Université
de Montpellier

Centre International
d'Études Supérieures en
Sciences Agronomiques

Rapport de stage

Diplôme de Master 2
Mention Biodiversité — Écologie — Évolution
Parcours DARWIN : Biologie Évolutive & Écologie

2020

Génomique de l'adaptation au climat des pommiers cultivés et sauvages : Recherche de traces de sélection dans des gènes impliqués dans la floraison

Justine Floret

Laboratoire d'accueil CIRAD, UMR AGAP
Équipe AFEF

Sous la direction de
Bouchaib Khadari (AGAP, AFEF),
Stéphanie Sibidé-Bocs (AGAP),
Amandine Cornille (GQE-Moulon, ATIP)

Soutenu les 11 et 12 juin 2020.
Membres du jury signataires du PV d'examen :

Emmanuel DOUZERY	Vincent RANWEZ	Emmanuelle JOUSSELIN	Thomas LENORMAND	Cyrille VIOLLE
Président	*Représentant Montpellier SupAgro*	*Directrice des Études*	*Directeur des Études*	*Directeur des Études*

Résumé

Une approche par gène candidats liés à la floraison chez 254 variétés de pommier cultivé (*Malus domestica)* et 88 de ses proches parents sauvages (*Malus orientalis, Malus sieversii, Malus sylvestris, Malus baccata*) nous a permis d'explorer la variation phénologique de la floraison en lien avec la domestication du pommier. Pour cela, nous disposons des dates de floraison sur quatre années et de 18 663 variants génétiques détectés sur les gènes de floraison des cinq espèces. L'étude des dates de floraison des espèces sauvages et cultivés a révélé que les pommiers sauvages fleurissent plus précocement que les pommiers cultivés. Avec le logiciel PCAdapt qui permet de détecter des signaux de sélection positive liés à l'adaptation locale, des variants génétiques ont été détectés chez le pommier cultivé, d'autres sont présents chez ses proches parents sauvages *M. sieversii, M. orientalis* et *M. sylvestris*. Ces découvertes apportent un nouvel aperçu du processus de domestication du pommier pour un des caractère phénotypique clé, la floraison.

Mots clés : Floraison, domestication, signatures génomique de sélection positive, balayage sélectif, *Malus domestica*

Candidate gene approach applied to 254 cultivated apples (*Malus domestica*) and 88 crop wild apple relatives (*Malus orientalis, Malus sieversii, Malus sylvestris, Malus baccata*) allowed us to study the genetic variation underlying flowering time variation in apple. To that aim, we used flowering dates recorded over four years and 18 663 SNP located in flowering time genes for the five apple species. We used the software PCAdapt to detect strong and recent signals of positive selection that can be associated to domestication or local adaptation, while taking into account population structure. Few genetic variants detected were found to be specific to the cultivated apple genetic group, others SNP were also present in its wild relatives *M. sieversii, M. orientalis* and *M. sylvestris*. These findings are bring new insight into the process of apple domestication and on the genetic variants involved.

Key words: Flowering time, domestication, genomic signature of positive selection, selective sweeps, *Malus domestica*

Remerciement

Je tiens à remercier mes directeur.es de mémoire Bouchaib Khadari, Stéphanie Bocs et Amandine Cornille. Je vous remercie pour votre aide, votre patience et vos différents conseils tant sur les différentes analyses que sur la rédaction. Merci pour votre précieux soutien lors du confinement qui ce ne fut pas une période toujours facile.

Je remercie également l'UMR AGAP (INRAe, CIRAD, Institut Agro – Montpellier Supagro) et Gis Fruits pour m'avoir permis de faire ce stage et plus particulièrement l'équipe AFEF. Merci à vous pour votre accueil et vos conseils. Je tiens à remercier particulièrement Evelyne Costes, Fernando Andres-Lalaguna, Benoit Pallas et Isabelle Farrera pour nos nombreux échanges qui m'ont permis de prendre du recul sur mes données et de les aborder sous un angle différent.

Je souhaiterai exprimer ma gratitude à toute l'équipe ID, pour votre accueil chaleureux et votre soutien lors du confinement qui fut une aide précieuse également.

Merci à toute l'équipe de la ferme du Moulons malgré la brièveté du mon séjour, notamment à Xi Long pour sa patience face à toutes mes questions.

Pour finir je souhaiterai remercier toute l'équipe pédagogique du master Darwin pour m'avoir permis d'intégrer ce master et de m'avoir laissée ma chance, ainsi qu'à tous mes camarades de promotion sans qui les difficultés rencontrées auraient été bien plus difficiles à franchir.

Contribution personnelle

Pour mon stage, j'avais à disposition un jeu de données génomiques correspondant à du séquençage par capture de gènes candidats ciblés. Une des premières étapes a été de comprendre comment ces données avaient été conçues et dans quel but afin de les exploiter au mieux. Cette étape fut enrichissante, elle m'a permise de comprendre comment mes données avaient été produites, dans quel but, et ainsi mieux prendre en main ma problématique de recherche sur l'adaptation de gènes de floraison chez le pommier. J'ai pu tester différentes approches dans mes analyses avec différents logiciels afin d'explorer au mieux mes données et de trouver les plus adaptées à ma problématique.

Mon travail fut décomposé en trois grandes parties : les analyses génomiques ciblées sur la structure des populations, les analyses phénologiques et la recherche de traces de sélection. Pour chacune d'elle, j'ai pu aborder des outils d'analyses complémentaires et explorer différentes facettes de ma problématique. Un des enjeux a été de ne pas perdre le fils rouge de mon hypothèse de départ et veiller à équilibrer ces différentes parties.

