein LOC103441514	2.00E+01	1.65E-149	57.75%	-	
	MDP0000135063	homeobox protein knotted-1-like 3 isoform x2	2.00E+01	0.00%	92.45%	C:nucleus; F:sequence-specific DNA binding transcription factor activity; F:sequence-specific DNA binding; P:regulation of	-

					transcription, DNA-templated; C:transcription factor complex; P:regulation of transcription, DNA-templated	
MDP0000135177	formin-like protein 14	20	2.89E-24	67.90%	-	
MDP0000135249	disease resistance protein rga3	2.00E+01	0.00%	80.85%	F:ADP binding; F:protein binding	-
MDP0000135386	s-adenosyl-l-methionine-dependent methyltransferases superfamily protein	20	1.77E-15	90.45%	F:methyltransferase activity; P:rRNA methylation	-
MDP0000136517	probable serine threonine-protein kinase at1g01540	20	0	84.75%	F:MAP kinase kinase kinase activity; F:ATP binding; P:activation of MAPKK activity; P:transmembrane receptor protein serine/threonine kinase signaling pathway; P:serine family amino acid metabolic process	EC:2.7.11.25; EC:2.7.11
MDP0000136519	subtilisin-like protease	20	2.98E-60	68.30%	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21
MDP0000136520	subtilisin-like protease	20	6.34E-108	81.85%	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21
MDP0000137177	cation calcium exchanger 2-like	2000.00%	0.00%	81.85%	C:integral component of membrane; P:transmembrane transport	-
MDP0000137179	protein pns1-like	20	0	79.70%	C:integral component of membrane	-
MDP0000137181	c2 domain-containing family protein	2.00E+01	0.00%	88.45%	F:protein binding	-
MDP0000140854	receptor-like protein 12	20	1.73E-112	78.50%	-	
MDP0000143114	---NA---	0.00%			-	
MDP0000143130	uncharacterized acetyltransferase at3g50280-like	20	0	86.40%	F:transferase activity, transferring acyl groups other than amino-acyl groups; P:metabolic process	-
MDP0000156152	protein far1-related sequence 5-like	20	3.07E-22	63.90%	F:organic cyclic compound binding; F:heterocyclic compound binding	-
MDP0000158507	nigrin b-like	2.00E+01	0.00%	67.15%	F:rRNA N-glycosylase activity; P:negative regulation of translation	EC:3.2.2.22
MDP0000159264	pentatricopeptide repeat-containing protein at5g66520-like	2000.00%	1.14E-31	76.45%	-	
MDP0000160601	heterogeneous nuclear ribonucleoprotein 1	2000.00%	0.00%	64.85%	F:nucleic acid binding; F:nucleotide binding	-
MDP0000161197	embryonic protein dc-8-like	2000.00%	6.16E-109	70.70%	P:embryo development ending in seed dormancy; C:cellular_component	
MDP0000164429	3-phosphoshikimate 1-carboxyvinyltransferase 2	2.00E+01	0.00%	89.35%	C:chloroplast stroma; F:3-phosphoshikimate 1-carboxyvinyltransferase activity; P:chorismate biosynthetic process; P:response to herbicide; P:tryptophan biosynthetic process; P:tyrosine biosynthetic process; P:L-phenylalanine biosynthetic process	EC:2.5.1.19
MDP0000173824	coiled-coil domain-containing protein	2.00E+01	2.16E-45	65.60%	-	
MDP0000174414	subtilisin-like protease	20	0	84.90%	F:serine-type endopeptidase activity; F:2-alkenal reductase [NAD(P)] activity; P:proteolysis; P:oxidation-reduction process	EC:3.4.21; EC:1.3.1.74
MDP0000176158	PREDICTED: uncharacterized protein LOC103423163	2000.00%	0.00%	84.35%	-	
MDP0000178796	glycerol kinase-like	20	0	94.00%	F:glycerol kinase activity; P:response to molecule of bacterial origin; P:carbohydrate metabolic process; P:glycerol-3-phosphate metabolic process; P:response to microbial	EC:2.7.1.30

					phytotoxin; P:phosphorylation; P:defense response to bacterium; P:response to karrikin; P:glycerolipid metabolic process	
MDP0000178799	nadh dehydrogenase	2000.00%	2.35E-180	95.45%	C:mitochondrial respiratory chain complex I; F:zinc ion binding; F:oxidoreductase activity; F:2 iron, 2 sulfur cluster binding; P:response to oxidative stress; P:oxidation-reduction process	-
MDP0000179991	tmv resistance protein n-like isoform x2	20	5.03E-26	61.85%	-	
MDP0000183373	potassium channel skor	20	5.88E-166	91.15%	F:outward rectifier potassium channel activity; P:potassium ion transmembrane transport; C:voltage-gated potassium channel complex; F:protein binding	-
MDP0000183375	embryonic protein dc-8-like	20	8.65E-90	66.95%	-	
MDP0000183836	embryonic protein dc-8-like	2.00E+01	0.00%	69.95%	P:embryo development ending in seed dormancy; C:cellular_component	
MDP0000184339	uncharacterized acetyltransferase at3g50280-like	20	7.46E-105	87.95%	C:cytosol; F:transferase activity, transferring acyl groups other than amino-acyl groups; P:metabolic process	-
MDP0000185314	PREDICTED: uncharacterized protein LOC103446519	9	8.39E-27	63.22%	-	
MDP0000185315	probable s-adenosylmethionine-dependent methyltransferase at5g38100	20	4.27E-36	70.35%	F:methyltransferase activity	-
MDP0000186418	mannosyl-oligosaccharide glucosidase gcs1-like	2.00E+01	1.59E-162	89.25%	F:alpha-1,4-glucosidase activity; F:mannosyl-oligosaccharide glucosidase activity; P:epidermal cell differentiation; P:root epidermal cell differentiation; P:starch metabolic process; P:sucrose metabolic process; P:galactose metabolic process; C:glucosidase II complex	EC:3.2.1.106; EC:3.2.1.20; EC:3.2.1
MDP0000191641	wd-repeat isoform partial	20	4.06E-93	83.45%	-	
MDP0000192054	heterogeneous nuclear ribonucleoprotein 1-like	20	0	79.15%	F:nucleic acid binding; F:nucleotide binding	-
MDP0000192819	soluble diacylglycerol acyltransferase	20	0	71.20%	F:diacylglycerol O-acyltransferase activity; P:triglyceride biosynthetic process; C:cytosol; F:transferase activity; P:metabolic process; F:transferase activity, transferring acyl groups	
MDP0000195139	disease resistance protein rga3	2000.00%	0.00%	80.35%	F:protein binding; F:ADP binding	-
MDP0000195887	probable polyribonucleotide nucleotidyltransferase chloroplastic	20	4.70E-48	68.80%	F:transferase activity; P:RNA metabolic process	-
MDP0000195987	probable receptor protein kinase tmk1	20	1.65E-49	71.85%	C:membrane; F:nucleotide binding; F:protein serine/threonine kinase activity; P:signal transduction; P:phosphorylation; P:serine family amino acid metabolic process	EC:2.7.11
MDP0000196140	PREDICTED: uncharacterized protein LOC103410720 isoform X2	2.00E+01	0.00%	88.95%	C:plasma membrane	-
MDP0000197246	dna-directed rna polymerases iv and v subunit 6a-like	2.00E+01	3.68E-59	77.25%	C:DNA-directed RNA polymerase IV complex; C:DNA-directed RNA polymerase V complex; C:DNA-directed RNA polymerase II, core complex; F:DNA binding; F:DNA-directed RNA polymerase activity; P:transcription, DNA-templated; C:nucleolus; P:purine nucleobase metabolic process; P:pyrimidine nucleobase metabolic process	EC:2.7.7.6

MDP0000197247	phosphatidylinositol n-acetylglucosaminyltransferase subunit c	20	3.79E-129	89.45%	C:integral component of membrane; F:phosphatidylinositol N-acetylglucosaminyltransferase activity; P:GPI anchor biosynthetic process; P:pollen germination; P:pollen tube growth	EC:2.4.1; EC:2.4.1.198
MDP0000198805	protein kinase pinoid	20	1.51E-57	69.05%	C:cell surface; F:protein serine/threonine kinase activity; F:identical protein binding; P:response to light stimulus; P:auxin-activated signaling pathway; P:auxin polar transport; P:positive gravitropism; P:root hair initiation; P:root hair elongation; P:cotyledon development; P:response to karrikin; P:serine family amino acid metabolic process; F:ATP binding; P:protein phosphorylation	EC:2.7.11
MDP0000202018	probable receptor protein kinase tmk1	20	3.87E-42	96.80%	C:plasma membrane; C:integral component of membrane; F:transmembrane receptor protein serine/threonine kinase activity; F:ATP binding; P:protein phosphorylation; P:transmembrane receptor protein serine/threonine kinase signaling pathway; P:serine family amino acid metabolic process	EC:2.7.11
MDP0000203903	probable polyribonucleotide nucleotidyltransferase chloroplastic	20	0	88.60%	C:chloroplast stroma; F:3'-5'-exoribonuclease activity; F:RNA binding; F:polyribonucleotide nucleotidyltransferase activity; P:mRNA catabolic process; P:negative regulation of isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway; P:chlorophyll biosynthetic process; P:cellular response to phosphate starvation; P:carotene biosynthetic process; P:xanthophyll biosynthetic process; P:chloroplast RNA processing; P:RNA phosphodiester bond hydrolysis, exonucleolytic; P:regulation of RNA metabolic process; P:purine nucleobase metabolic process; P:pyrimidine nucleobase metabolic process	EC:3.1; EC:3.1.13; EC:3.1.15; EC:2.7.7.8
MDP0000209135	heat stress transcription factor b-4	2.00E+01	0.00%	74.75%	C:nucleus; F:sequence-specific DNA binding transcription factor activity; F:sequence-specific DNA binding; P:regulation of transcription, DNA-templated; P:response to heat; C:transcription factor complex; P:regulation of transcription, DNA-templated	-
MDP0000209670		100.00%	0.00%	100%	F:protein binding; P:regulation of transcription, DNA-templated	-
MDP0000209674	haloacid dehalogenase-like hydrolase domain-containing protein 3	2000.00%	1.65E-38	91.15%	F:hydrolase activity; P:metabolic process	-
MDP0000219303	transcription factor hec2-like	20	1.32E-98	68.55%	F:protein dimerization activity; P:gynoecium development	-
MDP0000230533	PREDICTED: uncharacterized protein LOC103441111	2000.00%	0.00%	88.75%	C:plasma membrane	-
MDP0000238968	protein chloroplastic	20	0	70.75%	-	
MDP0000243236	cation calcium exchanger 2-like	2000.00%	0.00%	76.60%	C:integral component of membrane; P:transmembrane transport	-
MDP0000243380		100.00%	0.00%	100%	F:protein binding	-

MDP0000245028	probable serine threonine-protein kinase at1g01540	20	0	86.35%	F:protein serine/threonine kinase activity; F:non-membrane spanning protein tyrosine kinase activity; F:ATP binding; P:peptidyl-tyrosine phosphorylation; P:serine family amino acid metabolic process	EC:2.7.10; EC:2.7.11; EC:2.7.10.2
MDP0000249205	proline-rich protein 12-like	20	5.26E-31	64.55%	-	
MDP0000249209	xylosyltransferase 1-like	20	0	89.75%	C:membrane; F:acetylglucosaminyltransferase activity; P:metabolic process; P:double fertilization forming a zygote and endosperm; P:pollen tube development	EC:2.4.1
MDP0000249539	disease resistance protein rga3	2.00E+01	0.00%	83.25%	F:protein binding; F:ADP binding	-
MDP0000249581	dna-directed rna polymerase subunit beta	2.00E+01	0.00%	82.95%	-	
MDP0000254312	potassium channel skor	20	0	93.20%	F:outward rectifier potassium channel activity; P:potassium ion transmembrane transport; C:voltage-gated potassium channel complex; F:protein binding	-
MDP0000255896	probable sucrose-phosphate synthase 4	20	4.38E-14	93.10%	C:plasma membrane; F:sucrose-phosphate synthase activity; P:sucrose metabolic process; P:biosynthetic process; P:starch metabolic process	EC:2.4.1.14; EC:2.4.1
MDP0000258367	---NA---	0.00E+00			F:phosphoric diester hydrolase activity; P:lipid metabolic process	EC:3.1
MDP0000268468	protein root initiation defective 3	20	2.33E-35	51.10%	F:protein binding	-
MDP0000269048	maltase- intestinal	2000.00%	5.18E-105	77.30%	-	
MDP0000273271	disease resistance protein rga3	2000.00%	1.19E-144	82.90%	F:protein binding	-
MDP0000274556	probable mitochondrial-processing peptidase subunit beta	20	1.14E-71	92.65%	F:metalloendopeptidase activity; F:metal ion binding; P:proteolysis	EC:3.4.24
MDP0000275168	PREDICTED: uncharacterized protein LOC103407361	2.00E+01	1.17E-102	54.00%	F:zinc ion binding	-
MDP0000275789	protein srg1-like	20	1.01E-68	69.70%	F:oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors	EC:1.14.11
MDP0000276022	PREDICTED: uncharacterized protein LOC103401439	2.00E+01	0.00%	86.30%	F:triglyceride lipase activity; P:lipid catabolic process; P:glycerolipid metabolic process	EC:3.1.1; EC:3.1.1.1; EC:3.1.1.3
MDP0000278543	50s ribosomal protein l25-like	2000.00%	8.99E-175	88.00%	C:ribosome; F:structural constituent of ribosome; F:5S rRNA binding; P:translation; P:ribosome biogenesis; F:protein binding; P:signal transduction	-
MDP0000280551	pyrophosphate-energized vacuolar membrane proton pump-like	20	0	96.80%	C:membrane; F:inorganic diphosphatase activity; F:hydrogen-translocating pyrophosphatase activity; P:proton transport; P:transmembrane transport; P:oxidative phosphorylation	EC:3.6.1.1; EC:3.6.1
MDP0000281041	potassium channel skor-like isoform x2	20	8.17E-65	88.05%	F:outward rectifier potassium channel activity; P:potassium ion transmembrane transport; C:voltage-gated potassium channel complex	-
MDP0000281449	alpha beta-hydrolases superfamily	2000.00%	0.00%	82.80%	F:triglyceride lipase activity; P:lipid catabolic process; P:glycerolipid metabolic process	EC:3.1.1; EC:3.1.1.1; EC:3.1.1.3
MDP0000282711	pra1 family protein f4-like	20	2.82E-90	70.80%	-	

MDP0000283955	pentatricopeptide repeat-containing protein at1g26500	2000.00%	0.00%	85.40%	-	
MDP0000283985	---NA---	0.00E+00			F:RNA binding; F:nucleotide binding	-
MDP0000284296	nad -binding rossmann-fold superfamily protein isoform 1	2.00E+01	0.00%	84.35%	F:catalytic activity; P:metabolic process	-
MDP0000284298	probable receptor protein kinase tmk1	20	2.28E-93	69.70%	F:ATP binding; P:protein phosphorylation; F:protein serine/threonine kinase activity	EC:2.7.11
MDP0000290186	disease resistance protein rga3	2.00E+01	0.00%	78.85%	F:ADP binding; F:protein binding	-
MDP0000291076	60s ribosomal protein l19-1	2.00E+01	8.33E-114	97.75%	C:nucleolus; C:plasma membrane; C:plasmodesma; C:cytosolic large ribosomal subunit; F:structural constituent of ribosome; P:translation; P:ribosome biogenesis	-
MDP0000291077	proline--trna ligase	20	0	93.30%	C:cytosol; C:plasmodesma; C:membrane; F:proline-tRNA ligase activity; F:ATP binding; P:prolyl-tRNA aminoacylation; P:arginine metabolic process; P:proline metabolic process	EC:6.1.1; EC:6.1.1.15
MDP0000291386	phd finger protein male meiocyte death 1	20	0	76.80%	F:zinc ion binding; P:reproduction; P:single-organism cellular process; F:protein binding	-
MDP0000291387	potassium channel skor-like isoform x3	20	6.95E-42	86.35%	F:outward rectifier potassium channel activity; P:potassium ion transmembrane transport; C:voltage-gated potassium channel complex	-
MDP0000292500	probable polyribonucleotide nucleotidyltransferase chloroplastic	20	0	87.55%	C:chloroplast stroma; F:3'-5'-exoribonuclease activity; F:RNA binding; F:polyribonucleotide nucleotidyltransferase activity; P:mRNA catabolic process; P:negative regulation of isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway; P:chlorophyll biosynthetic process; P:cellular response to phosphate starvation; P:carotene biosynthetic process; P:xanthophyll biosynthetic process; P:chloroplast RNA processing; P:RNA phosphodiester bond hydrolysis, exonucleolytic; P:regulation of RNA metabolic process; P:purine nucleobase metabolic process; P:pyrimidine nucleobase metabolic process	EC:3.1; EC:3.1.13; EC:3.1.15; EC:2.7.7.8
MDP0000292776	wd-repeat isoform partial	20	2.88E-44	56.85%	F:protein binding	-
MDP0000293040	replication factor c subunit 3	20	0	96.70%	F:nucleotide binding; F:DNA binding; F:nucleoside-triphosphatase activity; P:DNA replication	EC:3.6.1; EC:3.6.1.15
MDP0000294613	---NA---	0.00E+00			F:transferase activity, transferring acyl groups other than amino-acyl groups	-
MDP0000294614	uncharacterized acetyltransferase at3g50280-like	20	0	80.30%	F:transferase activity, transferring acyl groups other than amino-acyl groups; F:DNA binding; P:regulation of transcription, DNA-templated	-
MDP0000297288	trafficking protein particle complex subunit 9 isoform x1	20	9.50E-125	88.90%	-	
MDP0000298555	protein gamete expressed 1	20	0	80.55%	-	
MDP0000303781	f-box lrr-repeat protein 13-like isoform x2	20	9.39E-134	80.15%	-	
MDP0000303782	cobw domain-containing protein 1-like	2.00E+01	0.00%	84.70%	C:plastid chromosome; C:chloroplast stroma	-
MDP0000304326	triphosphate tunel metalloenzyme 3-like	20	1.15E-130	85.45%	-	

MDP0000304327	nbs-lrr disease resistance protein	2.00E+01	0.00%	79.95%	F:ADP binding; F:diacylglycerol kinase activity; P:protein kinase C-activating G-protein coupled receptor signaling pathway; P:metabolic process; F:NAD+ kinase activity	EC:2.7.1.23; EC:2.7.1.107
MDP0000304601	disease resistance protein rga3	2000.00%	0.00%	80.85%	F:protein binding; F:ADP binding	-
MDP0000310375	maltase- intestinal	2.00E+01	0.00%	83.10%	F:zinc ion binding	
MDP0000310376	cationic amino acid transporter vacuolar-like	2.00E+01	0.00%	90.95%	C:plant-type vacuole membrane; C:integral component of membrane; F:amino acid transmembrane transporter activity; P:amino acid transmembrane transport	-
MDP0000315270	ankyrin repeat-containing protein at3g12360-like	2000.00%	1.02E-145	84.75%	F:coenzyme binding; P:coenzyme A metabolic process; F:hydroxymethylglutaryl-CoA reductase (NADPH) activity; P:oxidation-reduction process; F:protein dimerization activity; P:transcription, DNA-templated	EC:1.1.1.34
MDP0000320739		100.00%	0.00%	100%	F:protein dimerization activity	-
MDP0000322089	PREDICTED: uncharacterized protein LOC103324616	2.00E+01	0.00%	62.60%	F:nucleic acid binding	-
MDP0000328704	---NA---	0.00%			-	
MDP0000331298	hsp40 cysteine-rich domain superfamily protein isoform 1	2.00E+01	6.00E-87	83.65%	F:heat shock protein binding; F:unfolded protein binding	-
MDP0000338651	protein unc-13-c-like protein	20	1.10E-18	94.55%	F:molecular_function; P:biological_process; C:cellular_component	
MDP0000343581	---NA---	0.00%			-	
MDP0000359637	---NA---	0.00%			-	
MDP0000360924	---NA---	0.00E+00			-	
MDP0000361342	pollen-specific leucine-rich repeat extensin-like protein 1	13	4.58E-29	70.92%	-	
MDP0000389293	probable adp-ribosylation factor gtpase-activating protein agd6	20	1.81E-19	94.15%	C:nucleus; C:cytosol; F:ARF GTPase activator activity; F:zinc ion binding; F:transaminase activity; P:regulation of ARF GTPase activity; P:positive regulation of GTPase activity	-
MDP0000439832	uncharacterized oxidoreductase at4g09670-like	20	0	85.25%	F:oxidoreductase activity; P:oxidation-reduction process	-
MDP0000440151	potassium channel skor-like isoform x2	20	2.47E-39	87.90%	F:outward rectifier potassium channel activity; P:potassium ion transmembrane transport; C:voltage-gated potassium channel complex	-
MDP0000445632	28s ribosomal protein mitochondrial	2.00E+01	6.48E-47	81.60%	C:ribosome; F:structural constituent of ribosome; P:translation; P:ribosome biogenesis	-
MDP0000469664	lrr receptor-like serine threonine-protein kinase rch1	2000.00%	0.00%	85.65%	C:integral component of membrane; F:protein kinase activity; F:ATP binding; P:protein phosphorylation; F:protein binding	-
MDP0000512999	glycine dehydrogenase mitochondrial	20	4.77E-40	84.95%	C:mitochondrion; C:chloroplast thylakoid; C:chloroplast stroma; C:chloroplast envelope; C:apoplast; F:glycine dehydrogenase (decarboxylating) activity; F:ATP binding; F:pyridoxal phosphate binding; P:glycine catabolic process; P:response to cadmium ion; P:oxidation-reduction process; P:L-serine metabolic process; P:threonine metabolic process; P:biosynthetic process	EC:1.4.4.2