J'ai eu la chance d'être encadrée par trois chercheurs : Bouchaib Khadari (généticien des populations), Stéphanie Bocs ép. Sidibe (bioinformaticienne en omique végétale intégrative) et Amandine Cornille (génomicienne des populations chez les pommiers). J'ai pu m'intégrer dans différentes équipes et discuter avec des chercheurs de domaines différents, me permettant d'enrichir ma culture scientifique et d'acquérir des compétences dans les approches d'analyse de génomique des populations. Bouchaib et Stéphanie m'encadraient au quotidien au Cirad à Montpellier et Amandine à distance depuis le laboratoire GQE-Le Moulons à Gif-sur-Yvette (91), quelques déplacements y ont été programmés. Être encadrée par trois chercheurs aux compétences complémentaires a permis d'envisager les questions sous différents angles et d'explorer au mieux mes données. Nous avions un rendez-vous hebdomadaire par visioconférence afin de faire le point sur mes avancées. Je devais aussi veiller à faire des comptes-rendus réguliers par mail afin que tous puissent suivre en même temps le déroulement de mes analyses.

COVID-19

Lors du confinement, j'ai pu être en télétravail dans de bonnes conditions mis à part un ordinateur portable 32 bits. En effet, mon stage ayant été conçu avec des séjours à Gif-sur-Yvette, j'étais déjà équipée pour télé-travailler avec un ordinateur portable et une connexion VPN. Chez moi je disposais d'une connexion internet suffisante pour réaliser mes analyses bioinformatiques et génomiques sur le cluster du CIRAD. Nous avions également l'habitude de travailler à distance avec A. Cornille, nous étions donc déjà organisées. Nous avons maintenu nos rendez-vous hebdomadaires par visioconférence (avec toujours au moins deux des trois encadrants présents) et les comptes-rendus par mail. J'ai pu être guidée à distance pour la prise en main de certains logiciels, mais ce fut moins évident que si cela avait été le cas en présentiel et certaines analyses ont été ralenties. En outre, ma seconde mission à Gif-sur-Yvette a été annulée. Certains échanges ont été un peu ralentis, notamment avec d'anciens chercheurs de l'équipe qui étaient moins disponibles, mais toujours joignables. Nous avons donc eu des difficultés à réunir toutes les informations nécessaires sur les jeux de données. Différentes personnes ont travaillé sur ces jeux : contacter les personnes intéressées pour trouver certaines informations ne fut pas toujours évidents et nous avons des choix qu'il nous reste à confirmer mais qui au final n'ont pas bloqué l'analyse poussée des résultats.

Table des matières

Introduction

La floraison est une étape clé du cycle reproducteur des plantes à fleurs. Sous des conditions écologiques favorables, la floraison permet alors le développement d'un nombre conséquent de fleurs. Elle contribue ainsi au succès reproducteur de la plante et à sa valeur sélective (Chuine et Beaubien 2001). La floraison est une étape phénologique clé d'une plante qui est contrôlée par plusieurs voies, toutes interdépendantes (Srikanth et Schmid 2011) :

- des voies exogènes telle que la vernalisation (l'exposition au froid), la photopériode (la quantité et la qualité de la lumière,), la température et ;
- des voies endogènes : voie gibbérelline (le besoin d'acide gibbérellique), voie autonome (régulateurs endogènes), l'âge de la plante.

La date de floraison des plantes pérennes et annuelles est donc liée au climat local. Différentes études ont pu lier la floraison à l'adaptation locale de populations par différentes approches (Price et al. 2020). Des études se basent sur des loci de caractères quantitatifs (*quantitative trait loci, QTL*) comme chez l'orge sauvage, *Hordeum spontaneum*, où un QTL associé à la date de floraison est impliqué dans l'adaptation locale (Wang et al. 2018 ; Verhoeven et al. 2008). Chez un modèle d'étude pérenne comme le peuplier, *Populus tremula,* le rôle de la floraison dans l'adaptation au climat local a été montré en utilisant cette fois ci trois autres approches (Wang et al. 2018). Cette étude s'appuie en effet sur la confrontation des résultats obtenus par des approches différents : i) deux études en génomique d'association (*EGA*) pour examiner le lien entre variation génétique et variation phénotypique; faire le lien entre ces variants génétiques et le climat; ii) l'analyse des balayages sélectifs par analyse de composantes principales (ACP) qui permet de prendre en compte la structuration des populations. Cette dernière approche, implémentée dans le logiciel PCAdapt (Privé et al. s. d.), permet de détecter des SNP liés à l'adaptation locale tout prenant en compte les faux positifs pouvant être induits par la structuration des populations (Tiffin et Ross-Ibarra 2014). Cette approche a déjà été utilisée pour détecter des SNP sélectionnés lors de la domestication de l'olivier *Olea europaea* par exemple (Gros-Balthazard et al. 2019).

Le rôle majeur la floraison dans l'adaptation au climat en fait un caractère phénotypique crucial à prendre en compte dans les programmes d'amélioration variétale et plus généralement en agronomie. En effet, induite au moment le plus favorable, la floraison contribue à un meilleur rendement en maximisant le nombre de fleurs, participant ainsi à une production maximale de fruits et de graines (Andrés et Coupland 2012). Lors de la

domestication des plantes, de nombreux traits ont été sélectionnés : sensibilité à la photopériode, augmentation des rendements, taille des fruits et des graines, perte de la dormance des graines (Meyer, DuVal, et Jensen 2012; Gaut, Díez, et Morrell 2015). Ces nombreux changements morphologiques différencient les espèces cultivées de leurs ancêtres et parents sauvages, on parle de syndrome de la domestication (Miller et Gross 2011). La domestication s'accompagne aussi d'un coût : une perte de diversité génétique par rapport aux parents sauvages et par la fixation de variants délétères non létaux occasionnés par un goulot d'étranglement (Gaut, Díez, et Morrell 2015). Les espèces pérennes sont moins impactées par ce coût que les espèces annuelles. Plusieurs facteurs peuvent expliquer ceci : un goulot d'étranglement moins prononcé (pommier (Cornille et al. 2012), olivier (Gros-Balthazard et al. 2019)), de nombreuses introgressions avec des parents sauvages limitant l'effet du goulot d'étranglement initial, une propagation clonale et non par les graines dû aux caractéristiques particulières des espèces pérennes favorisant les flux de gènes intra-cultivé (phase juvénile longue, temps de génération longs, présence de systèmes d'auto-incompatibilité) (Gaut, Díez, et Morrell 2015).

Le pommier cultivé (*Malus domestica* Borkh.) fait partie des espèces fruitières les économiquement importantes des régions tempérées; la pomme l'un des fruits les plus consommés dans le monde ces cinq dernières années (http://www.fao.org/faostat). Le pommier cultivé est un modèle d'étude pour comprendre les processus de domestication des espèces pérennes (Cornille et al. 2014). Il fut domestiqué à partir du pommier sauvage d'Asie Centrale, *Malus sieversii Lebed.*, il y a environ 4000 ans dans les montagnes de Tian Shan. Ensuite, le pommier cultivé a été dispersé vers l'Europe le long des Routes de la Soie et a subi de nombreuses introgressions substantielles par *Malus sylvestris* Mill. et *Malus orientalis* Uglitz. (Cornille et al. 2012; 2014). Les études génétiques montrent que ces introgressions ont été si importantes qu'aujourd'hui *M. domestica* serait plus proche génétiquement de *M. sylvestris*, parent sauvage européen, que de son ancêtre asiatique *M. sieversii* (Cornille et al. 2012). Plusieurs gènes, impliqués dans la taille des fruits, leur fermeté et leur acidité, montrent des traces de signatures de sélection positive récente en lien avec la domestication du pommier (Yao et al. 2015; Duan et al. 2017). Certaines études se sont basées sur les microARN, comme celle par Yao et al (2015) qui s'est focalisée sur le micro-RNA 172 connu chez *Arabidopsis thaliana* pour stimuler la croissance des siliques. Une autre étude s'est basée sur le séquençage en courts fragments du génome de plusieurs variétés de pommier cultivé et de ses apparentés sauvages eurasiatiques (Duan et al. 2017). Cependant cette étude est limitée

car les auteurs n'ont pas pris en compte les introgressions de *M. domestica* par ses parents sauvages. Si des caractères liés aux qualités organoleptiques du fruit ont pu être sélectionnés au cours de la domestication, certains des ces caractères sont aussi retrouvés chez les parents sauvages (Yao et al. 2015), une sélection naturelle pré-domestication accentuée lors de la domestication a pu se produire.

Malgré le rôle majeur de la floraison dans la valeur adaptative d'une plante et son implication dans l'adaptation locale (Chuine et Beaubien 2001*;* Price et al. 2020)*)*, l'existence de signatures de sélection pour des gènes liés à la floraison au cours de la domestication de *M. domestica* est encore une question ouverte. Une telle étude permettrait de savoir si au cours de la domestication, la sélection d'individus dont la floraison était adaptée au climat local a été favorisée; ou si la domestication s'est focalisée sur des traits autour de la qualité des fruits ou d'autres caractères d'intérêts. Dans l'étude de Duan et al (2017), les gènes liés à la floraison ne sont pas apparus comme sélectionnés lors de la domestication. Une approche par gènes candidats permettrait de cibler les gènes impliqués dans la floraison et ainsi pouvoir détecter un signal de sélection qui aurait pu être manqué dans les études précédentes ou masqué par le signal plus fort d'autres gènes sélectionnés. Deux questions peuvent être examinées dans une telle étude: i) Est-ce que la variation allélique des SNP inclus dans les gènes impliqués dans la floraison traduit des traces de sélection liées à la domestication du pommier ? ii) Est-ce que ces signatures de sélection positive sont variables aux niveaux intra et interspécifiques et selon les groupes de génotypes cultivés et sauvages?

Le but de mon stage a été d'examiner ces questions par une étude des signatures de sélection positive récentes dans des gènes candidats liés à la floraison, tout en prenant en compte la structuration génétique des populations. Pour cela, je disposais d'un jeu de données génomiques unique basé sur l'analyse de 254 cultivars français de pommier cultivé ("core-collection" de variétés à pomme type "dessert"; (Lassois et al. 2016)) et de 88 génotypes appartenant aux quatre espèces sauvages apparentées au pommier cultivé: *M. orientalis*, *M. sieversii*, *M. sylvestris*, *M. baccata* Borkh.. (parent sauvage plus éloigné) (Höfer et al. 2014). Les régions codantes de plusieurs gènes candidats potentiellement liés à la floraison ont été séquencées en utilisant la méthodologie de génotypage par séquençage (*GBS*) par capture d'ADN ciblé (Soriano 2017 non publié ; Kozarewa et al. 2015). Les gènes étudiés sont des séquences homologues de gènes d'*Arabidopsis thaliana* connus comme étant impliqués dans le contrôle de la floraison. Tout d'abord j'ai analysé la variabilité du trait phénotypique date