MDP0000514448	vq motif-containing	20	1.54E-118	65.95%	-	
MDP0000516523	---NA---	0.00%			-	
MDP0000519674	subtilisin-like protease	20	8.55E-90	70.50%	P:proteolysis; F:serine-type endopeptidase activity	EC:3.4.21
MDP0000523812	transcription factor spatula-like isoform x1	20	6.22E-132	79.90%	F:protein dimerization activity	-
MDP0000527046	esterase vc_a0580-like	2000.00%	5.08E-91	88.85%	C:peroxisome; P:phylloquinone biosynthetic process	-
MDP0000531724	phosphate transporter pho1 homolog 10-like	20	3.46E-61	67.25%	C:integral component of membrane; P:protein phosphorylation; F:protein kinase activity; F:ATP binding	-
MDP0000572169	zinc finger protein 4-like	20	1.54E-37	68.90%	F:metal ion binding	-
MDP0000577338	PREDICTED: uncharacterized protein LOC103423750	2.00E+01	1.17E-66	73.05%	-	
MDP0000636876	tmv resistance protein n-like	20	0	79.90%	F:ADP binding; P:defense response; P:signal transduction; F:protein binding	-
MDP0000650075	probable boi-related e3 ubiquitin-protein ligase 2	20	6.20E-176	82.00%	F:zinc ion binding; F:protein binding	-
MDP0000652388	cation calcium exchanger 1-like	2.00E+01	0.00%	80.20%	C:integral component of membrane; P:transmembrane transport	-
MDP0000711911	nigrin b-like	2000.00%	0.00%	65.75%	F:rRNA N-glycosylase activity; P:negative regulation of translation	EC:3.2.2.22
MDP0000713910	transcription factor bhlh95	20	2.28E-168	70.40%	F:protein dimerization activity	-
MDP0000717184	protein suppressor of npr1- constitutive 1-like	20	9.07E-163	78.35%	F:protein binding	-
MDP0000717791	probable boi-related e3 ubiquitin-protein ligase 2	20	6.18E-178	83.95%	F:zinc ion binding; F:protein binding	-
MDP0000729521	uncharacterized acetyltransferase at3g50280-like	20	0	81.25%	F:transferase activity, transferring acyl groups other than amino-acyl groups; P:metabolic process	-
MDP0000735372	magnesium-protoporphyrin ix monomethyl ester	2.00E+01	2.85E-135	91.85%	C:chloroplast thylakoid; C:chloroplast envelope; F:metal ion binding; F:magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase activity; P:chloroplast organization; P:photosynthesis; P:chlorophyll biosynthetic process; P:oxidation-reduction process; P:regulation of tetrapyrrole metabolic process	EC:1.14.13.81; EC:1.14.13
MDP0000745770	u-box domain-containing protein 21-like	20	7.62E-82	78.50%	C:ubiquitin ligase complex; F:ubiquitin-protein transferase activity; P:protein ubiquitination	-
MDP0000747281	btb poz and math domain-containing protein 4-like	2000.00%	0.00%	91.70%	C:cytosol; P:cellular response to water deprivation; P:cellular response to salt stress; F:protein binding	-
MDP0000753788	40s ribosomal protein s13	2.00E+01	7.64E-66	97.80%	C:nucleolus; C:endoplasmic reticulum; C:ribosome; C:membrane; F:structural constituent of ribosome; P:cytokinesis by cell plate formation; P:translation; P:leaf morphogenesis; P:trichome morphogenesis; P:ribosome biogenesis	-
MDP0000760132	pentatricopeptide repeat-containing protein at1g26500	2000.00%	0.00%	85.60%	-	
MDP0000775126	PREDICTED: pinin-like	9.00E+00	1.73E-39	80.67%	-	

	MDP0000784090	tmv resistance protein n-like	20	4.04E-19	54.60%	-	
	MDP0000784168	mitogen-activated protein kinase kinase kinase 3-like	2.00E+01	0.00%	79.65%	F:protein serine/threonine kinase activity; F:ATP binding; P:protein phosphorylation; P:serine family amino acid metabolic process	EC:2.7.11
	MDP0000806017	---NA---	0.00%			-	
	MDP0000810351	disease resistance protein rga3	20	0.00%	81.05%	F:protein binding; F:ADP binding	-
	MDP0000818448	disease resistance protein rga3	2000.00%	1.87E-142	80.85%	F:protein binding	-
	MDP0000835932	mitogen-activated protein kinase kinase kinase 3-like	2.00E+01	0.00%	79.65%	F:protein serine/threonine kinase activity; F:ATP binding; P:protein phosphorylation; P:serine family amino acid metabolic process	EC:2.7.11
	MDP0000848029	protein root initiation defective 3-like	20	1.37E-38	79.80%	-	
	MDP0000851135	pentatricopeptide repeat-containing protein at5g66520-like	20	6.94E-28	78.55%	-	
	MDP0000853127	arogenate dehydrogenase chloroplastic-like	2.00E+01	0.00%	88.20%	F:prephenate dehydrogenase (NADP+) activity; F:prephenate dehydrogenase activity; P:tyrosine biosynthetic process; P:oxidation-reduction process; P:tryptophan biosynthetic process; P:L-phenylalanine biosynthetic process	EC:1.3.1.12; EC:1.3.1.13
	MDP0000857821	protein chloroplastic	20	1.75E-51	93.95%	-	
	MDP0000867534	PREDICTED: uncharacterized protein LOC103950428	2.00E+01	6.98E-52	77.40%	-	
	MDP0000869168	pi-plc x domain-containing protein at5g67130-like	20	7.31E-96	82.05%	F:phosphoric diester hydrolase activity; P:lipid metabolic process	EC:3.1
	MDP0000879254	replication factor c subunit 3	20	1.06E-111	97.10%	F:nucleotide binding; F:DNA binding; F:nucleoside-triphosphatase activity; P:DNA replication	EC:3.6.1; EC:3.6.1.15
	MDP0000879258	splicing factor 3b subunit	20	1.30E-23	95.60%	-	
	MDP0000893755	cationic amino acid transporter vacuolar-like	2000.00%	0.00%	90.85%	C:plant-type vacuole membrane; C:integral component of membrane; F:amino acid transmembrane transporter activity; P:amino acid transmembrane transport	-
	MDP0000921094	subtilisin-like protease	20	0	80.70%	C:extracellular region; F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21
Bitterness	MDP0000052541	embryo defective 1381 isoform 1	20	0	90.35%	-	
	MDP0000119516	3-ketodihydrosphingosine reductase-like	20	0	88.50%	F:oxidoreductase activity; P:metabolic process	-
	MDP0000123824	26s proteasome non-atpase regulatory subunit 7 homolog a-like	20	4.55E-10	58.85%	P:leaf morphogenesis; C:cytosol; P:embryo development ending in seed dormancy; C:proteasome complex	
	MDP0000124900	methionine aminopeptidase chloroplastic	20	0	86.90%	P:proteolysis; F:aminopeptidase activity; F:metalloexopeptidase activity	EC:3.4.11
	MDP0000127054	ethylene-responsive transcription factor erf003-like	20	1.58E-101	84.35%	P:regulation of transcription, DNA-templated; F:sequence-specific DNA binding transcription factor activity; F:DNA binding; C:transcription factor complex; P:regulation of transcription, DNA-templated	-
	MDP0000128281	vacuolar protein 8	20	0	87.00%	F:protein binding	-
	MDP0000129011	transcription factor bhlh68-like isoform x1	20	0	77.40%	F:protein dimerization activity	-

MDP0000131356	probable carboxylesterase 6	20	0	74.50%	P:metabolic process; F:hydrolase activity	-
MDP0000132527	sucrose synthase	20	0	88.30%	P:sucrose metabolic process; P:biosynthetic process	-
MDP0000134560	probable histone-lysine n-methyltransferase atxr3	20	3.65E-64	80.25%	F:methyltransferase activity; P:methylation	-
MDP0000135679	thiamine-repressible mitochondrial transport protein thi74-like	20	0	78.10%	C:integral component of membrane	-
MDP0000135680	probable monogalactosyldiacylglycerol chloroplastic	20	0	87.65%	P:lipid glycosylation; F:carbohydrate binding; F:transferase activity, transferring hexosyl groups; P:glycolipid biosynthetic process	-
MDP0000136728	methionine aminopeptidase chloroplastic	20	0	86.70%	F:aminopeptidase activity; P:proteolysis; F:metalloexopeptidase activity	EC:3.4.11
MDP0000139500	probable histone-lysine n-methyltransferase atxr3	20	3.66E-64	80.25%	F:methyltransferase activity; P:methylation	-
MDP0000141005	serine threonine-protein phosphatase 2a 65 kda regulatory subunit a beta isoform	20	0	94.55%	F:binding	-
MDP0000145027	tyrosine-protein phosphatase	20	2.37E-158	86.40%	F:protein tyrosine phosphatase activity; P:protein dephosphorylation; P:tyrosine metabolic process	EC:3.1.3.16; EC:3.1; EC:3.1.3.48; EC:3.1.3.41
MDP0000148855	protein glutamine dumper 5-like	20	2.18E-94	76.40%	-	
MDP0000154158	gdt1-like protein 4	20	4.03E-35	96.05%	C:membrane	-
MDP0000155087	uncharacterized loc101206567	20	0	89.80%	C:endosome; C:trans-Golgi network	-
MDP0000155673	spindle and kinetochore-associated 2	20	1.92E-75	90.05%	-	
MDP0000155674	dynamin-related protein 3a-like isoform x1	20	0	80.80%	F:GTPase activity; F:GTP binding	EC:3.6.1; EC:3.6.1.15
MDP0000155675	protein tic chloroplastic-like	20	0	90.00%	F:chlorophyllide a oxygenase [overall] activity; P:oxidation-reduction process; F:2 iron, 2 sulfur cluster binding	EC:1.13.12; EC:1.14.13.122
MDP0000157412	ctl-like protein ddb_g0274487	20	0	88.70%	C:integral component of membrane	-
MDP0000159583	PREDICTED: uncharacterized protein LOC103416009	20	2.32E-113	61.15%	-	
MDP0000160232	disease resistance protein rga3	20	0	78.55%	F:protein binding; F:ADP binding	-
MDP0000160621	probable mitochondrial adenine nucleotide transporter btl3	20	1.71E-32	55.95%	C:integral component of membrane; C:membrane; P:transmembrane transport; P:transport	
MDP0000167338	duf868 family protein	20	0	73.80%	C:plasma membrane	
MDP0000167343	probable flavin-containing monooxygenase 1	20	0	86.15%	F:NADP binding; F:flavin adenine dinucleotide binding; P:oxidation-reduction process; F:N,N-dimethylaniline monooxygenase activity	EC:1.14.13.8; EC:1.14.13
MDP0000171928	leucoanthocyanidin reductase-like	20	2.52E-131	92.90%	F:leucoanthocyanidin reductase activity; P:oxidation-reduction process	EC:1.17.1; EC:1.17.1.3
MDP0000171929	leucoanthocyanidin reductase	20	1.97E-22	83.60%	F:leucoanthocyanidin reductase activity; P:oxidation-reduction process	EC:1.17.1; EC:1.17.1.3
MDP0000180004	allene oxide cyclase chloroplastic-like	20	2.11E-140	85.80%	F:isomerase activity; C:chloroplast	-
MDP0000180005	methionine aminopeptidase chloroplastic-like	20	0	87.90%	F:aminopeptidase activity; F:metalloexopeptidase activity; P:proteolysis	EC:3.4.11
MDP0000181021	---NA---	0			-	
MDP0000181352	cytochrome c-type biogenesis	20	2.05E-103	91.40%	C:mitochondrial inner membrane; C:protein complex; F:oxidoreductase activity; P:embryo development; P:oxidation-reduction process	-

MDP0000182500	cysteine-rich receptor-like protein kinase 10	20	0	68.00%	F:ATP binding; P:protein phosphorylation; F:protein serine/threonine kinase activity; P:serine family amino acid metabolic process	EC:2.7.11	
MDP0000188242	dof zinc finger	20	2.11E-96	70.50%	F:translation elongation factor activity; C:eukaryotic translation elongation factor 1 complex; F:DNA binding; P:regulation of transcription, DNA-templated; C:ribosome; P:regulation of translational elongation	-	
MDP0000188336	mediator of rna polymerase ii transcription subunit 25-like isoform x2	20	0	83.25%	P:signal transduction; P:response to external stimulus; P:response to red or far red light; P:regulation of flower development; P:positive regulation of biological process	-	
MDP0000188337	lysine histidine transporter-like 6	20	0	91.80%	C:integral component of membrane	-	
MDP0000188338	40s ribosomal protein s3a-2-like	20	8.33E-93	64.70%	F:structural constituent of ribosome; P:translation; C:ribosome; P:ribosome biogenesis	-	
MDP0000190319	protein transport protein sec24-like at4g32640	20	3.14E-59	67.70%	P:protein transport; P:intracellular transport	-	
MDP0000193411	cysteine-rich rlk isoform 1	20	1.18E-74	70.90%	P:protein phosphorylation; F:protein kinase activity; F:ATP binding	-	
MDP0000198209	leucine-rich repeat receptor-like serine threonine-protein kinase at2g14440	20	5.91E-71	61.85%	C:endosome; C:vacuole; C:trans-Golgi network; C:plasma membrane	-	
MDP0000201211							
MDP0000201494	cysteine-rich receptor-like protein kinase 10	20	0	68.00%	F:ATP binding; P:protein phosphorylation; F:protein serine/threonine kinase activity; P:serine family amino acid metabolic process	EC:2.7.11	
MDP0000202669	zinc finger protein constans-like 4	20	0	86.50%	F:zinc ion binding; C:intracellular; F:protein binding	-	
MDP0000207420	palmitoyl-monogalactosyldiacylglycerol delta-7 chloroplastic-like	20	4.14E-81	83.50%	P:oxidation-reduction process; F:oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water	EC:1.14.19	
MDP0000207423	vacuolar protein 8	20	0	87.10%	F:protein binding; C:vacuolar proton-transporting V-type ATPase, V1 domain; F:proton-transporting ATPase activity, rotational mechanism; P:ATP hydrolysis coupled proton transport; P:oxidative phosphorylation	EC:3.6.1; EC:3.6.1.15	EC:3.6.1.3;
MDP0000208044	metallo-hydrolase oxidoreductase superfamily protein isoform 1	20	0	90.20%	C:chloroplast; F:hydrolase activity; P:metabolic process	-	
MDP0000208045	40s ribosomal protein s3a-2-like	20	6.72E-42	55.20%	C:ribonucleoprotein complex; C:intracellular non-membrane-bounded organelle; C:cytoplasmic part	-	
MDP0000208046	lysine histidine transporter-like 6	20	0	92.40%	C:integral component of membrane	-	
MDP0000208228	n-acylphosphatidylethanolamine synthase-like	20	1.49E-77	65.25%	F:transferase activity, transferring acyl groups; P:phospholipid metabolic process	-	
MDP0000215799	serine threonine-protein kinase tor	20	1.35E-49	78.95%	F:protein binding	-	
MDP0000215801	catalytic coenzyme binding protein	20	0	85.85%	C:plasma membrane	-	
MDP0000216289	palmitoyl-monogalactosyldiacylglycerol delta-7 chloroplastic-like	20	1.51E-95	89.45%	F:oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen	EC:1.14.19	

					to two molecules of water; P:oxidation-reduction process; P:lipid metabolic process	
MDP0000216620	40s ribosomal protein s3a-2-like	20	8.47E-56	59.10%	C:ribosome; P:translation; F:structural constituent of ribosome; P:ribosome biogenesis	-
MDP0000220114	mediator of rna polymerase ii transcription subunit 25-like isoform x2	20	0	72.40%	P:response to red or far red light; P:positive regulation of biological process; P:regulation of cellular process	-
MDP0000220174	upf0481 protein at3g47200-like	20	0	68.70%	-	
MDP0000220175	disease resistance protein rga3	20	0	78.10%	F:protein binding; F:ADP binding	-
MDP0000220176	protein mitochondrial-like	20	3.52E-170	72.15%	F:single-stranded DNA binding; P:DNA replication	-
MDP0000220177	mitogen-activated protein kinase kinase kinase yoda-like	20	0	76.00%	P:protein phosphorylation; F:protein tyrosine kinase activity; F:ATP binding	EC:2.7.10
MDP0000220179	mitogen-activated protein kinase kinase kinase yoda-like	20	0	74.30%	P:protein phosphorylation; F:protein kinase activity; F:ATP binding	-
MDP0000220181	gamma-tubulin complex component 5-like isoform x1	20	0	82.00%	P:microtubule cytoskeleton organization; C:spindle pole; C:microtubule organizing center	-
MDP0000221435	cytochrome p450 cyp82d47-like	20	0	81.45%	F:oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; F:heme binding; P:oxidation-reduction process; F:iron ion binding	-
MDP0000221436	probable transcription factor kan4	20	0	61.10%	F:DNA binding; F:chromatin binding; C:chromatin	-
MDP0000221442	ent-kaurene chloroplastic-like	20	0	88.65%	F:heme binding; F:oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; P:oxidation-reduction process; F:iron ion binding	-
MDP0000221444	dof zinc finger	20	4.95E-148	69.65%	P:regulation of transcription, DNA-templated; F:DNA binding	-
MDP0000221451	nucleobase-ascorbate transporter 3-like	20	0	89.10%	C:membrane; F:transporter activity; P:transmembrane transport	-
MDP0000222306	probable flavin-containing monooxygenase 1	20	0	86.05%	P:oxidation-reduction process; F:flavin adenine dinucleotide binding; F:N,N-dimethylaniline monooxygenase activity; F:NADP binding	EC:1.14.13.8; EC:1.14.13
MDP0000222317	dna mismatch repair protein msh6	20	1.65E-76	89.85%	P:mismatch repair; F:mismatched DNA binding; F:ATP binding	-
MDP0000224592		1	0	100%	F:DNA binding; P:regulation of transcription, DNA-templated	-
MDP0000227833	3-ketodihydrosphingosine reductase-like	20	0	88.50%	P:metabolic process; F:oxidoreductase activity	-
MDP0000231832	gdt1-like protein 4	20	8.37E-69	89.00%	C:membrane	-
MDP0000231836	---NA---	0			-	
MDP0000233229	transcription factor pif1	20	0	76.35%	-	
MDP0000235023	mediator of rna polymerase ii transcription subunit 25-like isoform x2	20	1.28E-143	74.45%	P:response to red or far red light; P:positive regulation of biological process; P:regulation of cellular process	-
MDP0000235369	eukaryotic peptide chain release factor subunit 1-3	20	7.30E-100	67.85%	C:cytoplasm; F:translation release factor activity, codon specific; P:translational termination; F:translation release factor activity; C:cytosol; C:plasma membrane	
MDP0000239834	allene oxide cyclase chloroplastic-like	20	2.11E-140	85.80%	F:isomerase activity; C:chloroplast	-