de pleine floraison, observé sur plusieurs années chez les pommiers cultivés et sauvages. Puis j'ai réalisé une analyse de structuration génétique des populations basée sur 5 154 polymorphismes d'un seul nucléotide (*single nucleotide polymorphism, SNP*). Ensuite j'ai fait une recherche de détection de traces de sélection en cherchant des signatures de sélection positive récente, tout en tenant compte de la structuration génétique du pommier cultivé et de ses apparentés sauvages. La prise en compte de la structuration des populations était indispensable afin d'éviter des effets confondant liés à l'histoire démographique des espèces: aux histoires de divergence, à la dérive et aux flux de gènes (Tiffin et Ross-Ibarra 2014; Duforet-Frebourg, Bazin, et Blum 2014). Les résultats obtenus pourront être mis en perspective avec une étude EGA pour le trait "date de débourrement" sur la même core-collection de pommiers cultivés, réalisée par mon équipe d'accueil (Ousmane 2019, non publié).

Matériels et méthodes

Un résumé des jeux de données et des différentes analyses réalisées est présenté en Figure 1.

1- Données de dates de floraison

La première étape fut d'explorer les données de phénologie acquises avant mon stage afin d'examiner s'il existe une différence dans les dates de floraison entre les quatre espèces sauvages et l'espèce cultivée. Pour répondre à cette question, nous disposions des dates de pleine floraison, stade phénologique noté lorsque 50% des boutons floraux sont présents sur l'arbre. Ces données, notées en nombre de jours à partir du 1er janvier de l'année, sur deux collections:

- la collection allemande située à Dresden (latitude: 51.77, longitude: 11.14) composée d'individus des **espèces sauvages N=67**, (19 *M. baccata*, 8 *M. orientalis*, 20 *M. sieversii* et 20 *M. sylvestris*). Les données correspondent à des *dates de floraison* **pour un réplicat par génotype de 2011 à 2014.**

- la "core collection" à l'INRAE Mauguio (latitude: 43.62, longitude: 4.01) composée de **235 individus de pommiers cultivés *M. domestica.*** Ces données correspondent à des *dates de floraison* **de 2016 à 2019** avec **deux réplicas par génotype les deux premières années (2016, 2017) puis quatre réplicats par génotype (2018, 2019)**.

Afin d'homogénéiser ces jeux de données, un seul réplicat par génotype pour la collection des pommiers sauvages, et plusieurs réplicats pour la collection des pommiers cultivés, nous