MDP0000241811	serine threonine-protein phosphatase 2a 65 kda regulatory subunit a beta isoform	20	0	97.70%	F:binding	-
MDP0000242294	probable plastidic glucose transporter 3 isoform x1	20	2.08E-19	88.70%	C:integral component of membrane; F:transmembrane transporter activity; P:transmembrane transport	-
MDP0000243706	malonate-- ligase	20	0	73.45%	P:metabolic process; F:catalytic activity; P:spindle assembly; C:HAUS complex	-
MDP0000244253	copper transport protein	20	7.15E-53	89.75%	F:metal ion binding; P:metal ion transport	-
MDP0000247199	protein plant cadmium resistance 2-like	20	6.99E-113	80.40%	P:pollen sperm cell differentiation	-
MDP0000247659	peroxiredoxin chloroplastic-like	20	1.31E-142	89.20%	P:oxidation-reduction process; F:antioxidant activity; F:oxidoreductase activity	-
MDP0000248043	vq motif-containing protein	20	2.81E-103	75.65%	-	
MDP0000250967	ctl-like protein ddb_g0274487	20	0	90.25%	P:regulation of transcription, DNA-templated; F:sequence-specific DNA binding transcription factor activity; F:sequence-specific DNA binding; C:transcription factor complex; P:regulation of transcription, DNA-templated	-
MDP0000251295	1-aminocyclopropane-1-carboxylate oxidase	20	0	93.25%	F:oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors; P:oxidation-reduction process	EC:1.14.11
MDP0000252114	aluminum-activated malate transporter 4	20	0	84.30%	P:malate transport	-
MDP0000257928	protein glutamine dumper 5-like	20	6.33E-76	75.65%	-	
MDP0000257929	cbl-interacting serine threonine-protein kinase 6-like	20	0	83.90%	F:protein kinase activity; P:protein phosphorylation; F:ATP binding; F:omega peptidase activity	EC:3.4.19
MDP0000257931	3-hydroxyisobutyryl- hydrolase-like protein mitochondrial	20	0	82.35%	F:catalytic activity; P:metabolic process	-
MDP0000258718	branched-chain-amino-acid aminotransferase chloroplastic-like	20	0	83.35%	F:branched-chain-amino-acid transaminase activity; P:isoleucine catabolic process; P:leucine catabolic process; P:valine catabolic process; P:isoleucine biosynthetic process; P:leucine biosynthetic process; P:valine biosynthetic process; P:pantothenate biosynthetic process	EC:2.6.1.42
MDP0000261040		1	0	100%	F:substrate-specific transmembrane transporter activity; P:transmembrane transport; C:integral component of membrane	-
MDP0000261658	f-box protein at1g55000-like	20	4.56E-33	83.90%	-	
MDP0000262602	auxin-responsive protein iaa27-like	20	0	77.75%	P:regulation of transcription, DNA-templated; C:nucleus; F:protein dimerization activity	-
MDP0000263035	protein enhanced disease resistance 2-like	20	0	88.20%	C:nucleus	-
MDP0000264232	---NA---	0			P:oxidation-reduction process; C:membrane; F:oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor; F:peroxidase activity; P:peroxidase reaction; P:response to oxidative stress	EC:1.11.1.7
MDP0000266451	l-type lectin-domain containing receptor kinase -like	20	0	81.05%	P:protein phosphorylation; F:ATP binding; F:carbohydrate binding; F:protein	EC:2.7.11

					serine/threonine kinase activity; P:serine family amino acid metabolic process	
MDP0000266452	uncharacterized loc101212813	20	0	78.20%	-	
MDP0000266453		1	0	100%	P:oxidation-reduction process; P:response to oxidative stress; F:glutathione peroxidase activity; P:RNA processing; F:RNA binding; F:nucleotide binding; P:cell cycle; P:glutathione metabolic process; P:peroxidase reaction	EC:1.11.1.9; EC:1.11.1.7
MDP0000266454		1	0	100%	F:RNA binding; F:protein binding	-
MDP0000270281	protein fizzy-related 3	20	1.83E-50	90.30%	F:protein binding	-
MDP0000274967	---NA---	0			-	
MDP0000275365	transmembrane fragile-x-f-associated protein	20	0	72.85%	F:metal ion binding	-
MDP0000276832	transposase tnp2	20	1.82E-19	59.00%	-	
MDP0000286166	dna mismatch repair protein msh6-like	20	0	69.40%	F:mismatched DNA binding; P:mismatch repair; F:ATP binding	-
MDP0000286637	integral membrane hpp family protein	20	3.55E-127	82.65%	C:chloroplast inner membrane	-
MDP0000291732	---NA---	0			-	
MDP0000297646	ethylene-responsive transcription factor erf003-like	20	2.73E-103	84.50%	F:sequence-specific DNA binding transcription factor activity; P:regulation of transcription, DNA-templated; F:DNA binding; C:transcription factor complex; P:regulation of transcription, DNA-templated	-
MDP0000298053	leucine-rich repeat receptor-like serine threonine-protein kinase at2g14440	20	0	83.85%	C:endosome; C:vacuole; C:trans-Golgi network; C:plasma membrane; F:kinase activity; P:phosphorylation	-
MDP0000300090	haus augmin-like complex subunit 2	20	0	91.60%	F:substrate-specific transmembrane transporter activity; P:transmembrane transport; C:integral component of membrane; P:spindle assembly; P:microtubule organizing center organization	-
MDP0000300939	septum site-determining protein mind chloroplastic	20	1.16E-119	74.00%	C:chloroplast part; F:nucleotide binding; F:ATPase activity; F:identical protein binding; P:cellular component organization; P:single-organism cellular process	EC:3.6.1; EC:3.6.1.3; EC:3.6.1.15
MDP0000300940	---NA---	0			-	
MDP0000302718	PREDICTED: uncharacterized protein LOC103417303	1	5.37E-22	100.00%	-	
MDP0000303379	PREDICTED: uncharacterized protein LOC103416007	20	7.97E-99	60.30%	-	
MDP0000303483	protein transparent testa 1-like	20	0	82.30%	F:nucleic acid binding; F:metal ion binding	-
MDP0000303496	tmv resistance protein n-like	20	0	72.30%	P:signal transduction; F:protein binding; F:ADP binding	-
MDP0000311232	14-3-3-like protein gf14 kappa isoform x1	20	2.80E-29	87.15%	F:protein domain specific binding	-
MDP0000311636	disease resistance protein rga3	20	2.16E-21	71.45%	C:membrane; F:transporter activity; P:transmembrane transport	-
MDP0000312598		1	0	100%	F:DNA binding; F:catalytic activity; F:pyridoxal phosphate binding	-

MDP0000312601	lysine histidine transporter-like 8	20	0	97.70%	C:integral component of membrane	-
MDP0000312602	ent-kaurene chloroplastic-like	20	0	79.65%	P:oxidation-reduction process; F:heme binding; F:oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; F:iron ion binding	-
MDP0000312604	serpin-zx-like	20	0	92.70%	C:extracellular space	-
MDP0000312606	disease resistance protein rga3	20	0	79.40%	F:protein binding; F:ADP binding	-
MDP0000315002	dof zinc finger	20	1.31E-96	75.65%	F:DNA binding; P:regulation of transcription, DNA-templated; C:eukaryotic translation elongation factor 1 complex; F:translation elongation factor activity; C:ribosome; P:regulation of translational elongation	-
MDP0000319726	transmembrane fragile-x-f-associated protein	20	0	83.60%	F:protein dimerization activity	-
MDP0000321326	glycosyl hydrolase family 17 family protein	20	0	87.90%	P:carbohydrate metabolic process; F:hydrolase activity, hydrolyzing O-glycosyl compounds	-
MDP0000328545	---NA---	0			-	
MDP0000332721	---NA---	0			-	
MDP0000334761	palmitoyl-monogalactosyldiacylglycerol delta-7 chloroplastic-like	20	1.79E-79	88.80%	P:lipid metabolic process	-
MDP0000337690	60s ribosomal protein l39-1	20	1.73E-29	97.60%	C:cytosolic large ribosomal subunit; F:structural constituent of ribosome; P:translation; P:ribosome biogenesis	-
MDP0000340082	transmembrane protein 147-like	20	1.30E-169	95.95%	C:mitochondrion; C:Golgi apparatus	-
MDP0000348117	---NA---	0			-	
MDP0000353444	rna polymerase beta subunit	20	1.70E-100	90.95%	-	
MDP0000357895	---NA---	0			-	
MDP0000375685	transcription factor rax3	20	0	69.25%	F:chromatin binding; F:DNA binding; C:chromatin	-
MDP0000376284	leucoanthocyanidin reductase-like	20	0	91.50%	-	
MDP0000376285	leucoanthocyanidin reductase	20	1.97E-22	83.60%	F:leucoanthocyanidin reductase activity; P:oxidation-reduction process	EC:1.17.1; EC:1.17.1.3
MDP0000385620	ribosomal protein l12 atp-dependent clp protease adaptor protein family protein isoform 1	20	2.45E-111	90.20%	P:protein catabolic process	-
MDP0000385622	40s ribosomal protein s3a-like	20	1.87E-52	75.60%	P:translation; F:structural constituent of ribosome; C:ribosome; P:ribosome biogenesis	-
MDP0000409515	l-type lectin-domain containing receptor kinase -like	20	0	72.20%	F:carbohydrate binding; F:protein kinase activity; P:protein phosphorylation; F:ATP binding	-
MDP0000422612	serine threonine-protein kinase chloroplastic	20	1.65E-31	75.70%	F:protein serine/threonine kinase activity; F:ATP binding; P:protein phosphorylation; P:serine family amino acid metabolic process	EC:2.7.11
MDP0000422617	pax3- and pax7-binding protein 1	20	0	81.80%	P:regulation of transcription, DNA-templated; F:DNA binding; C:nucleus; F:sequence-specific DNA binding transcription factor activity; C:transcription factor complex; P:regulation of transcription, DNA-templated	-

MDP0000425402	o-glycosyl hydrolases family 17 isoform partial	20	0	79.10%	-	
MDP0000445291	clathrin assembly family protein	20	0	71.10%	F:1-phosphatidylinositol binding; F:clathrin binding; P:clathrin coat assembly; C:clathrin-coated vesicle	-
MDP0000487985	u2 snrnp-associated surp motif-containing isoform x1	20	4.06E-171	88.00%	F:nucleotide binding; F:RNA binding; P:RNA processing	-
MDP0000506508	small ubiquitin-related modifier 1-like	20	7.87E-66	96.40%	F:protein binding	-
MDP0000523718	60s ribosomal protein l13-3-like	20	2.21E-27	77.90%	C:ribosome; P:translation; F:structural constituent of ribosome; P:ribosome biogenesis	-
MDP0000525232	probable protein phosphatase 2c 59	20	9.78E-161	86.50%	F:catalytic activity	-
MDP0000559531	septum site-determining protein mind chloroplastic	20	9.16E-26	80.55%	C:chloroplast stroma; C:chloroplast envelope; F:ATP binding; F:calcium-dependent ATPase activity; F:protein homodimerization activity; P:barrier septum site selection; P:ATP catabolic process; P:chloroplast fission	EC:3.6.1; EC:3.6.1.3; EC:3.6.1.15
MDP0000559532	septum site-determining protein mind chloroplastic	20	4.77E-21	93.50%	C:chloroplast stroma; C:chloroplast envelope; F:calcium-dependent ATPase activity; F:protein homodimerization activity; P:barrier septum site selection; P:ATP catabolic process; P:chloroplast fission	EC:3.6.1; EC:3.6.1.3; EC:3.6.1.15
MDP0000577455	saur family protein	20	6.80E-78	83.00%	-	
MDP0000600931	eukaryotic translation initiation factor 5a-2-like	20	1.66E-18	89.10%	C:nucleus; F:translation initiation factor activity; F:translation elongation factor activity; F:ribosome binding; P:translational frameshifting; P:peptidyl-lysine modification to peptidyl-hypusine; P:response to wounding; P:host programmed cell death induced by symbiont; P:defense response to bacterium; P:positive regulation of translational elongation; P:positive regulation of translational termination; P:response to cadmium ion; C:ribosome; P:regulation of translational initiation	-
MDP0000603942	leucoanthocyanidin reductase-like	20	0	91.50%	-	
MDP0000655848	glutamate synthase 1	20	1.58E-24	86.05%	C:chloroplast stroma; F:iron ion binding; F:FMN binding; F:glutamate synthase (NADH) activity; F:flavin adenine dinucleotide binding; F:iron-sulfur cluster binding; P:glutamate biosynthetic process; P:ammonia assimilation cycle; P:response to cadmium ion; P:developmental growth; P:oxidation-reduction process; P:electron transport	EC:1.4.1.14; EC:1.4; EC:1.4.1
MDP0000663430	zinc finger mym-type protein 5-like	20	1.93E-101	68.55%	F:binding	-
MDP0000703817	two-component response regulator	20	0	76.60%	F:chromatin binding; F:DNA binding; C:chromatin	-
MDP0000717027	14-3-3-like protein b	20	2.77E-29	91.35%	F:protein domain specific binding	-
MDP0000755936	peroxiredoxin chloroplastic-like	20	1.15E-103	95.00%	P:oxidation-reduction process; F:antioxidant activity; F:oxidoreductase activity	-
MDP0000755938	probable flavin-containing monooxygenase 1	20	1.81E-153	93.60%	P:oxidation-reduction process; F:N,N-dimethylaniline monooxygenase activity;	EC:1.14.13.8; EC:1.14.13

					F:NADP binding; F:flavin adenine dinucleotide binding	
MDP0000755991	polyol transporter 1	20	3.15E-35	67.40%	F:transmembrane transporter activity; C:integral component of membrane; P:transmembrane transport	-
MDP0000779630	palmitoyl-monogalactosyldiacylglycerol delta-7 chloroplastic-like	20	2.72E-34	90.15%	C:integral component of membrane; F:oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water; P:fatty acid biosynthetic process; P:oxidation-reduction process	EC:1.14.19
MDP0000794936	probable galacturonosyltransferase-like 3	20	0	89.75%	F:transferase activity, transferring glycosyl groups	-
MDP0000796893	palmitoyl-monogalactosyldiacylglycerol delta-7 chloroplastic-like	20	2.72E-34	90.15%	C:integral component of membrane; F:oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water; P:fatty acid biosynthetic process; P:oxidation-reduction process	EC:1.14.19
MDP0000806021	palmitoyl-monogalactosyldiacylglycerol delta-7 chloroplastic-like	20	6.29E-98	82.40%	F:oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water; P:oxidation-reduction process; P:lipid metabolic process	EC:1.14.19
MDP0000816018	palmitoyl-monogalactosyldiacylglycerol delta-7 chloroplastic-like	20	3.45E-29	95.05%	C:integral component of membrane; F:oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water; P:fatty acid biosynthetic process; P:oxidation-reduction process	EC:1.14.19
MDP0000816846	probable inactive receptor kinase at1g48480	20	3.19E-75	70.25%	C:membrane; F:nucleotide binding; F:protein kinase activity; P:metabolic process	-
MDP0000822021	adenosylhomocysteinase	20	0	98.40%	F:adenosylhomocysteinase activity; P:one-carbon metabolic process; P:methionine metabolic process	EC:3.3.1.1; EC:3.3.1
MDP0000836363	---NA---	0			-	
MDP0000859609	probable calcium-binding protein cml18	20	2.63E-114	81.90%	F:calcium ion binding	-
MDP0000868486	hypothetical protein PRUPE_ppa022986mg, partial	4	6.32E-27	60.75%	-	
MDP0000881783	transmembrane protein 147-like	20	1.35E-169	95.95%	C:mitochondrion; C:Golgi apparatus	-
MDP0000881805	probable thylakoidal processing peptidase chloroplastic	20	0	82.30%	F:serine-type peptidase activity; P:proteolysis; C:integral component of membrane	-
MDP0000912045	nucleobase-ascorbate transporter 3-like	20	2.81E-103	87.25%	F:transporter activity; C:membrane; P:transmembrane transport	-
MDP0000912056	---NA---	0			-	
MDP0000912059	thioredoxin-like protein chloroplastic	20	3.86E-133	82.65%	P:cell redox homeostasis; F:protein disulfide oxidoreductase activity; P:glycerol ether metabolic process; P:electron transport	-
MDP0000939633	ap2 erf and b3 domain-containing transcription factor rav1	20	0	78.35%	P:regulation of transcription, DNA-templated; F:sequence-specific DNA binding transcription	-

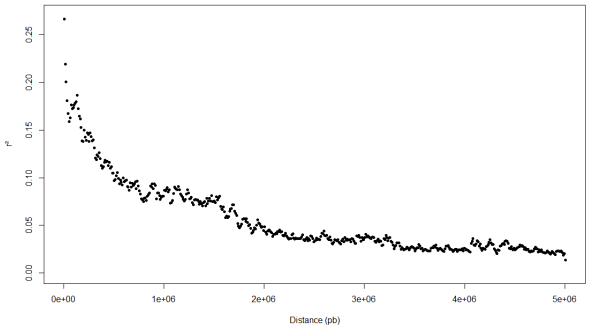
					factor activity; F:DNA binding; C:transcription factor complex; P:regulation of transcription, DNA-templated	
	MDP0000941318	bola-like protein 1	20	5.43E-117	76.60%	C:chloroplast	
	MDP0000952005	compass component swd1	20	6.61E-93	96.95%	F:protein binding	-
	MDP0000952010	kinesin-like protein kif11	20	0	84.70%	-	

[a] Mean Similarity

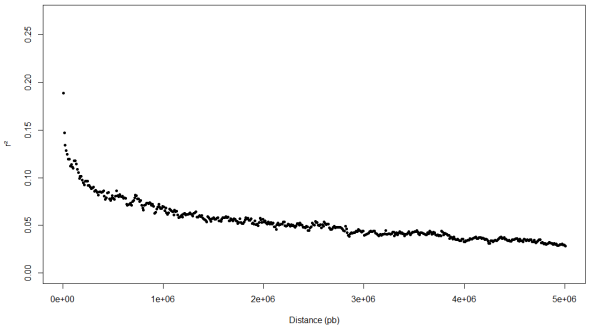
[b] Gene Ontology

Annexe 5 : Courbes de décroissance du déséquilibre de liaison par groupe de liaison (exprimé par le paramètre r^2), en fonction de la distance physique en utilisant les données de génotypage issus de la puce 480k SNPs sur la CC278