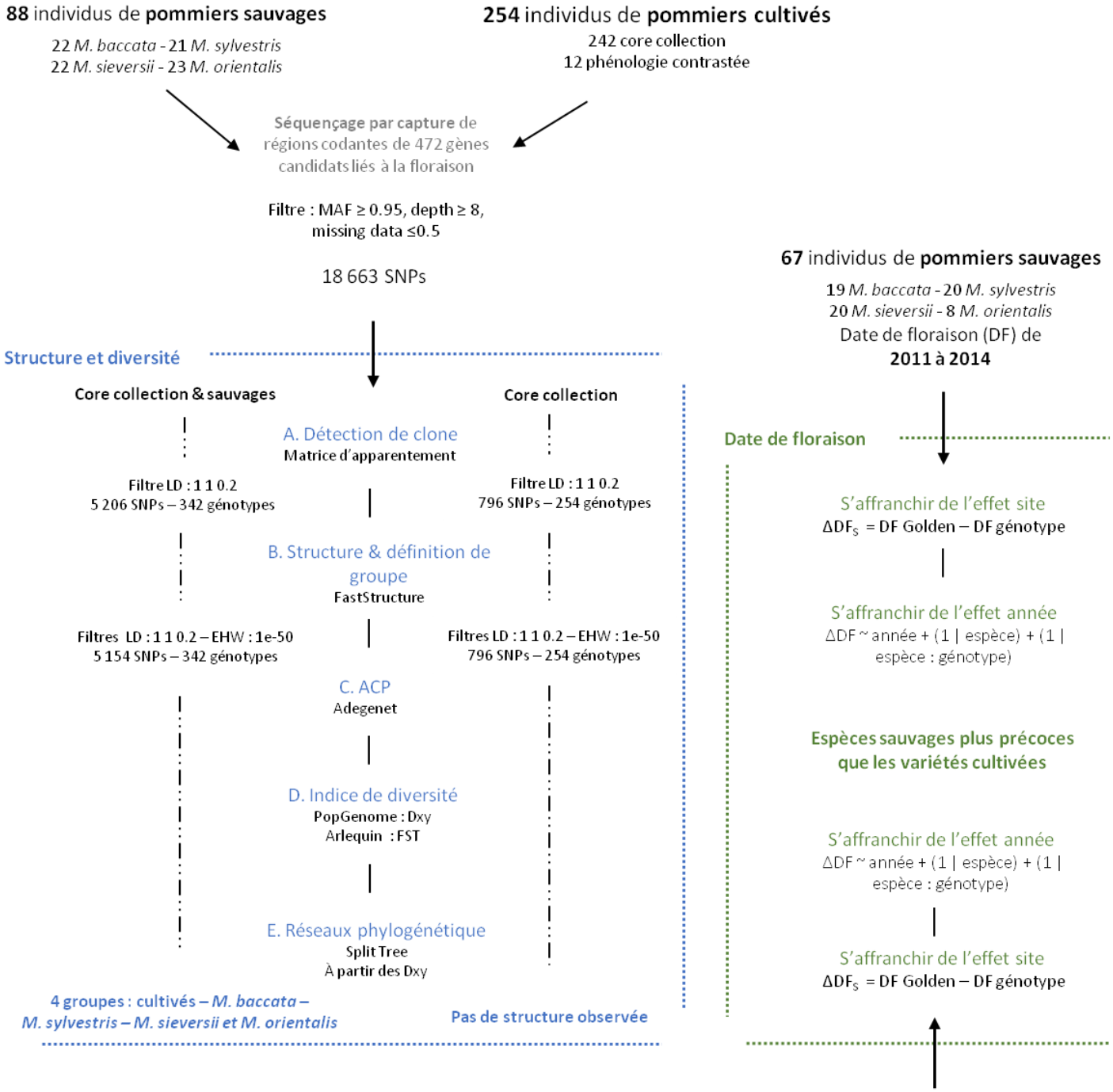


Figure 1 : Schéma récapitulatif des différentes analyses réalisées et du flux de données au cours des analyses

avons fait la moyenne des réplicats par génotype des individus cultivés *M. domestica*. Les *dates de floraison* des espèces sauvages et des pommiers cultivés étant hétérogènes de part leur lieux d'échantillonnage (Allemagne *versus* France) et leurs années d'échantillonnage (2011-2014 en Allemagne et 2016-2019 en France); il est nécessaire de s'affranchir de ces deux effets pour pouvoir les comparer. De plus les espèces sauvages étant présentes uniquement sur le site allemand et les variétés cultivées sur le site français, l'effet site et l'effet espèces sauvages - cultivés sont confondus, il faut donc aussi s'affranchir de ces effets.

a- S'affranchir de l'effet du site : Allemagne vs France

Le climat entre les deux sites (est de l'Allemagne - sud de la France) est différent, on peut donc s'attendre à un fort impact de l'environnement sur les dates de floraison. Nous n'avons pas pu modéliser l'effet du site sur l'ensemble du jeu de données car aucune espèce n'est présente sur les deux sites. Nous avons donc utilisé la date de pleine floraison de la variété de pommier cultivé Golden, présente sur les deux sites à toutes les années, comme date de référence. Nous avons soustrait la *date de floraison* du génotype observé *j* ($DF_{génotype(j)}$) à la *date de floraison* de la variété Golden (DF_{Golden}) pour l'année *i* :

$$(1) \quad \Delta DF_{(i,j)} = DF_{Golden(i)} - DF_{génotype\ (i,j)}$$

$\Delta DF_{(i,j)}$ représente le nombre de jour entre la *date de floraison* de Golden et la *date de floraison* du génotype j à l'année i. Nous pouvons alors comparer indirectement les espèces entre les deux sites : les espèces sauvages sont-elles plus précoces que la variété Golden ? Et comment les génotypes cultivés se positionnent par rapport à la variété Golden?

b- S'affranchir de l'effet année : 2011-2014 vs. 2016-2019

Le climat varie d'une année sur l'autre, impactant la date de floraison des espèces. Afin de s'affranchir de cet effet, nous avons réalisé un modèle linéaire mixte (*MLM*) avec le package lme4 (Bates et al. 2015) du logiciel R (*R Core Team, 2017*). Le modèle a été fait sur les ΔDF calculées précédemment afin de modéliser l'effet des années et de pouvoir s'exempter de celui-ci. Deux MLM seront utilisés, un par jeu de données :

$$(2) \quad \Delta DF_{Sauvage} \sim année + (1 \mid espèce) + (1 \mid espèce : génotype)$$

$$(3) \quad \Delta DF_{Cultivé} \sim année + (1 \mid génotype)$$

L'équation (2) modélise ΔDF du jeu de données sauvages ($\Delta DF_{Sauvage}$) avec l'année en effet fixe, l'espèce en effet aléatoire et l'effet niché du génotype au sein de chaque espèce. L'effet niché va permettre d'évaluer si la variance des ΔDF est plus due à l'espèce ou aux génotypes

au sein des espèces. L'équation (3) concerne le jeu de données "cultivé" en modélisant le ΔDF ($\Delta DF_{Cultivé}$). Comme pour le modèle précédent l'année est un effet fixe et le génotype un effet aléatoire (il n'y a qu'une espèce dans ce jeu de données : *M. domestica*). Il n'était pas possible d'estimer l'effet niché année : génotype car les génotypes sont les mêmes pour toutes les années. Pour mesurer l'effet des génotypes sur la *date de floraison* nous avons utilisé le jeu de données avec les réplicats afin d'avoir de la variabilité intra-génotypique et inter-génotypique :

(4) $\Delta DF_{Cultivé}$ ~année + (1 | génotype) + (1 | réplicat : génotypes)

Pour les modèles (2) et (3) les effets années ont été calculés. Lorsque les effets années étaient significatifs, la valeur de l'effet a alors été enlevée du ΔDF donnant ainsi un ΔDF ajustée : $\Delta DF_{ajusté}$.

c- Comparaison des $\Delta DF_{ajusté}$ pommiers sauvages vs. pommiers cultivés

Une fois les effets site et année pris en compte, nous avons comparé les $\Delta DF_{ajusté}$ des espèces sauvages et cultivés. Nous avons utilisé le test des rangs de Wilcoxon de comparaison de moyenne implémenté dans package Stat (v. 3.5.1) du logiciel R (*R Core Team, 2017*), fonction *pairwise.wilcox.test*, avec un ajustement de Bonferroni des p-value.