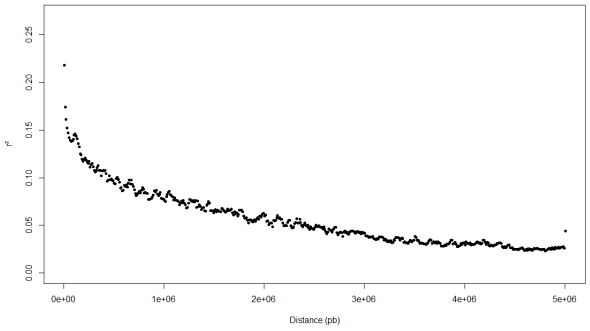
LG01



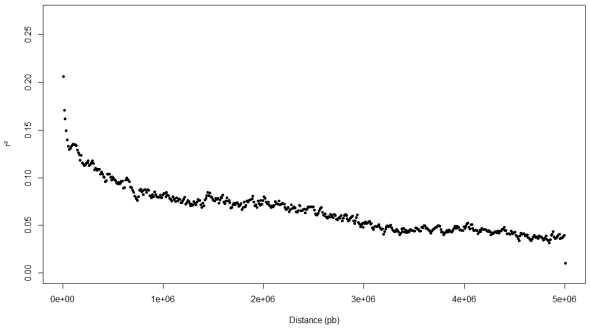
LG02



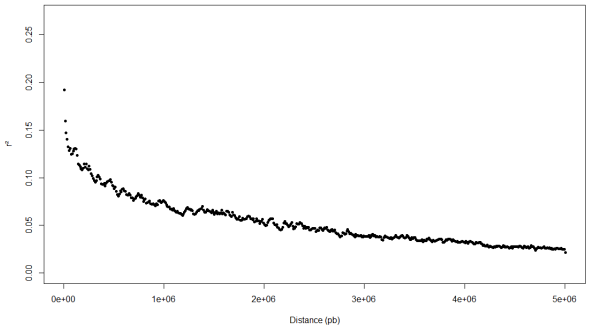
LG03



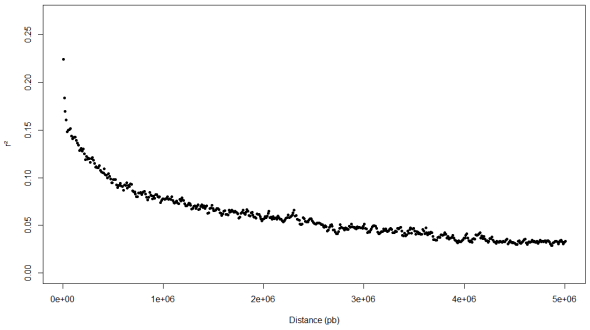
LG04



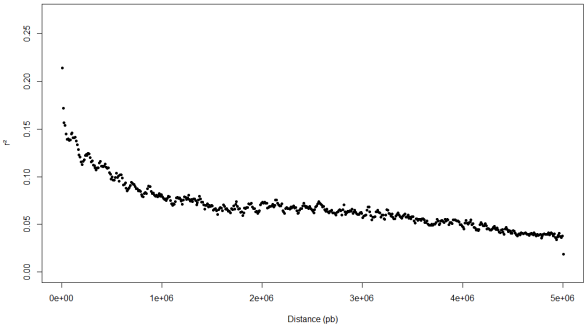
LG05



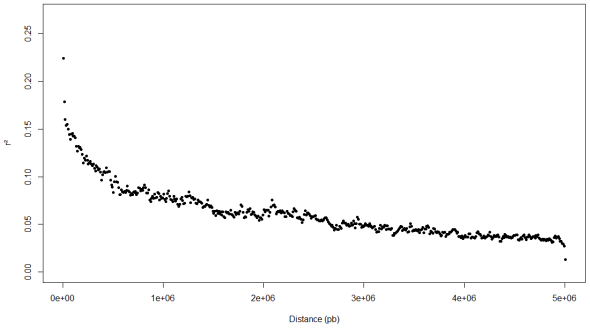
LG06



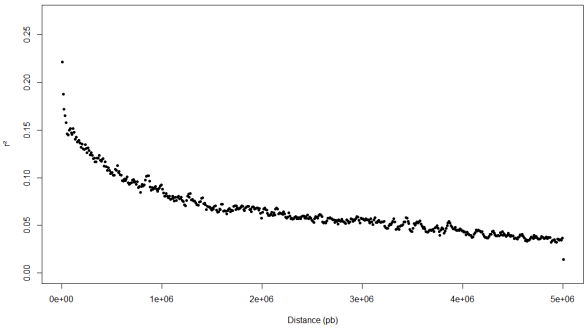
LG07



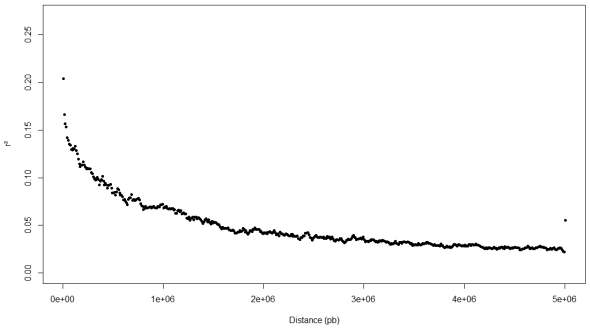
LG08



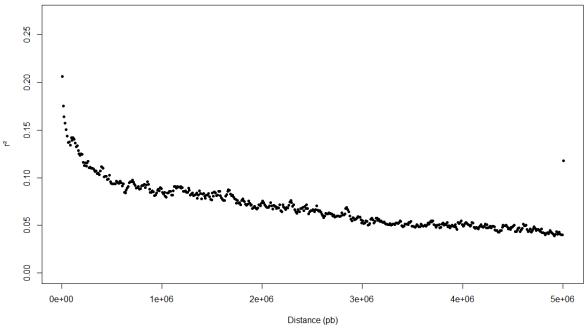
LG09



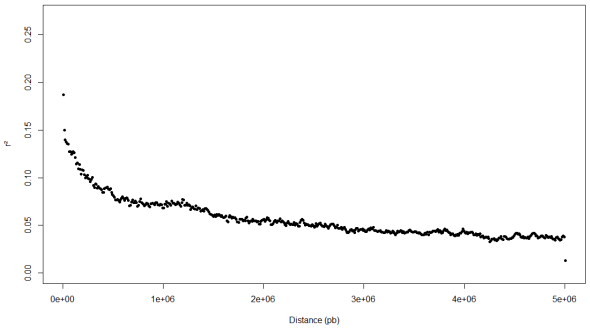
LG10



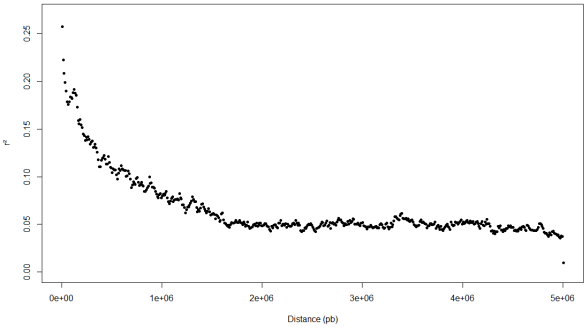
LG11



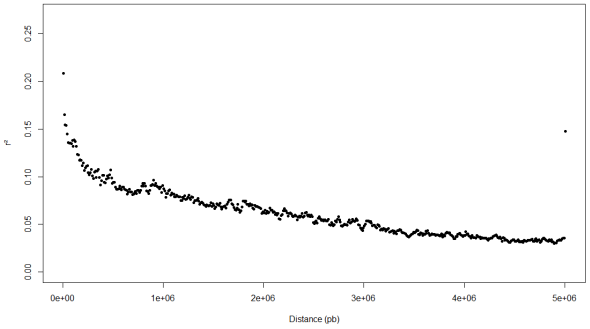
LG12



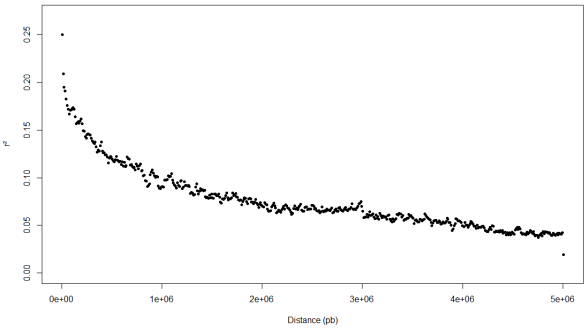
LG13



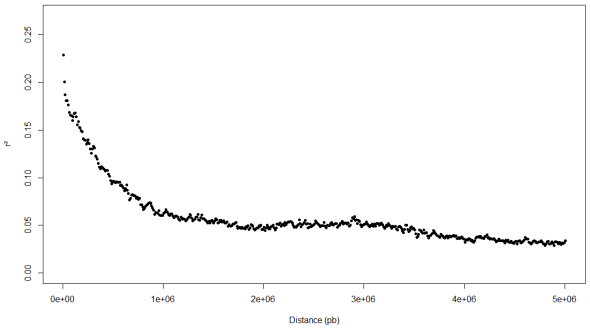
LG14



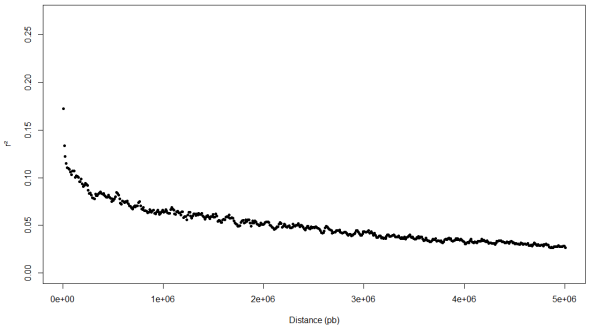
LG15



LG16

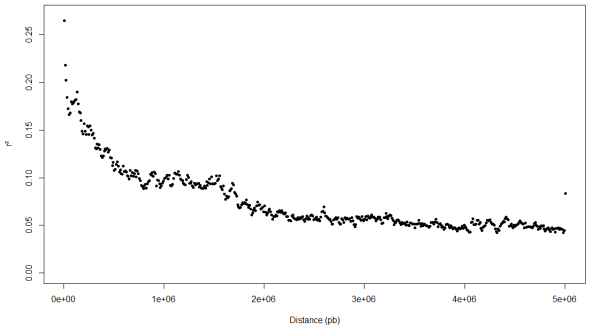


LG17

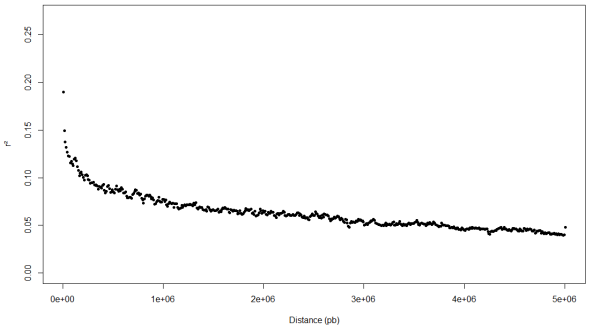


Annexe 6 : Courbes de décroissance du déséquilibre de liaison par groupe de liaison (exprimé par le paramètre r²), en fonction de la distance physique en utilisant les données de génotypage issus de la puce 480k SNPs sur la CC48

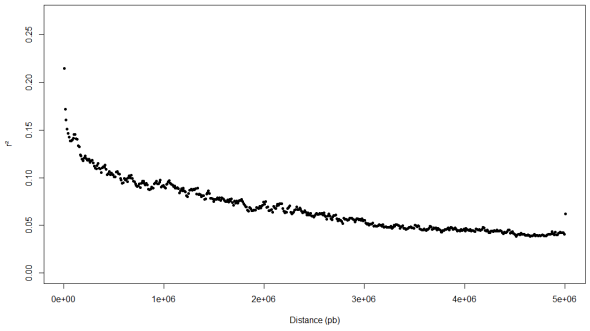
LG01



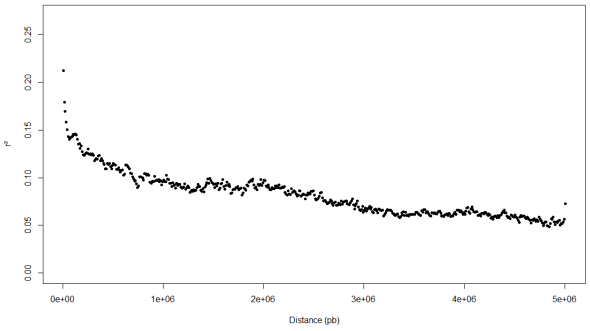
LG02



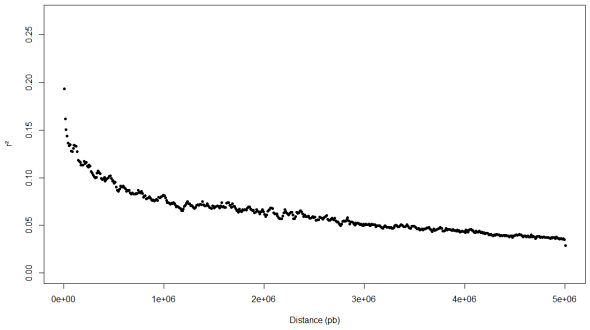
LG03



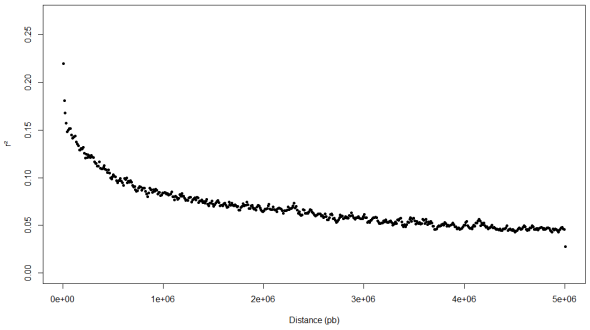
LG04



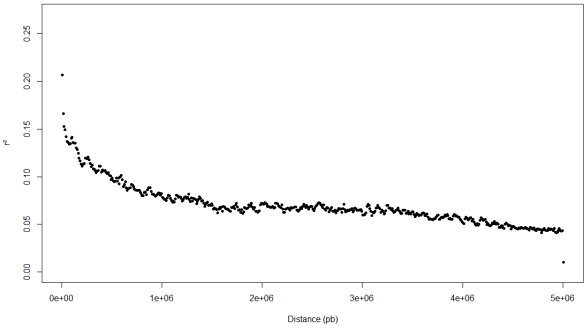
LG05



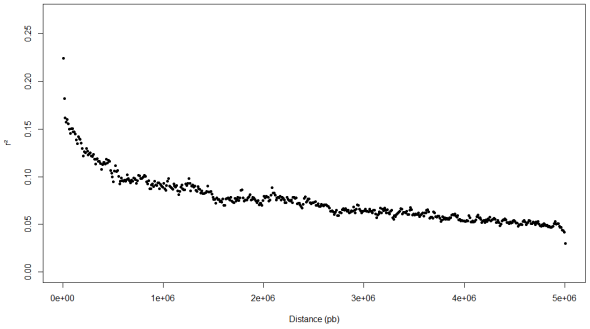
LG06



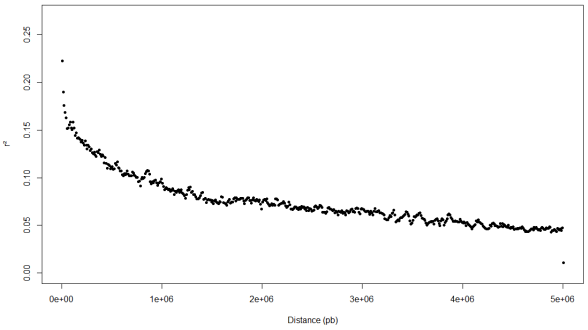
LG07



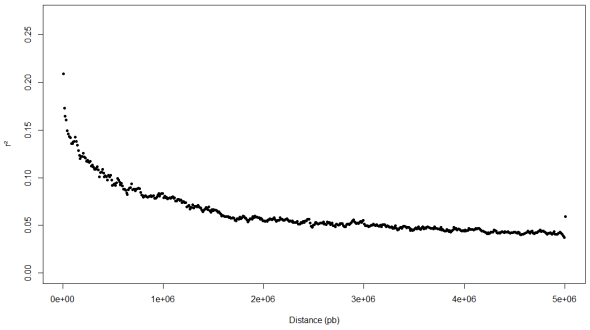
LG08



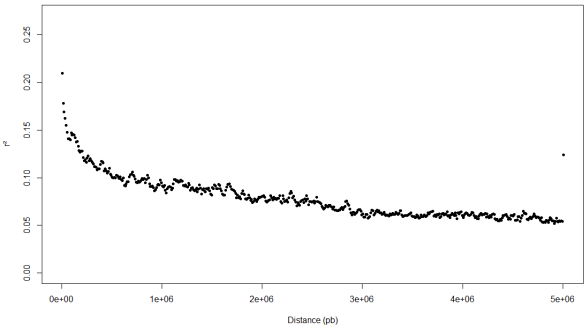
LG09



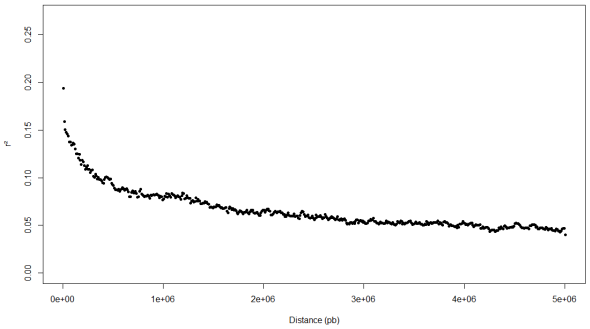
LG10



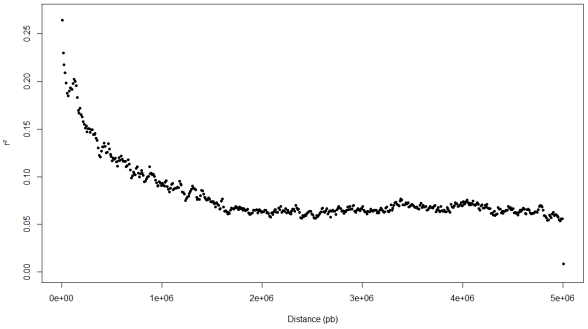
LG11



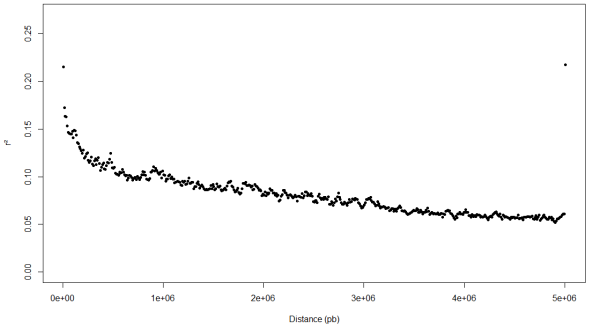
LG12



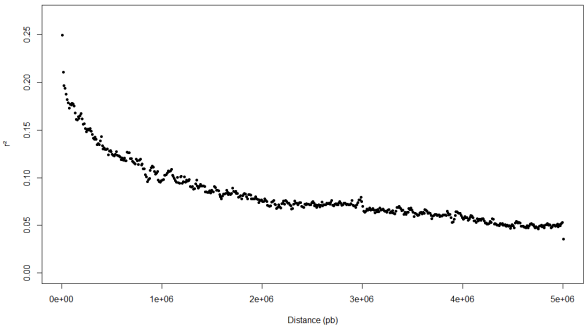
LG13



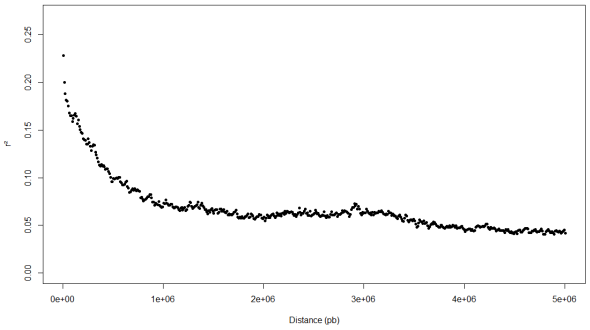
LG14



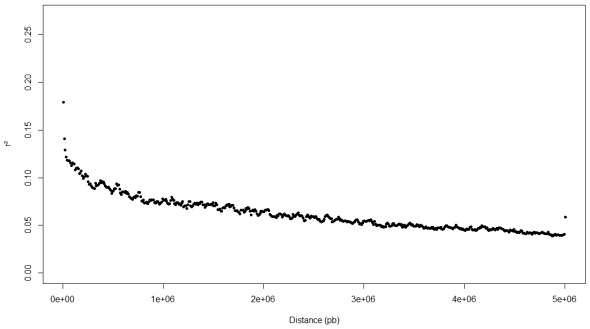
LG15



LG16

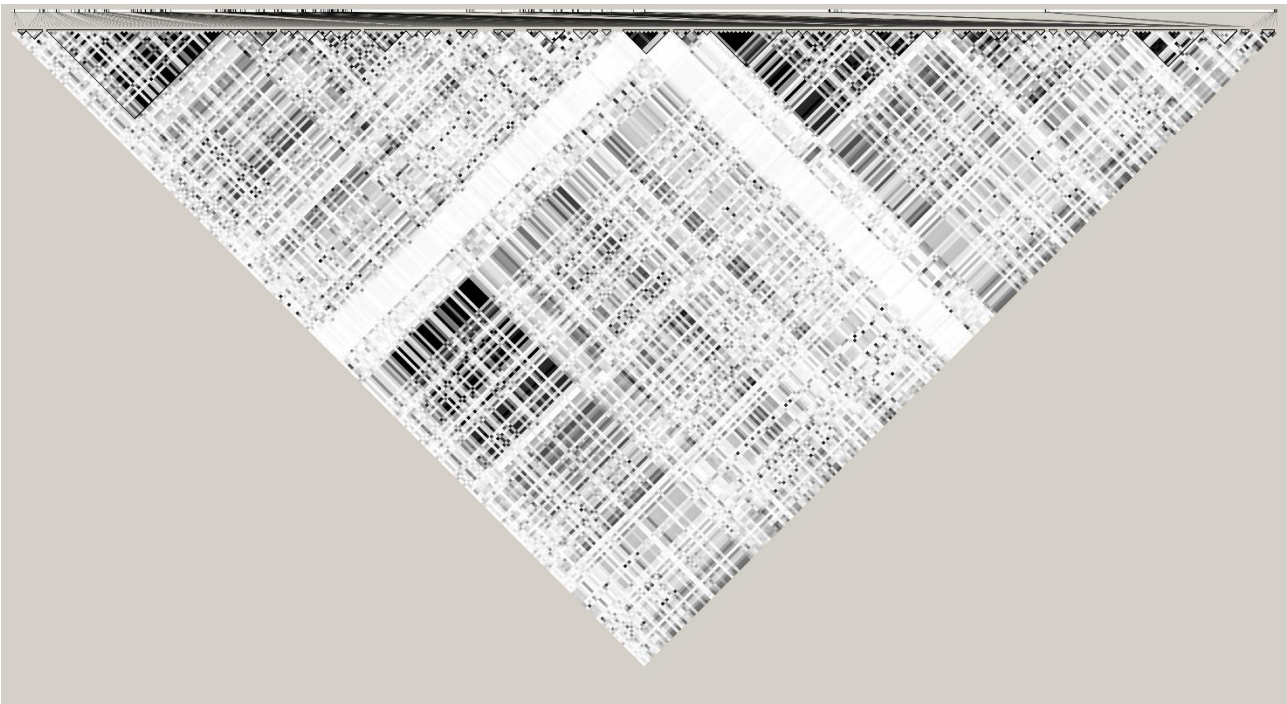


LG17

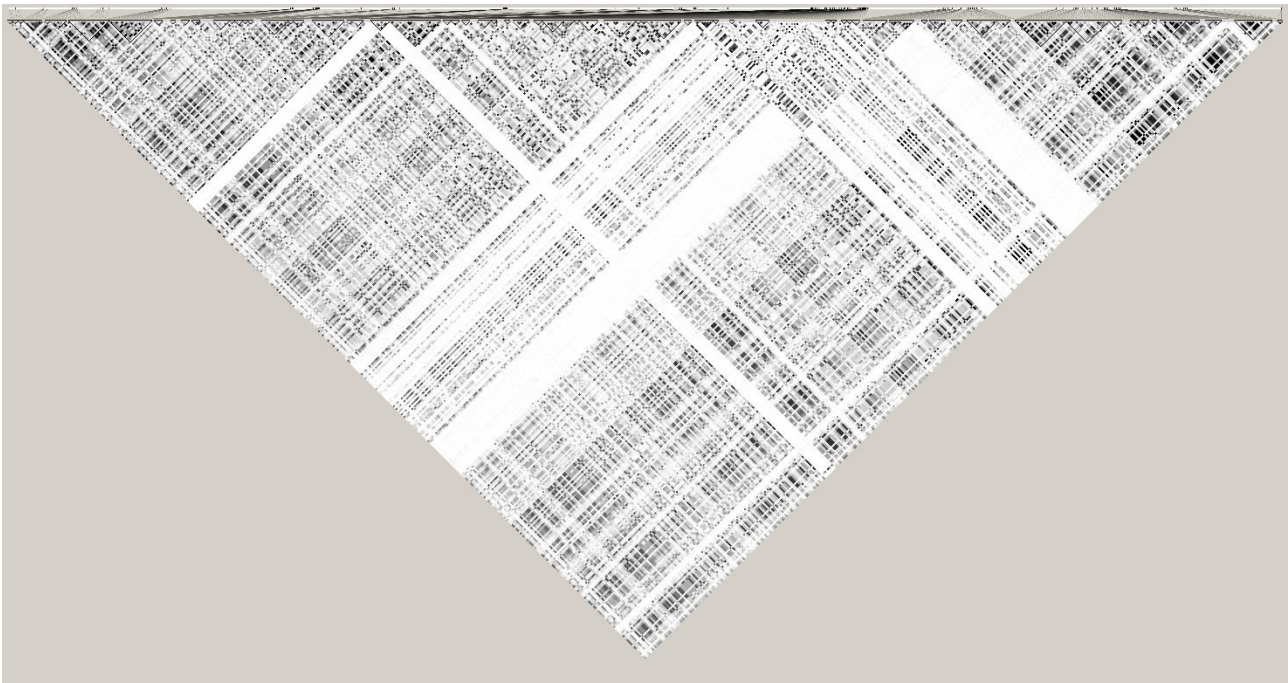


Annexe 7 : Représentation graphique des blocs de SNPs en déséquilibre de liaison dans plusieurs régions du génome du pommier en utilisant les marqueurs de la puce 480k SNPs