Le script R est disponible à cette adresse :

https://gitlab.southgreen.fr/sidibebocs/domestication_selection_trace/blob/master/FT_analysis.R

2- Les données de séquençage

Au total, nous avons travaillé sur un jeu de données "complet" incluant 342 génotypes de pommiers, dont 254 pommiers cultivés et 88 pommiers sauvages. Les pommiers sauvages (N=88, incluant : 22 *M. baccata*, 21 *M. sylvestris*, 22 *M. sieversii*, 23 *M. orientalis*) ont été récoltés à partir de la collection vivante allemande situé à Dresden (Allemagne). Les variétés de pommier cultivé, *M. domestica, (N*=254 variétés de pommiers) proviennent de la collection INRAE à Mauguio (France). Cette collection inclut 242 génotypes à dessert représentant 90% de la diversité allélique totale connue chez le pommier cultivé (Lassois et al. 2016), plus d'autres génotypes cultivés dont 12 connus pour leur phénologie contrastée qui sont inclus dans le jeu de données. Dans la suite du rapport, l'ensemble des variétés de pommier cultivé (N=254) sera noté *core collection*.

Les 342 génotypes, tous diploïdes, ont été séquencés avec la méthode de génotypage par séquençage (*GBS pour "genotyping by sequencing"*) par capture d'ADN ciblé ("*target*

capture"). Le premier groupe de gènes codant des polypeptides ciblés, contenait 472 gènes orthologues de gènes liés à la floraison chez *Arabidopsis thaliana*. Le second groupe contenait 251 gènes, ils ont été repérés comme gènes liés à la floraison chez le pommier à la suite à des études de détection de QTL sur des populations multi-parentales (Allard et al. 2016). Pour ce groupe de gènes codants, ce sont principalement les régions du génome correspondant aux parties exoniques qui ont été ciblées. Les données de séquence de type "courtes lectures" obtenues par Illumina ont été alignées sur le génome de référence du pommier cultivé variété Golden Delicious GDDH1.3 (annotations v1.1, https://iris.angers.inra.fr/gddh13/the-apple-genome-downloads.html). L'appel de SNP a été effectué à partir des ces alignements (Soriano, 2017 , non publié). Les SNP ont ensuite été filtrés : les SNP présentant moins de 5% de données manquantes, ayant une fréquence d'allèle mineur (minor allele frequency MAF) supérieure ou égale à 5%, une profondeur d'alignement d'au moins 8 et qui étaient bi-alléliques ont été conservés. C'est ainsi que 18 663 *SNP* de haute qualité ont été obtenus.

3- Structuration et diversité génétique des populations

Plusieurs analyses ont été réalisées afin de connaître la diversité et la structure génétique des populations chez les pommiers cultivés et sauvages représentés dans le jeu de données. La connaissance sur la structure génétique des populations étudiées est essentielle pour détecter les traces de sélection positive dans les génomes. La structure des populations peut en effet être source de facteurs confondants avec des signatures de sélection de certains SNP (Purcell et al. 2007; Verhoeven et al. 2008) ; Les analyses présentées ci dessous ont été réalisées sur deux jeux de données: 1) le jeu de données dit "complet" (incluant les espèces sauvages et le pommier cultivé, 342 génotypes au total), 2) le jeu de données incluant uniquement les génotypes de pommier cultivé *M. domestica*, 254 génotypes au total. Le détail des analyses est disponible à cette adresse :
https://gitlab.southgreen.fr/sidibebocs/domestication_selection_trace/wikis/Structure_analysis

a- Matrice d'apparentement génétique et recherche de clones

Nous avons réalisé une matrice d'identité par descendance (*IPD*) pour détecter la présence de clones (individus considérés comme génétiquement identiques ou très proches). L'un des deux individus de chaque paire de clone détectée a été retiré du jeu de données pour les analyses de diversité afin de tenir compte des hypothèses des modèles des logiciels comme FastStructure; la présence d'individus fortement apparenté peut en effet créer des faux

groupes génétiques. Nous avons ensuite écarté les SNP en déséquilibre de liaison (DL) trop élevé avec le logiciel PLINK v1.9 (Purcell et al. 2007). Le DL estimé à partir du génome GDDH13 chez le pommier cultivé est connu pour diminuer rapidement lorsque la distance augmente entre les marqueurs, notamment pour les gènes liés à la floraison (Urrestarazu et al. 2017). Nous avons paramétré le logiciel pour calculer le DL sur des fenêtres de 1 kbp (minimum admis par le logiciel) tous les 1 SNP pour un r^2 inférieur à 0.2, valeur de r^2 retrouvée dans la littérature ((Duan et al. 2017; Urrestarazu et al. 2017). Le paramètre DL est ainsi noté "DL : *1 1 0.2* dans le reste du rapport". La matrice d'apparentement génétique entre individus a été calculée pour le jeu de données complet (N=332, sauvages et domestiqués) avec le logiciel PLINK et le module *-genome*. L'indice PI-HAT du module, permet de calculer IPD de paires d'individus selon la formule est la suivante :

(5) PI-HAT = P(IPD=2) + 0.5*P(IPD=1)).