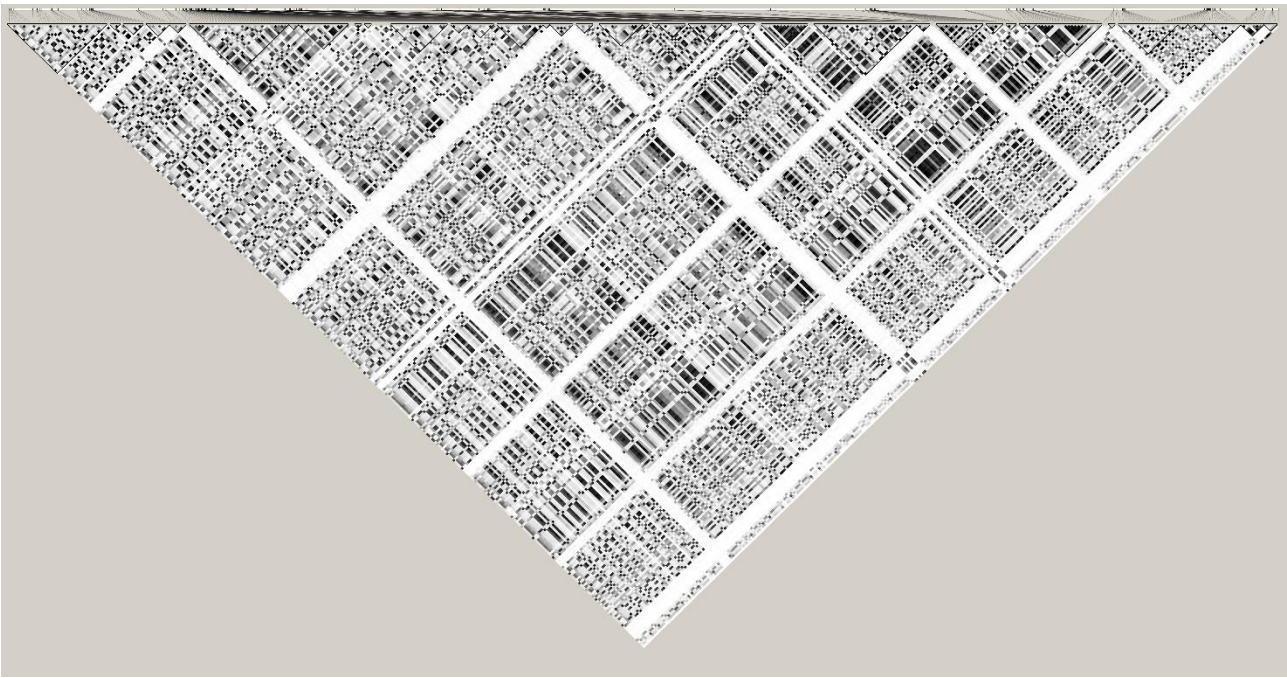
LG05



LG12



LG15



Annexe 8 : Tableau récapitulatif du nombre de fragments par individus et du nombre d'individus inclus dans les analyses sur les données de re-séquençage

Clone	Fragments	Clone	Fragments	Clone	Fragments
X0036	77	X1349	62	X6208	75
X0048	65	X1552	71	X6354	72
X0337	17	X1556	66	X6468	65
X0342	81	X1560	67	X6471	71
X0344	73	X1618	67	X6905	73
X0352	18	X1646	69	X6917	73
X0380	74	X1682	77	X6918	65
X0395	68	X1705	67	X6920	62
X0404	69	X1846	70	X7195	73
X0421	71	X1853	65	X7197	71
X0468	65	X1894	75	X7199	67
X0522	64	X1954	76	X7200	64
X0585	68	X1960	63	X7201	65
X0591	71	X1982	52	X7203	64
X0599	69	X2104	67	X7204	66
X0600	66	X2302	64	X7358	69
X0640	67	X2313	76	X7759	69
X0666	69	X2316	70	X8149	69
X0667	69	X2317	73	X8199	68
X0691	77	X2318	72	X8200	73
X0695	68	X2320	74	X8201	74
X0700	72	X2322	73	X8202	74
X0710	7	X2327	69	X8203	65
X0849	66	X2361	72	X8209	71
X0898	66	X2428	57	X8211	73
X0942	67	X2430	65	X8212	64
X0968	68	X2640	65	X8215	70
X1071	53	X2643	76	X8218	76
X1076	70	X2646	70	X8220	71
X1077	72	X2949	70	X8222	65
X1095	72	X2953	70	X8223	72
X1176	72	X2998	72	X8224	73
X1180	68	X3714	67	X8226	17
X1186	54	X4004	71	X8227	67
X1206	35	X4616	66	X8228	69
X1212	63	X4664	72	X8229	72
X1225	70	X4874	72	X8232	69
X1227	69	X4898	73	X8233	72
X1235	74	X4915	67	X8236	71
X1269	72	X4975	74	X8237	76
X1301	71	X6171	24	X8238	67
X1307	72	X6172	69	X8239	72
X1314	66	X6194	72	X8242	67

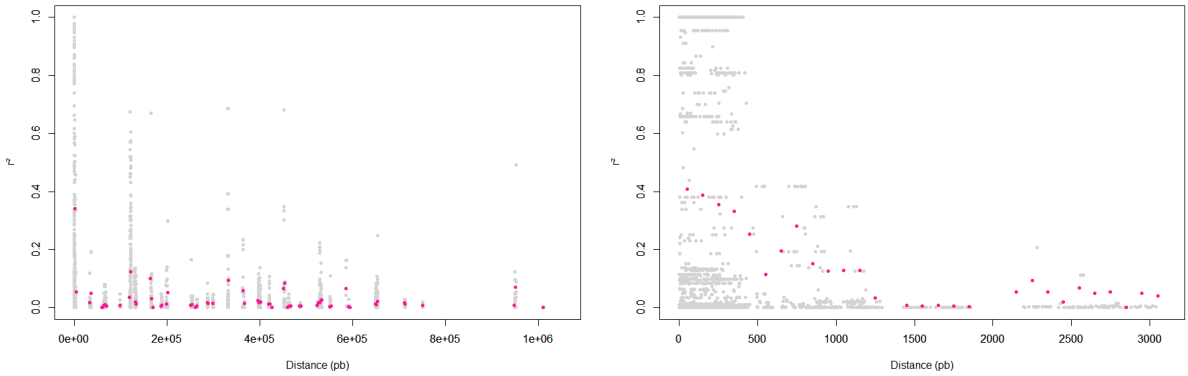
X1344	14	X6206	63	X8244	72
X8245	62	X8739	58	X9251	76
X8246	57	X8740	72	X9252	73
X8247	73	X8742	69	X9256	69
X8249	71	X8743	69	X9257	78
X8250	71	X8746	69	X9259	68
X8252	67	X8749	69	X9260	70
X8256	73	X8750	75	X9266	74
X8380	73	X8751	54	X9267	68
X8381	70	X8933	74	X9389	72
X8383	73	X8934	69	X9391	73
X8384	65	X8937	72	X9394	72
X8386	67	X8939	71	X9398	67
X8389	69	X8972	61	X9407	74
X8392	68	X9078	66	X9408	71
X8396	71	X9080	70	X9409	74
X8398	59	X9089	74	X9411	74
X8404	70	X9090	74	X9416	69
X8405	70	X9097	70	X9420	71
X8407	73	X9100	67	X9421	68
X8411	67	X9105	62	X9429	66
X8412	67	X9115	72	X9433	70
X8414	67	X9116	72	X9436	66
X8415	65	X9118	76	X9437	73
X8416	71	X9122	75	X8737	68
X8607	74	X9124	74	X8738	71
X8691	71	X9128	67	X9246	67
X8692	75	X9130	60	X9250	55
X8694	73	X9134	65		
X8697	75	X9135	70		
X8698	61	X9148	72		
X8699	67	X9151	67		
X8702	49	X9152	70		
X8703	77	X9166	71		
X8705	73	X9167	76		
X8706	74	X9171	57		
X8710	65	X9176	71		
X8713	70	X9177	70		
X8715	59	X9179	66		
X8717	74	X9185	72		
X8718	51	X9186	71		
X8719	66	X9190	68		
X8723	74	X9191	67		
X8724	72	X9195	67		
X8726	75	X9196	73		
X8734	73	X9198	74		
X8735	63	X9202	74		

Fragment	Clone	Fragment	Clone
MDP0000322294	0	MDP0000868785_A	228
MDP0000179513_A	118	MDP0000868782_B	134
MDP0000179513_B	187	MDP0000868782_A	229
MDP0000232836	0	MDP0000945841_B	243
MDP0000321036	0	MDP0000945841_A	110
MDP0000238940_A	244	MDP0000467552_A	232
MDP0000238940_B	150	MDP0000467552_B	234
MDP0000930114_A	111	MDP0000204278_A	105
MDP0000930114_B	216	MDP0000204278_B	242
MDP0000855736_A	143	MDP0000262892_B	243
MDP0000855736_B	241	MDP0000262892_A	224
MDP0000662922_A	123	MDP0000131641	0
MDP0000662922_B	211	MDP0000593517_A	239
MDP0000166405_A	158	MDP0000593517_B	166
MDP0000166405_B	234	MDP0000290079	0
MDP0000133269	0	MDP0000307848_A	80
MDP0000162814_B	206	MDP0000307848_B	240
MDP0000162814_A	220	MDP0000121185_B	131
MDP0000208326_A	246	MDP0000121185_A	164
MDP0000208326_B	115	MDP0000172931_A	174
MDP0000206447	0	MDP0000172931_B	147
MDP0000148339	0	MDP0000509183_A	242
MDP0000305238_A	185	MDP0000509183_B	232
MDP0000305238_B	142	MDP0000275915_B	244
MDP0000252859_A	179	MDP0000275915_A	137
MDP0000252859_B	241	MDP0000148169_B	242
MDP0000160327_B	167	MDP0000148169_A	242
MDP0000160327_A	221	MDP0000255648_B	219
MDP0000300483_A	180	MDP0000255648_A	242
MDP0000300483_B	144	MDP0000244520_A	62
MDP0000949487_A	243	MDP0000244520_B	229
MDP0000949487_B	88	MDP0000218438_B	238
MDP0000561267	0	MDP0000218438_A	152
MDP0000561265_B	243	MDP0000233031_A	238
MDP0000561265_A	146	MDP0000233031_B	241
MDP0000189315_A	165	MDP0000288817_A	238
MDP0000189315_B	211	MDP0000288817_B	239
MDP0000129923_A	170	MDP0000224380_A	147
MDP0000129923_B	234	MDP0000224380_B	237
MDP0000937158_A	212	MDP0000196196_A	51
MDP0000937158_B	218	MDP0000196196_B	222
MDP0000791555_B	196	MDP0000547261	0
MDP0000791555_A	244	MDP0000269326_B	232
MDP0000275088_A	235	MDP0000269326_A	240
MDP0000275088_B	216	MDP0000304854_A	237
MDP0000383616_A	129	MDP0000304854_B	244
MDP0000383616_B	175	MDP0000868785_B	228

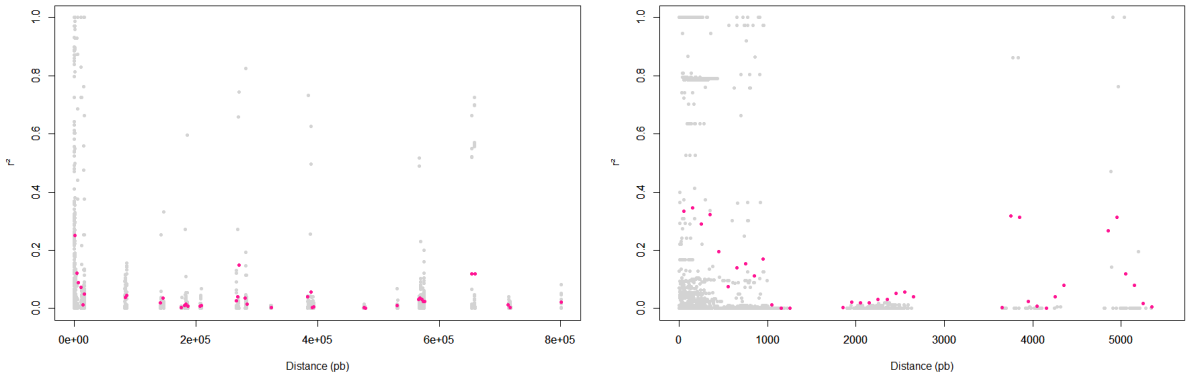
Annexe 9 : Courbes de décroissance du déséquilibre de liaison inter- (gauche) et intra-génique (droite) par région génomique (exprimé par le paramètre r^2), en fonction de la distance physique en utilisant les données de re-séquençage sur la CC278

les points gris représentent la totalité des comparaison entre toutes les paires de marqueurs ; les points roses représentent les r^2 moyens par incrément de 1 250 pb

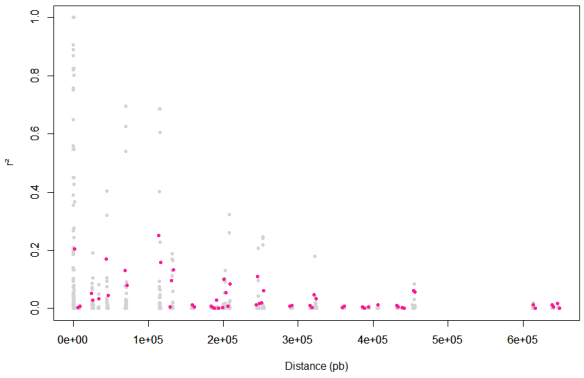
LG01

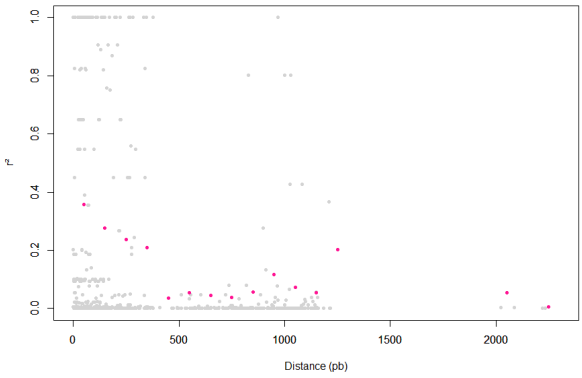


LG03

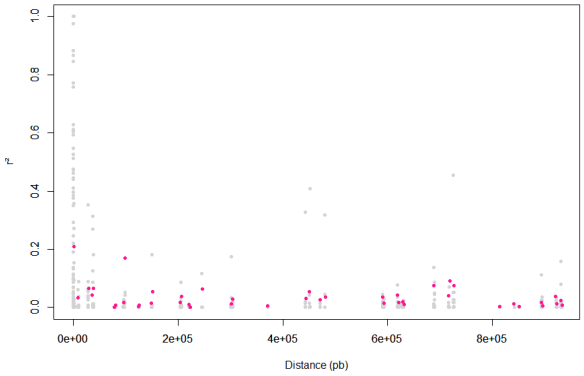


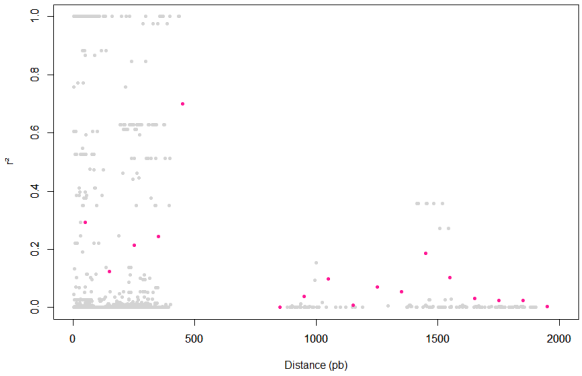
LG07



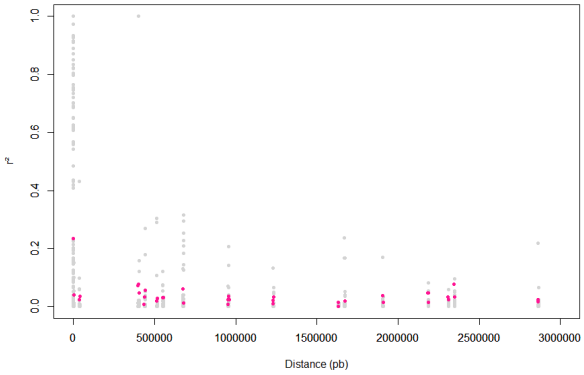


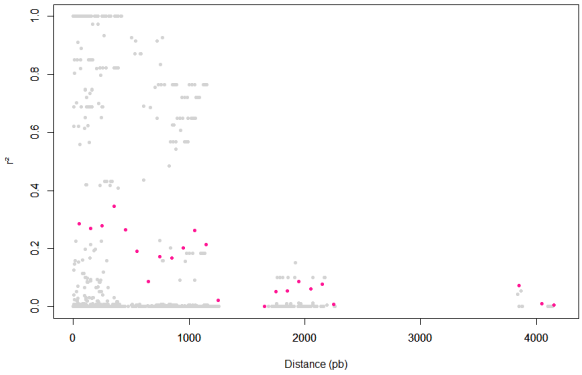
LG10



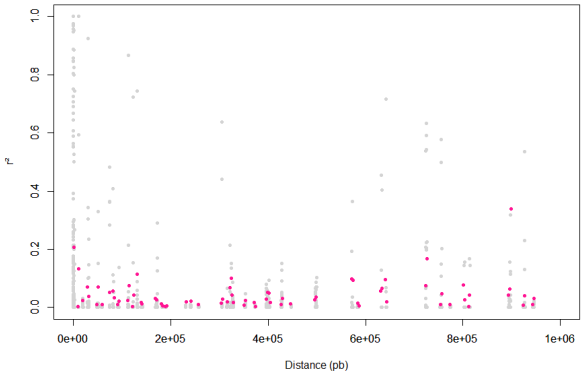


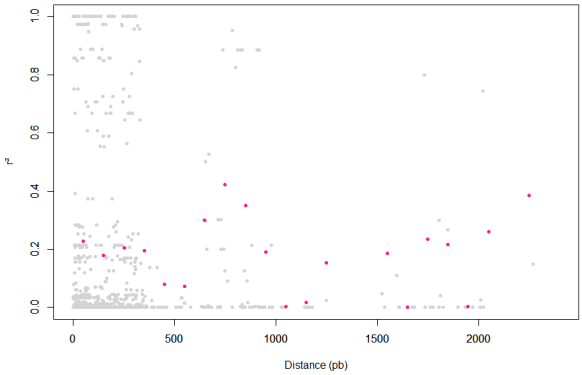
LG11





LG17





Annexe 10 : Représentation graphique des blocs de SNPs en déséquilibre de liaison dans cinq régions génomiques à partir des données de re-séquençage sur la CC278