Une valeur de PI-HAT de 1 signifie que les individus sont génétiquement identiques (clones), de 0,5 qu'ils sont parents-descendant ou pleins frères. Deux réplicas de deux variétés de *M. domestica* dans le jeu de données ont permis d'étalonner la détection des clones: les paires d'individus dont le PI-HAT était proche de celui des doublons étaient considérées comme des clones. Ainsi, nous avons gardé les paires d'individus dont le PI-HAT était supérieure à 0,4 (Muranty et al. 2020)

b- Structuration génétique des populations

Afin d'étudier la structuration génétique entre sauvages et cultivés, nous avons tout d'abord filtré les SNP présentant un fort DL (filtre : 1 1 0.2, identique à celui utilisé pour le calcul de l'IPD) et ne respectant pas l'équilibre d'Hardy-Weinberg (*EHW*). Ces SNP sont retirés du jeu de données afin de pouvoir réaliser l'analyse de structure. Le seuil pris pour EHW est de 1e-50 comme conseillé dans la documentation de l'outil PLINK, afin d'éviter de filtrer des variants associés à des traits déviant légèrement de l'équilibre de Hardy-Weinberg. Suite à ces deux filtres, 5 154 SNP ont été conservés pour les 333 génotypes sauvages et cultivés pour le jeu de données complet, et 796 SNP pour 249 génotypes cultivés pour la core collection. Nous avons inféré la structure génétique des populations avec le logiciel FastStructure (Raj et al 2014), avec ce jeu de données SNP pour inférer le nombre le plus probable de populations respectant l'équilibre de HW et les proportions d'admixture entre les individus. Le résultat a été visualisé avec le logiciel Structure Selector (*Liu JX 2018,* *https://lmme.ac.cn/StructureSelector/index.html*).

c- *Analyse en composante principale*

L'analyse en composante principale (ACP) a été réalisée à l'aide du package R Adegenet v2.1.0 (Jombart et Ahmed 2011) du logiciel R (*R Core Team, 2017*) à partir du jeu de données filtré pour le DL et l'EHW de 5 154 SNPs de 333 génotypes sauvages et domestiqués, et de 796 SNP de 249 génotypes domestiqués. Le script R est disponible à cette adresse : https://gitlab.southgreen.fr/sidibebocs/domestication_selection_trace/blob/master/PCA_ccw.R

d- *Diversité génétique*

Différents indices de diversité génétique ont été calculés sur les groupes (i.e. regroupant l'ensemble des individus assignés à plus de 80% dans un groupe donné) définis suite aux analyses précédentes, à savoir avec FastStructure. L'indice de fixation (*Fst*) a été calculé avec l'outil Arlequin v3.5.2.2 (Excoffier et Lischer, 2010) et la distance génétique Dxy avec le package R Popgenome v2.7.5 (Pfeifer et al 2014), les deux indices ont été inférés le jeu de données filtré : filtre LD 1 1 0.2 et filtre HWE : 1e-50. Le détail des analyses est disponible à cette adresse : https://gitlab.southgreen.fr/sidibebocs/domestication_selection_trace/wikis/Diversity_analysis et le script R Popgenome ici : https://gitlab.southgreen.fr/sidibebocs/domestication_selection_trace/blob/master/Popgenome_analysis.R

e- *Réseau phylogénétique*

Nous avons visualisé les relations génétiques entre populations inférées avec le logiciel FastSTRUCTURE avec le logiciel SplistTree v5.1.4 (Huson et Bryant, 2006).

4- Détection des traces de sélection

La détection des SNP liés à l'adaptation locale a été faite avec le logiciel PCAdapt Version 4.3.3 (Privé et al. s. d.) implémenté sur (*R Core Team, 2017*). PCAdapt est un modèle hiérarchique bayésien qui se base sur la détection de SNPs anormalement liés aux axes de l'ACP. Ces loci sont candidats à l'adaptation locale (Duforet-Frebourg, Bazin, et Blum 2014). Une fois les clones et individus mal assignés (Figure 3.A) retirés du jeu de données initial (N=342), l'analyse PACadapt a été réalisée sur 327 individus des espèces sauvages (N=79) et cultivés (N=248). Ensuite nous avons procédé aux analyses sur différents groupes de jeu de données à partir de 18 663 SNP :

- espèces sauvages et cultivés (N=327)

- espèces sauvages seules (N=79)
- *M. domestica, M. sieversii, M. orientalis* (N=288)
- *M. domestica, M. sieversii, M. orientalis, M. sylvestris* (N=308)
- espèce cultivée : *M. domestica* (N=248)

Pour chacun des groupes le raisonnement a été le même, il est basé sur les critères suivants: i) choix du nombre d'axe de l'ACP à garder selon la proportion de variance expliquée, ii) élimination des individus dont les valeurs sont aberrantes et ne correspondant pas à la structure génétique observée, iii) détection des SNP anormalement liés aux axes et correction de bonferroni des p-value, iv) élimination des SNP corrélés aux axes qui sont liés à la structuration des populations. PCAdapt nous a permis de nous affranchir des effets liés à la structure génétique pouvant être confondants avec l'adaptation locale, en choisissant les axes qui ne sont pas à ces effets (Tiffin et Ross-Ibarra 2014). Ainsi les SNP détectés sont liés à l'adaptation locale et non à des faux négatifs liés à la structuration génétique des populations. Si cette démarche peut paraître conservative, elle nous a fait manquer la détection de certains SNPs potentiellement liés à l'adaptation locale, elle nous a évité la détection de faux-positifs (Duforet-Frebourg et al. 2016). Le script R est disponible à cette adresse : https://gitlab.southgreen.fr/sidibebocs/domestication_selection_trace/blob/master/PCAdapt.R