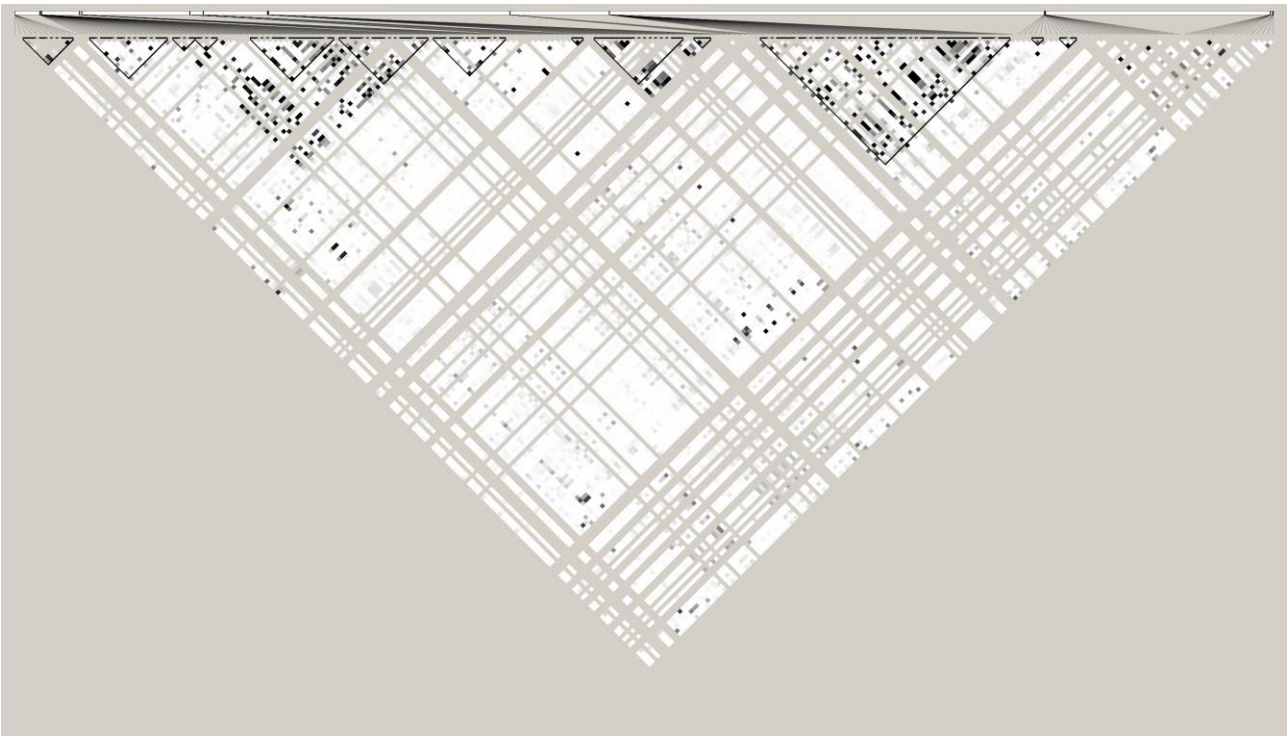
LG03

LG07

LG10

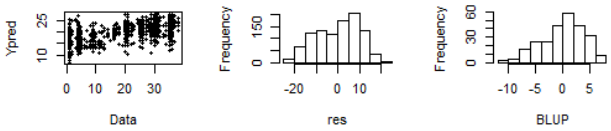
LG11

LG17

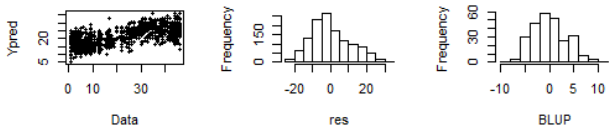


Annexe 11 : Graphiques des projections de la valeur génotypique estimée avec le BLUP contre la valeur réelle des phénotypes (gauche), de la distribution des résidus du modèle (milieu) et de la distribution des BLUP (droite) pour les données des tests de résistances

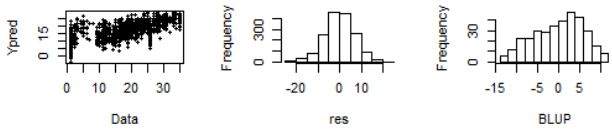
Résistance vis-à-vis de la souche EU-B04



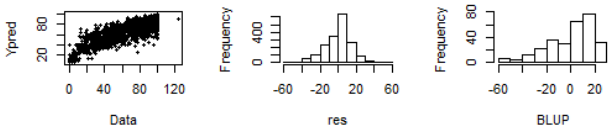
Résistance vis-à-vis de la souche 104



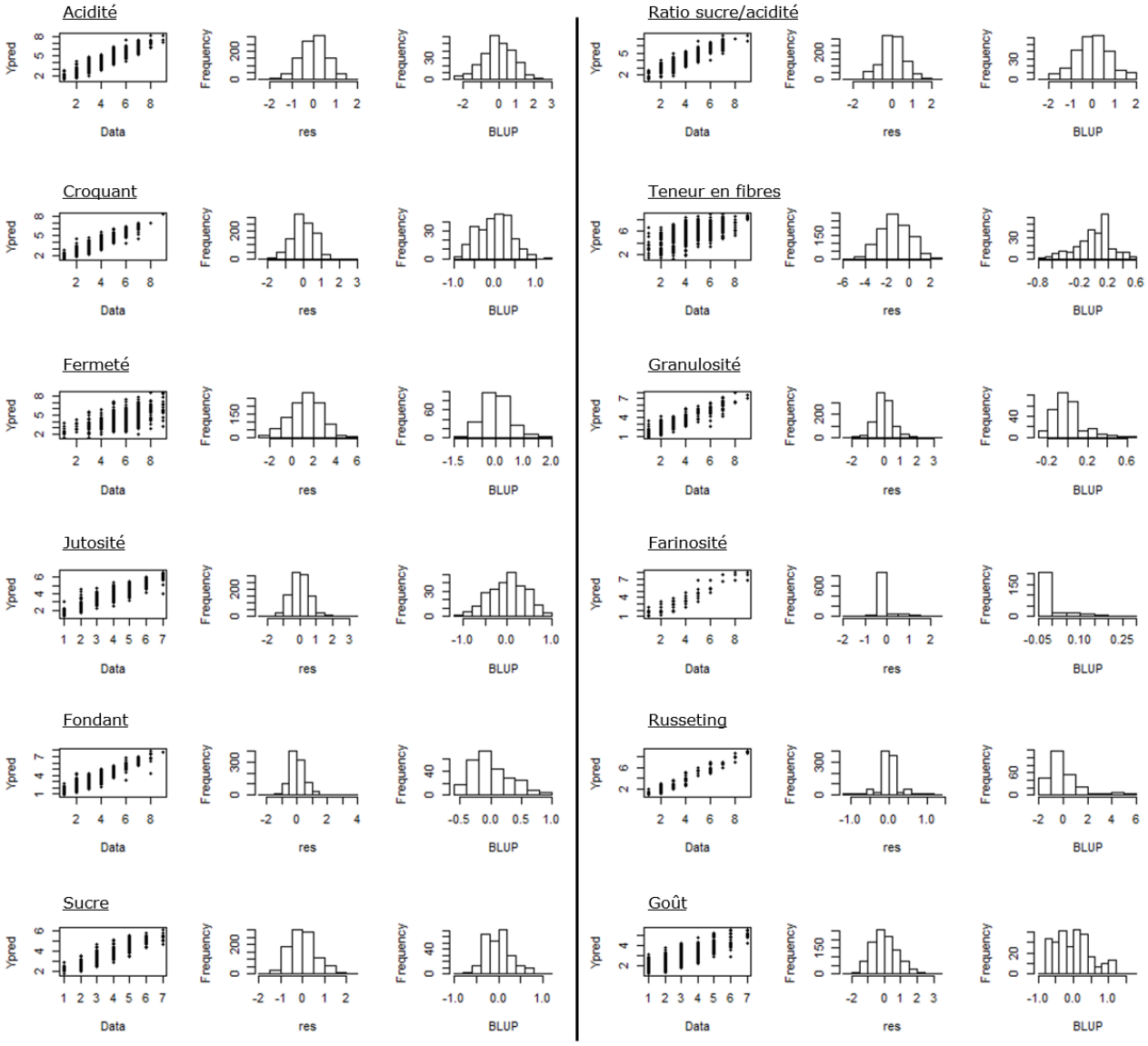
Résistance vis-à-vis du mélange de souches



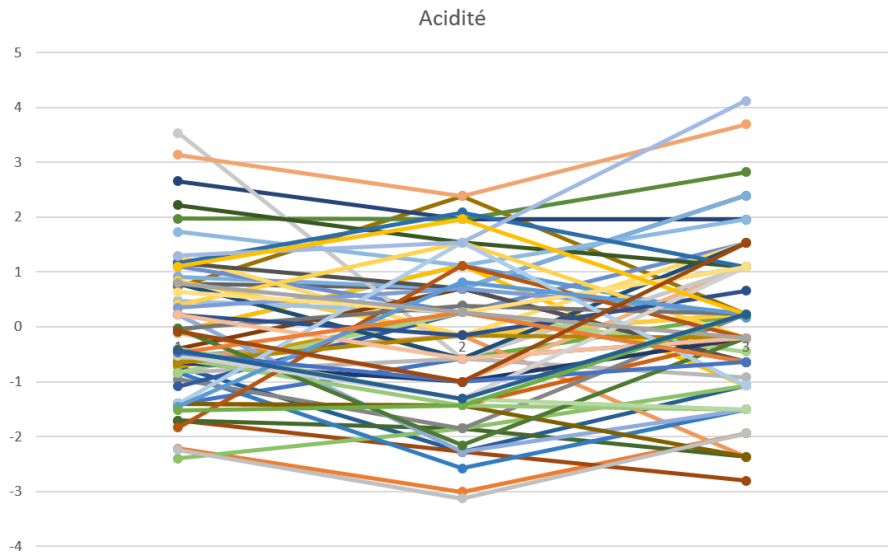
Résistance vis-à-vis d'*Erwinia amylovora*

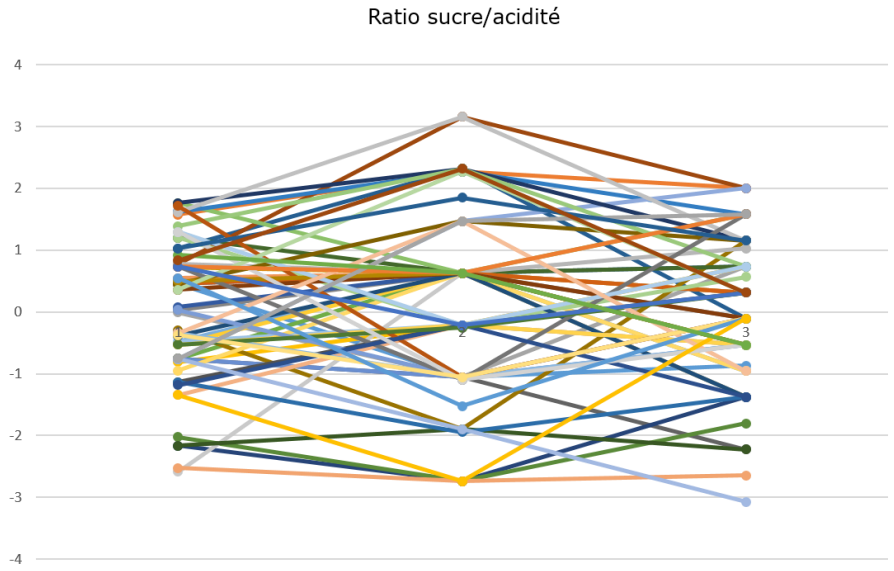


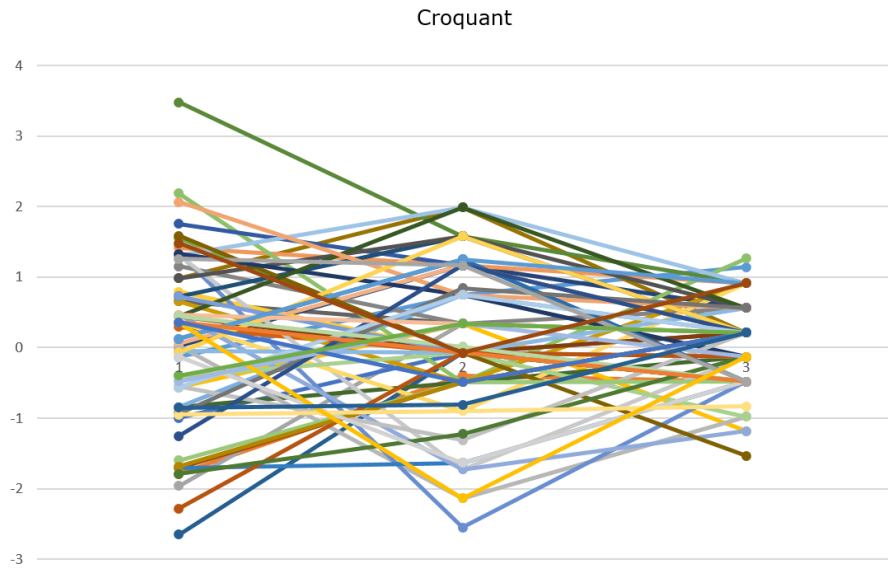
Annexe 12 : Graphiques des projections de la valeur génotypique estimée avec le BLUP contre la valeur réelle des phénotypes (gauche), de la distribution des résidus du modèle (milieu) et de la distribution des BLUP (droite) pour les données de qualité du fruit

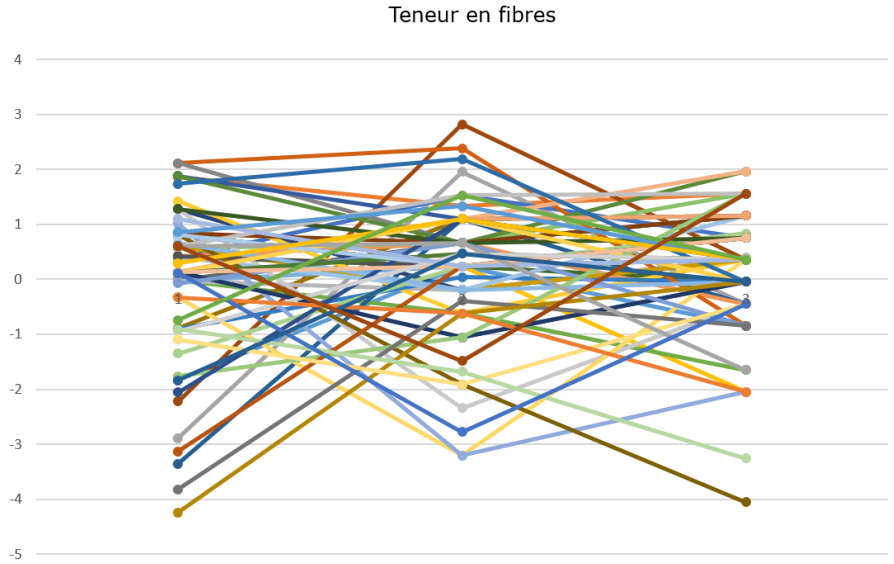


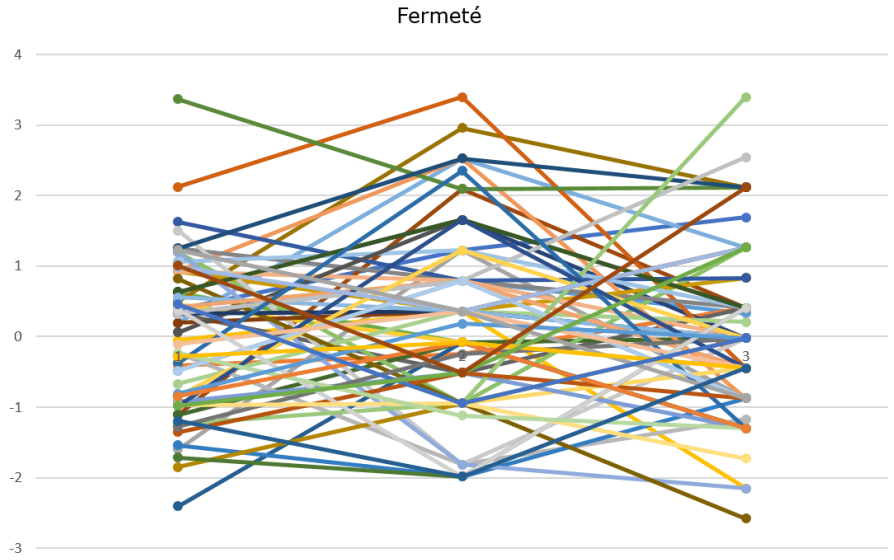
Annexe 13 : Visualisation de l'interaction entre l'effet génotype et l'effet année grâce aux BLUP des données phénotypiques des caractères de qualité du fruit pour les individus commun aux trois années de notation

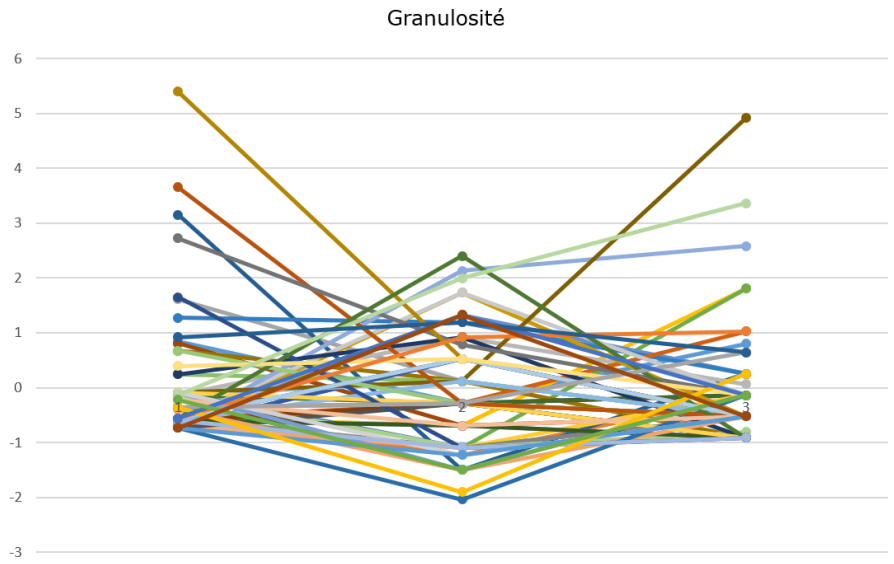


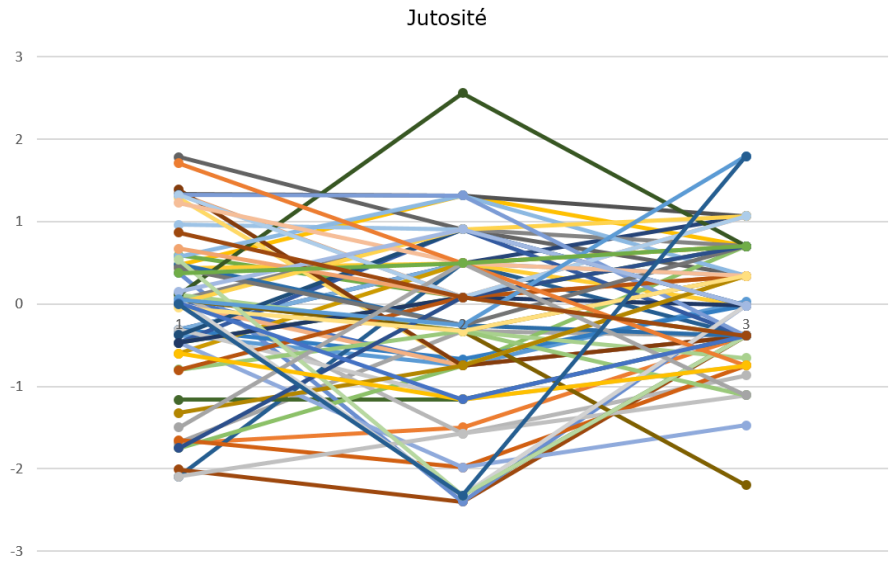


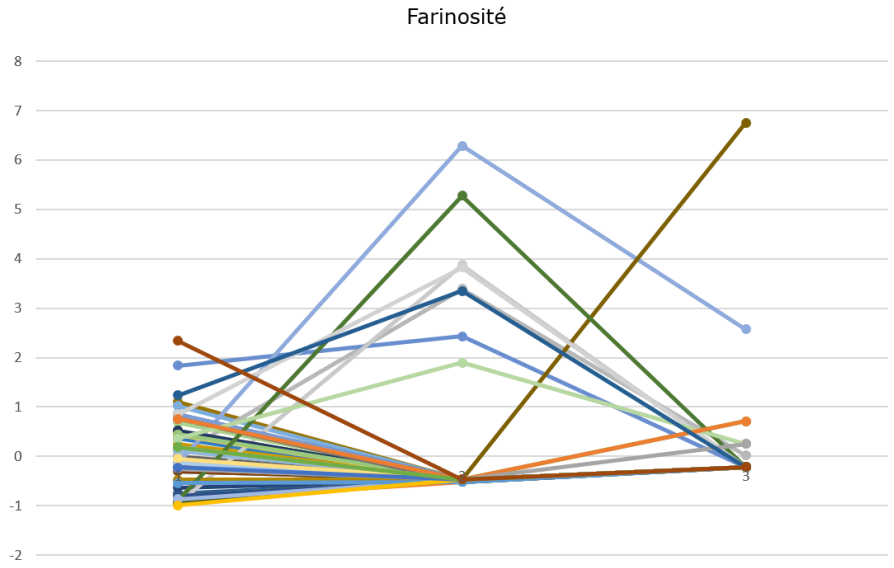

Croquant
4
3
2
1
0
-1
-2
-3
1
2
3

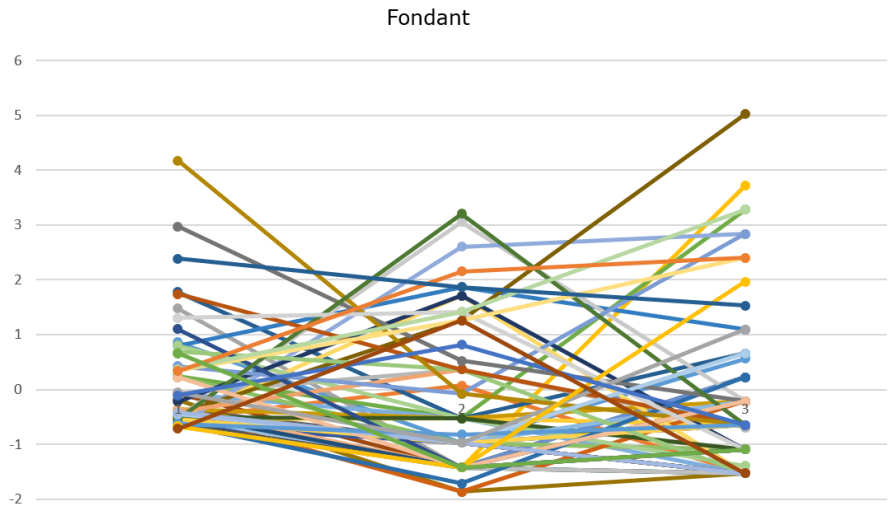

Teneur en fibres
4
3
2
1
0
-1
-2
-3
-4
-5
1
2
3

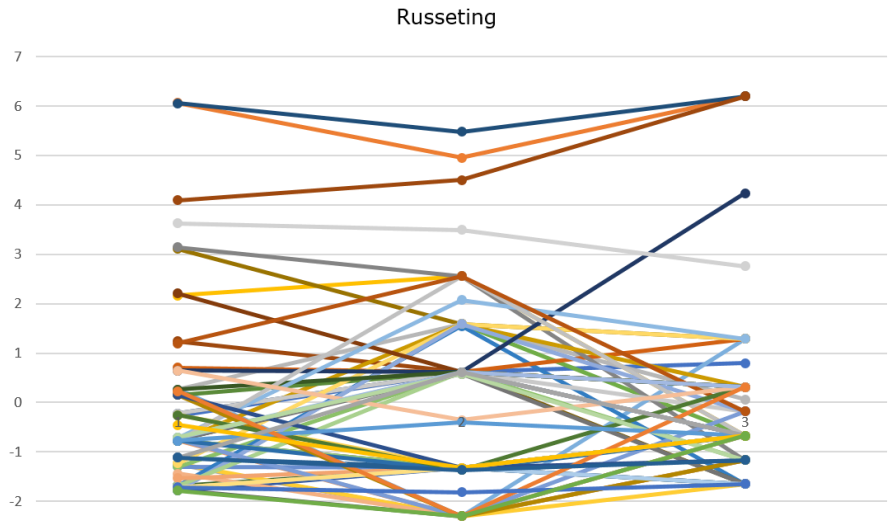

Fermeté
4
3
2
1
0
-1
-2
-3
1
2
3

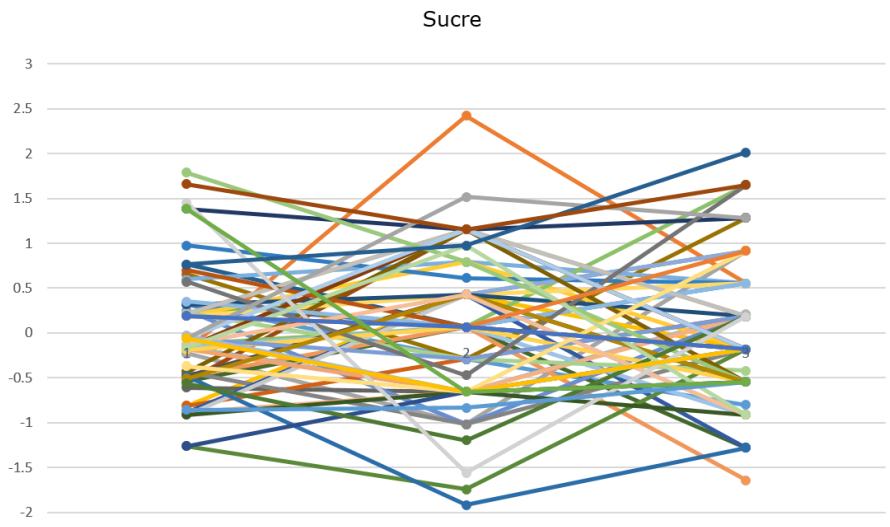

Granulosité
6
5
4
3
2
1
0
-1
-2
-3
1
2
3

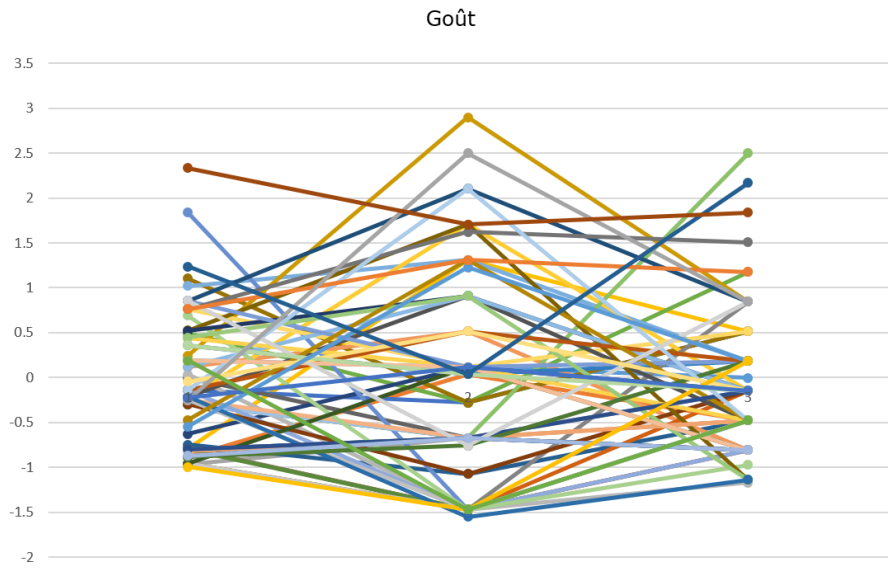

Jutosité
3
2
1
0
-1
-2
-3
1
2
3


Farinosité
8
7
6
5
4
3
2
1
0
-1
-2
3


Fondant
6
5
4
3
2
1
0
-1
-2
1
2
3

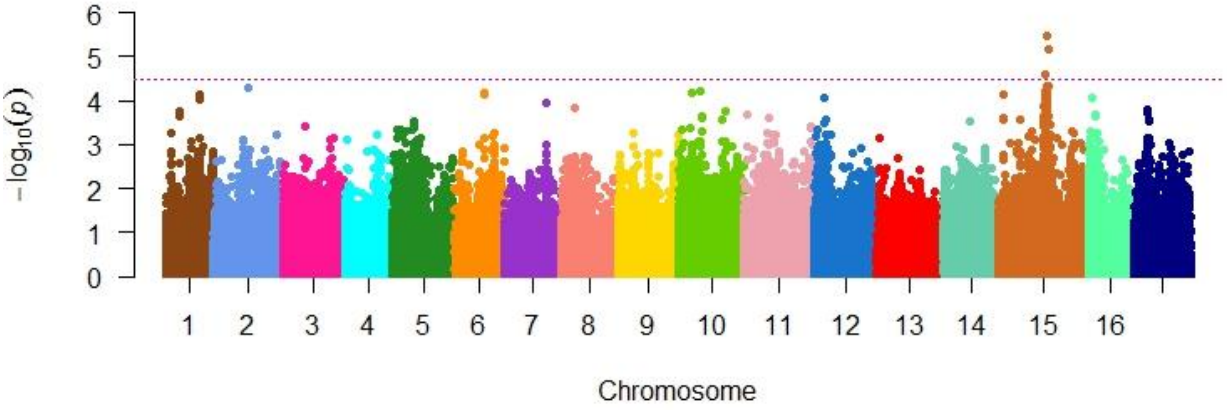

Russeting
7
6
5
4
3
2
1
0
-1
-2
1
2
3

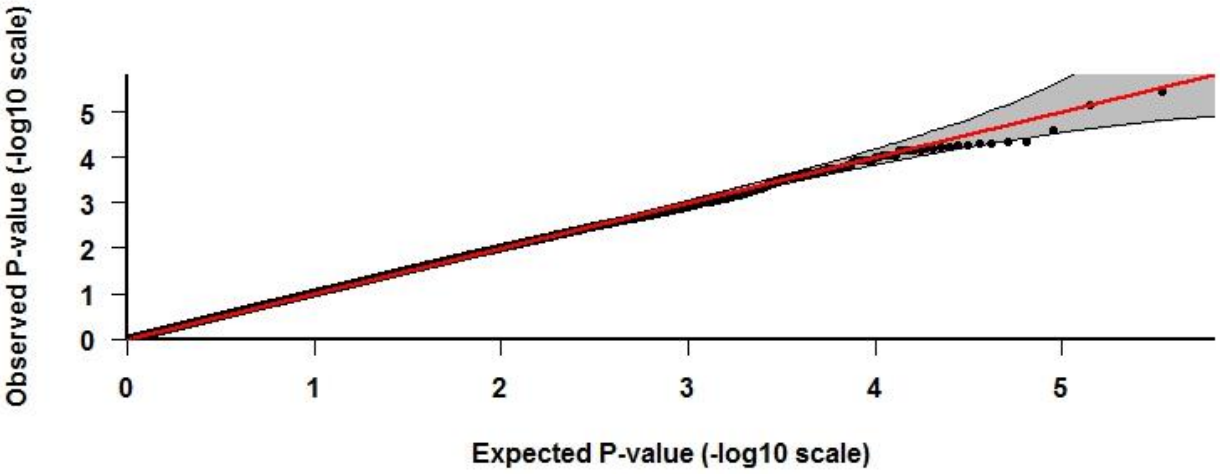

Sucre
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3


Goût
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3

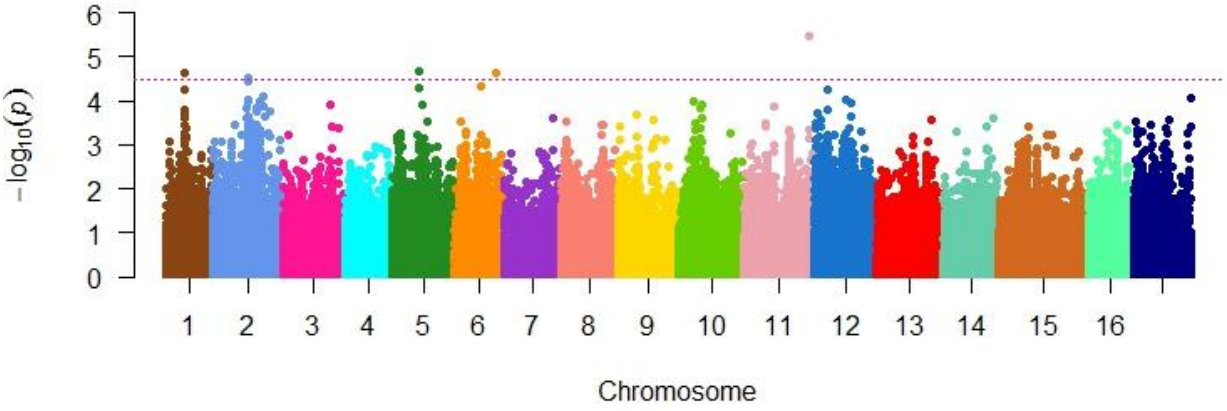
Annexe 14 : Manhattan plots et QQ plots des résultats des analyses d'association réalisées sur les données de résistance à la tavelure en corrigeant de la structure et de l'apparentement dans la CC278 avec les données de génotypage issues de la puce 480k SNPs

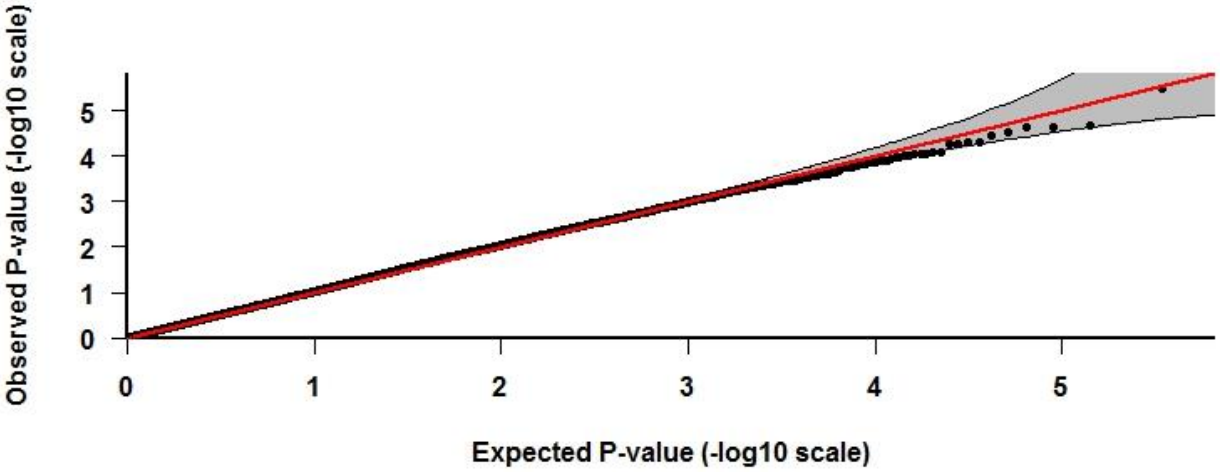
<u>Résistance vis-à-vis de la souche EU-B04</u>



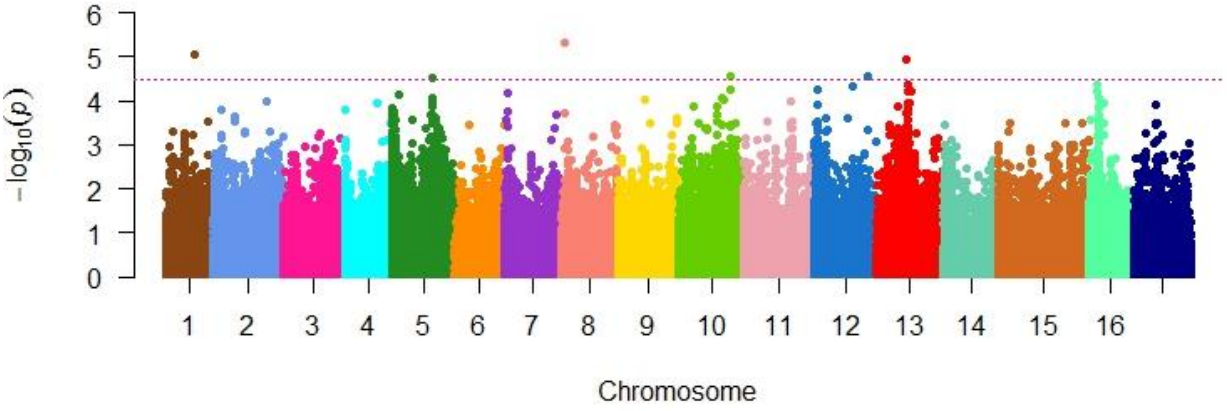


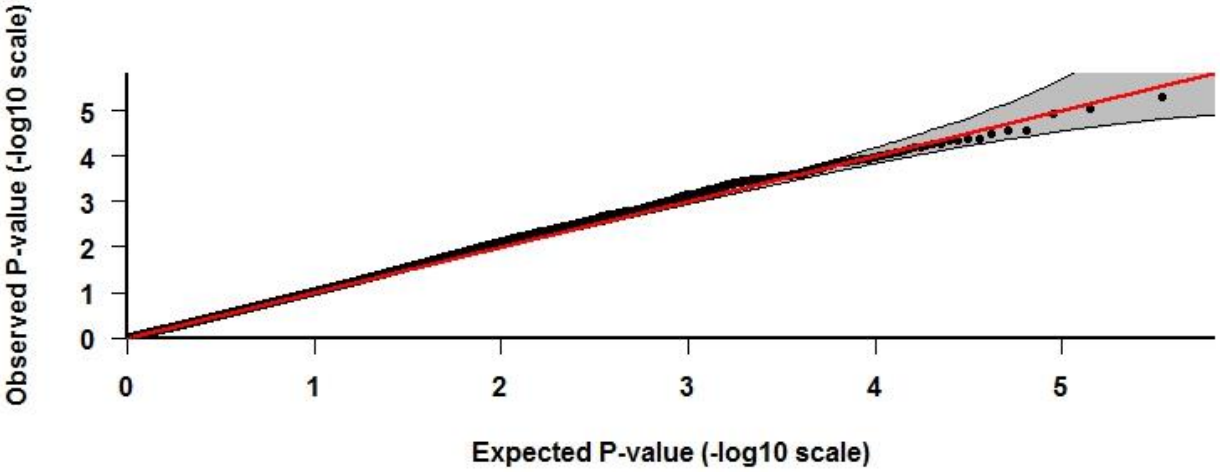
Résistance vis-à-vis de la souche 104





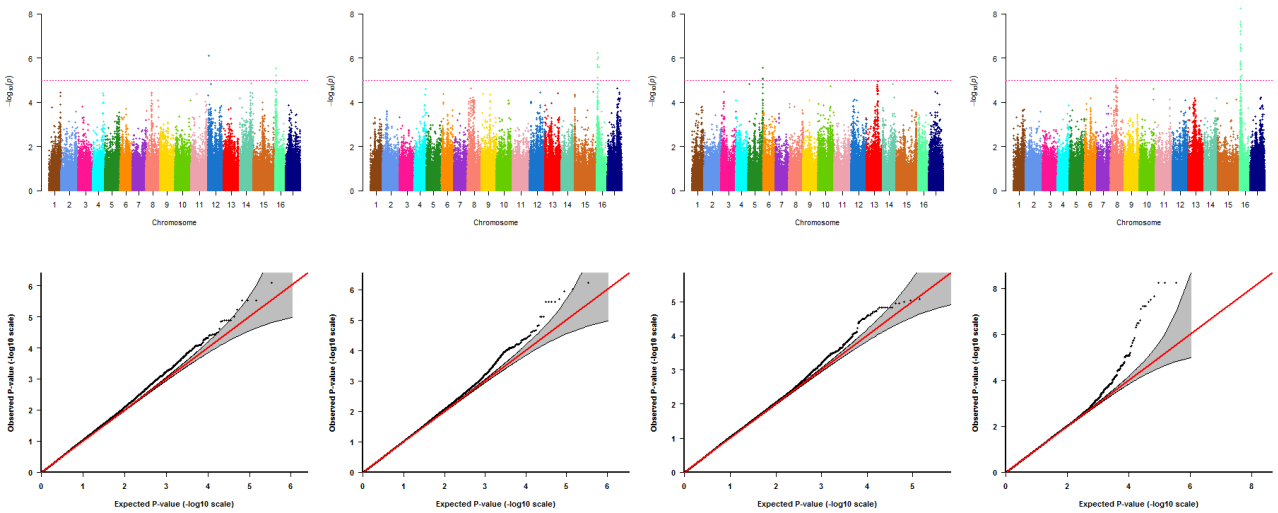
Résistance vis-à-vis du mélange de souches



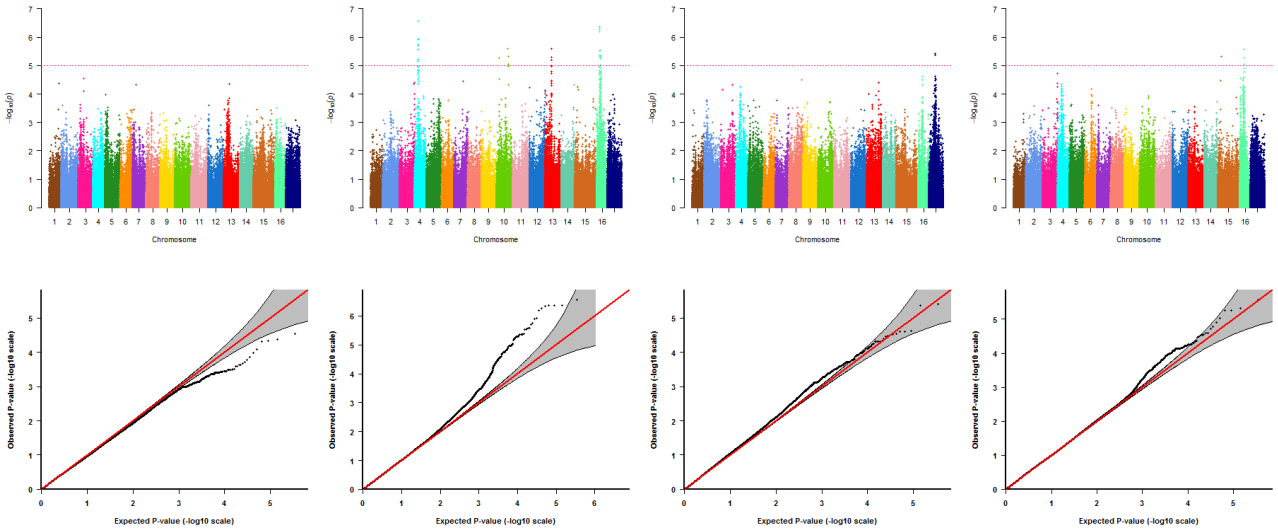


Annexe 15 : Manhattan plots et QQ plots des résultats des analyses d'association réalisées sur les données de qualité du fruit en corrigeant de la structure et de l'apparentement dans la CC278 avec les données de génotypage issues de la puce 480k SNPs

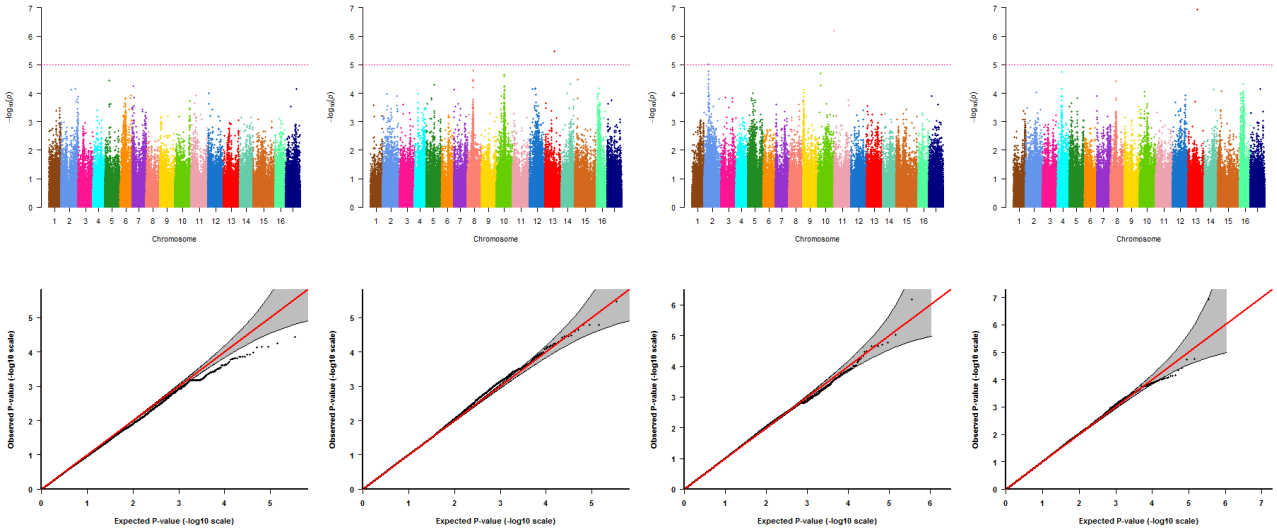
Ratio sucre/acidité



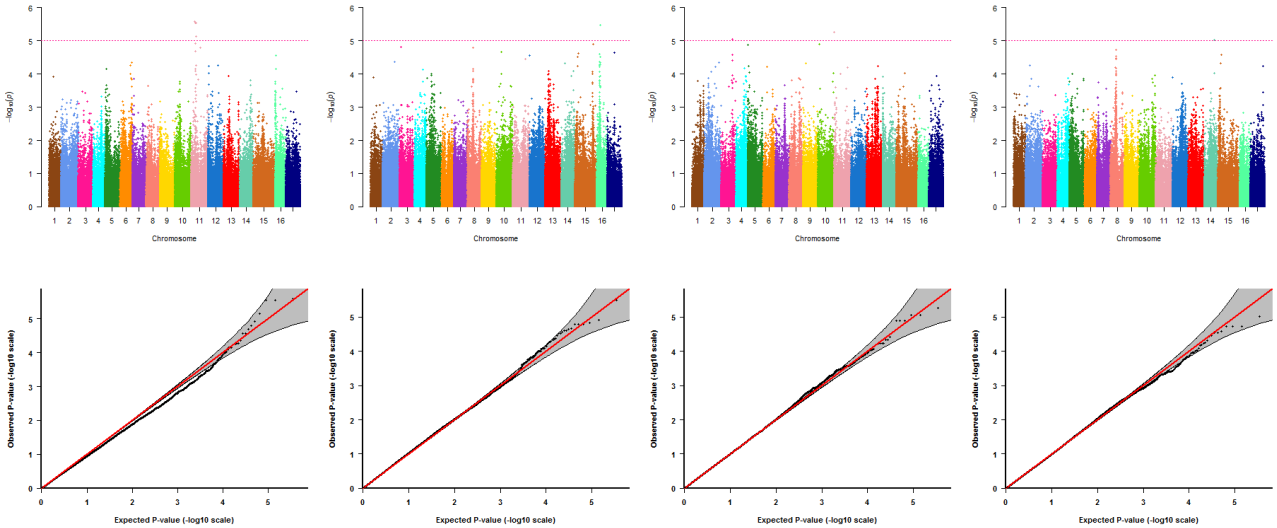
Croquant



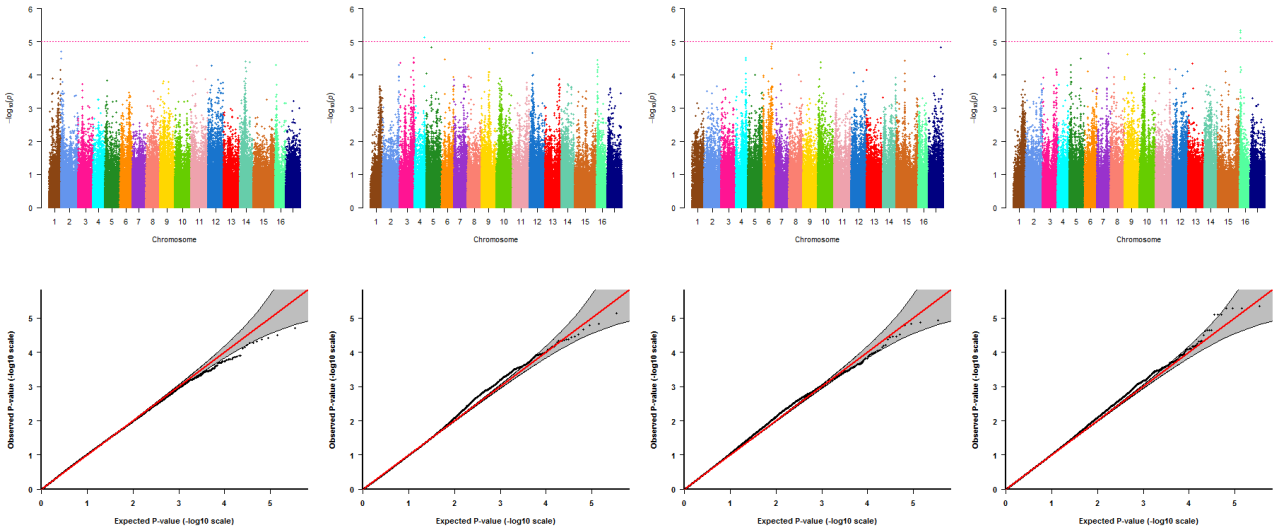
Teneur en fibres



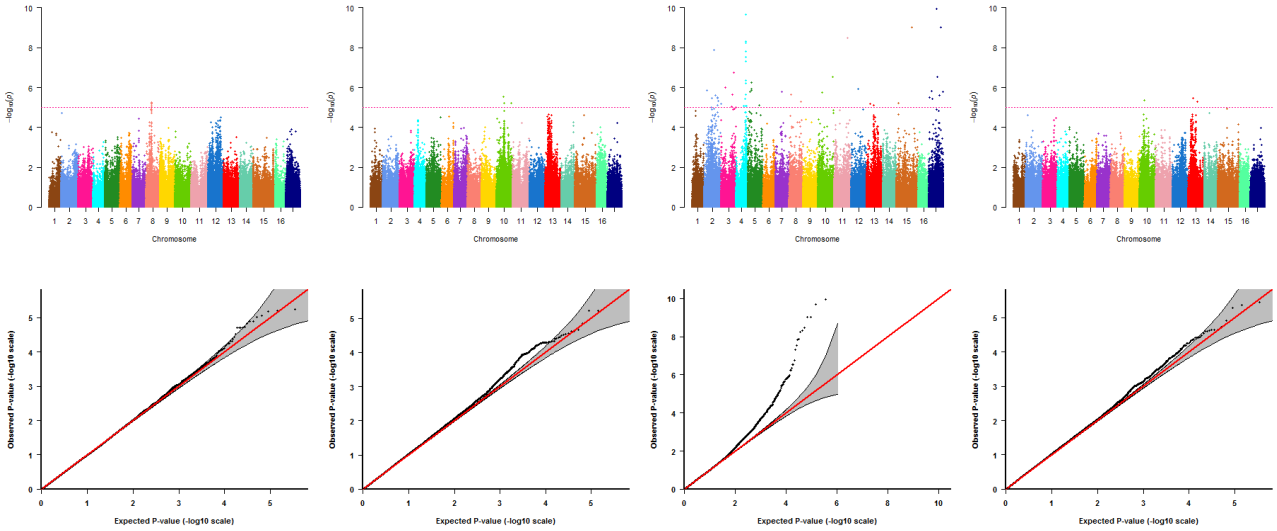
Granulosité



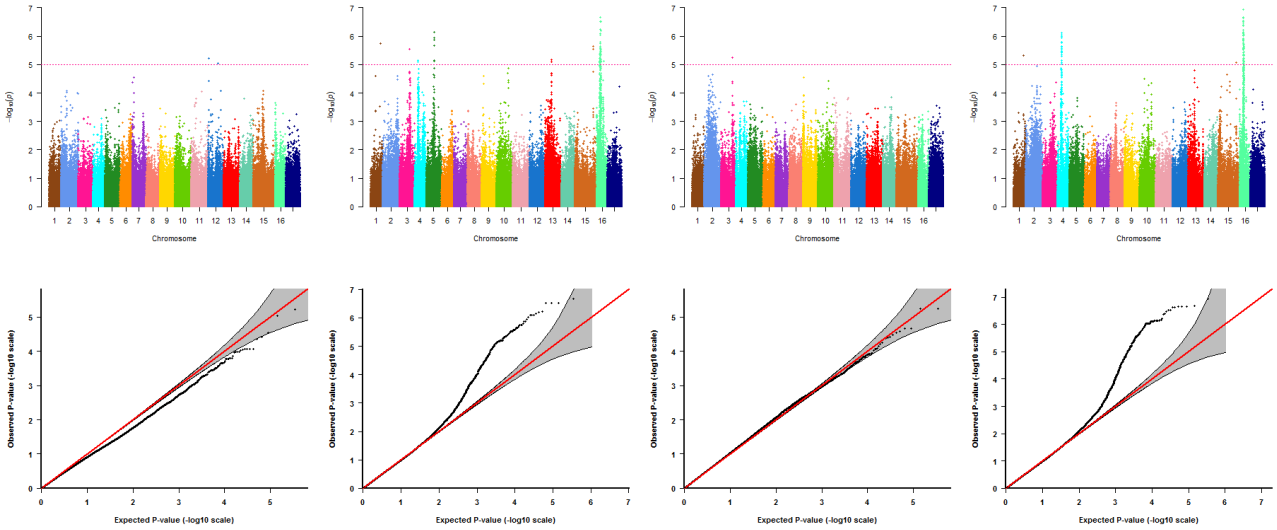
Jutosité



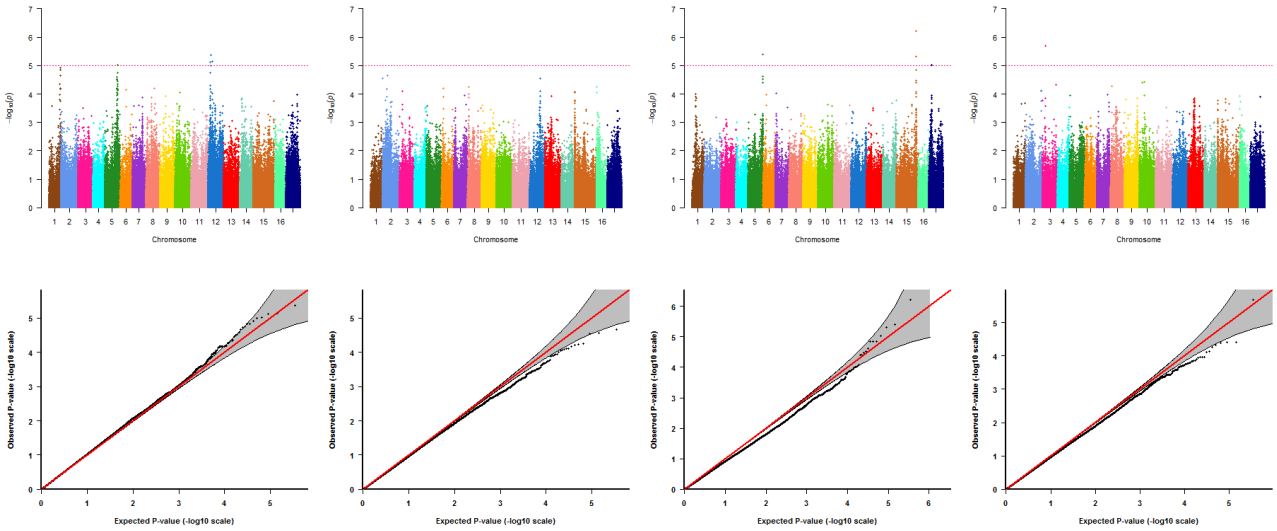
Farinosité



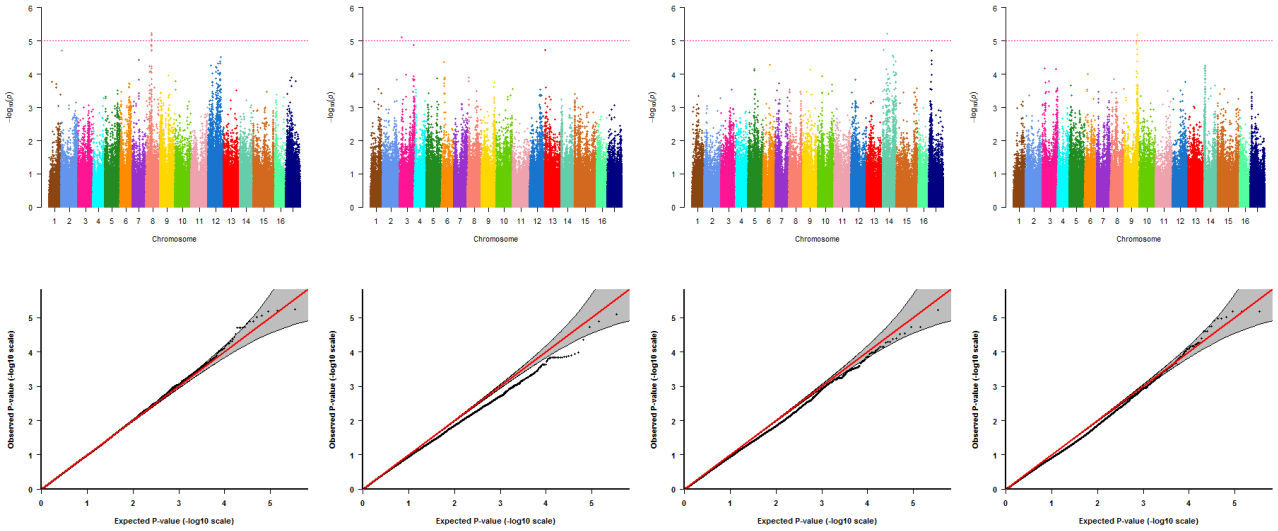
Fondant



Sucre



Goût


Chromosome
Chromosome
Chromosome
Chromosome
Observed P-value (-log10 scale)
Expected P-value (-log10 scale)
Observed P-value (-log10 scale)
Expected P-value (-log10 scale)
Observed P-value (-log10 scale)
Expected P-value (-log10 scale)
Observed P-value (-log10 scale)
Expected P-value (-log10 scale)

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

université angers

Thèse de Doctorat

Diane LEFORESTIER

Localisation de régions du génome du pommier contrôlant la variation de caractères de qualité du fruit et de résistance aux maladies : signatures de sélection et génétique d'association

Localization of genomic regions controlling the variation of fruit quality and disease resistance traits in apple: selection signatures and association genetics

Résumé

Depuis la domestication du pommier, l'homme a progressivement sélectionné des variétés plus performantes, notamment pour la qualité du fruit, la productivité ou la résistance aux pathogènes. Les bases génétiques de ces caractères ont été explorées par cartographie en descendances F1 ne permettant d'explorer qu'une infime partie de la diversité génétique disponible.

L'objectif de la thèse portait sur l'analyse des bases génétiques de caractères de qualité du fruit et de résistance du pommier à la tavelure et au feu bactérien dans des collections représentant une diversité plus large. Le génotypage de core collections de variétés anciennes s'est fait à l'aide de deux puces 8k et 480k SNPs ou grâce à du re-séquençage de gènes. Des traces de différenciation génétique entre pommes à cidre et à couteau ont été identifiées et partiellement reliées à la voie des polyphénols. Après analyse de l'étendue du déséquilibre de liaison à large et fine échelle, une approche de génétique d'association a permis l'identification de régions génomiques associées à la variation de plusieurs caractères de qualité du fruit, dont le haut du groupe de liaison 16 rassemblant l'acidité (locus *Ma*), la fermeté, la jutosité et l'amertume (gène *LAR*). Pour la résistance au feu bactérien, une région contenant un homologue du gène *NPR1* (activateur de défenses) a été identifiée.

Cette thèse a ainsi permis de préciser la localisation potentielle de QTLs identifiés préalablement par cartographie génétique et d'identifier de nouvelles ressources utiles dans de futurs programmes de sélection assistée par marqueurs.

Mots clés *Malus domestica* ; GWAS ; signatures de sélection ; qualité du fruit ; acidité ; polyphénols ; résistance ; *Venturia inaequalis* ; *Erwinia amylovora*

Abstract

Since apple domestication, humans have progressively selected improved varieties, especially for traits linked with fruit quality, productivity or resistance to pathogens. The genetic bases underlying these traits have been explored thanks to genetic mapping in F1 segregating populations that only allows the study of a small part of the available genetic diversity.

The aim of this work was to analyze the genetic bases of fruit quality and disease resistance against apple scab and fire blight, in collections of old apple varieties representing a much larger diversity. Genotyping of core collections was performed either with arrays of 8k and 480k SNPs or by resequencing of chosen genes. Signs of genetic differentiation were identified between cider and dessert apples and were partially linked to the polyphenols pathway. After studying linkage disequilibrium, both on a large and a small scale, an association genetics approach allowed the identification of genomic regions associated with the variation of several fruit quality traits. Especially, the top of linkage group 16 was found to be linked with acidity (locus *Ma*), firmness, juiciness and bitterness (*LAR* gene). Concerning the resistance of apple to fire blight, a region containing a homolog of the *NPR1* gene (defense activator) was identified.

This thesis allowed the refining of the putative localization of previously identified QTLs and the identification of new genetic resources that could be useful in future selection programs using marker assisted selection.

Key words *Malus domestica* ; GWAS ; signatures of selection ; fruit quality ; acidity ; polyphenols ; resistance ; *Venturia inaequalis* ; *Erwinia amylovora*

L'Université Nantes Angers Le Mans