Une fois les SNP associés aux bons axes détectés, ils ont été comparés entre groupes analysés. L'outil de visualisation Integrative Genomics Viewer (IGV) nous a permis de comparer les séquences génétiques entre les groupes analysés des SNP détectés (Robinson et al, 2011, https://software.broadinstitute.org/software/igv/home). Les détails de la transformation des données pour l'utilisation d'IGV est disponble ici : https://gitlab.southgreen.fr/sidibebocs/domestication_selection_trace/wikis/diversity_annotation_selection_to_function

Résultats

1- Variabilité des dates de floraison

a- S'affranchir de l'effet année : 2011-2014 vs. 2016-2019

Les résultats du premier MLM réalisé sur le ΔDF du jeu de données "sauvages" sont consignés dans le Tableau 1. Ces résultats décrivent la différence entre les dates de pleine floraison de la variété Golden et des 4 espèces sauvages ($\Delta DF_{Sauvage}$) selon l'année, les espèces et l'effet niché des génotypes au sein des espèces de 2011 à 2014 (Tableau 1). L'effet année global affecte significativement le ΔDF_S (p-value < 2.2e-16). Cependant, ce n'est pas le

cas de toutes les années, l'année 2013 n'affecte pas $\Delta DF_{Sauvage}$ (p-value = 0.81). L'effet année a donc ensuite été pris en compte pour les années 2011, 2012 et 2014 uniquement, pour l'année 2013 les valeurs de $\Delta DF_{Sauvage}$ n'ont pas été ajustées. La variance au sein de chaque année est dûe à la variance intra-espèce (3.04) plutôt qu'inter-espèce (0) (Annexe Figure S1).

Le deuxième MLM modélise la différence entre les dates de pleine floraison de la variété référence Golden et les variétés cultivées ($\Delta DF_{Cultivé}$), selon l'année et le génotype des variétés de 2016 à 2019 (Tableau 2). L'effet année affecte significativement $\Delta DF_{Cultivé}$ (p-value < 2.0e-16) et ceux pour toutes les années. Pour les deux jeux de données, le ΔDF ajusté a été alors calculé en retirant les estimations des effets significatifs année. Le modèle 4 : (4) $DF_{Cultivé}$ ~année + (1 | génotype) + (1 | réplica : génotype), montre que l'effet aléatoire du génotype (1 | génotype) à une variance de 53.20 et que l'effet niché du réplica dans le génotype (1 | réplica : génotype) est de 0.50. La variabilité inter-génotypes influence la date de floraison et l'effet des réplicas est moindre.

Tableau 1 : Modèle linéaire mixte réalisé sur la différence entre les dates de pleine floraison de la variété Golden et des quatre espèces sauvages : *Malus baccata, Malus sylvestris, Malus sieversii, Malus orientalis* ($\Delta DF_{Sauvage}$), *N*=67 de 2011 à 2014

(2) $\Delta DF_{Sauvage}$ ~ année + (1 | espèce) + (1 | espèce : génotype)

Effets fixes					Effets aléatoires		
Groupe	Estimation	Ecart type	df	Pr(>|t|)	Groupe	Variance	Ecart type.
année			3	2.20e-16			
					espèce :		
(Intercept)	1.35	0.37	201.57	0.000374	génotype	3.04	1.74
2012	1.82	0.41	180.69	1.33e-05	espèce	0	0
2013	0.10	0.40	181.86	0.810058	Residual	4.44	2.12
2014	4.59	0.41	180.02	< 2e-16			

Tableau 2 : Modèle linéaire mixte réalisé sur la différence entre les dates de pleine floraison de la variété référence Golden et des variétés cultivées : *M. domestica* (ΔDF_C), *N*=235 de 2016 à 2019

(3) $\Delta DF_{Cultivé}$ ~ année + (1 | génotype)							
	Effets fixes				Effets aléatoires		
Group	Estimation	Ecart type	df	Pr(>|t|)	Groupe	Variance	Ecart type
année			3	2.20e-16			
(Intercept)	-7.15	0.50	615.17	2.00e-16	génotype	24.53	4.95
2017	4.67	0.53	683.66	2.46e-16			
2018	4.50	0.52	683.50	2.00e-16	Residual	30.58	5.53
2019	5.38	0.52	683.19881	2.00e-16			

b- Comparaison des ΔDFajustés sauvage ($\Delta DF_{A.Sauvage}$) vs cultivé ($\Delta DF_{A.Cultivé}$)

Un test de comparaison de moyenne avec ajustement de Bonferroni des p-values nous a permis de comparer les $\Delta DF_{ajustés}$ entre les différentes années (Figure 2). Les 4 espèces sauvages sont toutes plus précoces que la variétés Golden ($\Delta DF_{A.Sauvage} > 0$), les variétés cultivées sont plus tardives ($\Delta DF_{A.Cultivé} < 0$). Hormis pour l'année 2016, les $\Delta DF_{ajustés}$ des variétés cultivées sont significativement différentes de celles des sauvages (p-value < 0.01). Cette comparaison indique que les pommiers sauvages fleurissent avant les pommiers cultivés (Figure 2).

(4) Structuration & diversité génétique

a- Détection des clones

L'indice PI-HAT de l'outil PLINK a permis de calculer l'IPD entre paires d'individus et de mettre en évidence ceux étant proches génétiquement. Les paires ayant une valeur de PI-HAT > 0.90 sont considérées comme des clones. Le jeu de données "complet" (sauvages et cultivés) contenait huit paires de génotypes avec des PI-HAT élevés. Ces huit paires représentaient soit des "doublons" résultant d'erreurs d'étiquetage lors du greffage ou de la plantation, soit pour certains, des contaminations survenues lors de l'extraction de l'ADN (Tableau 3).

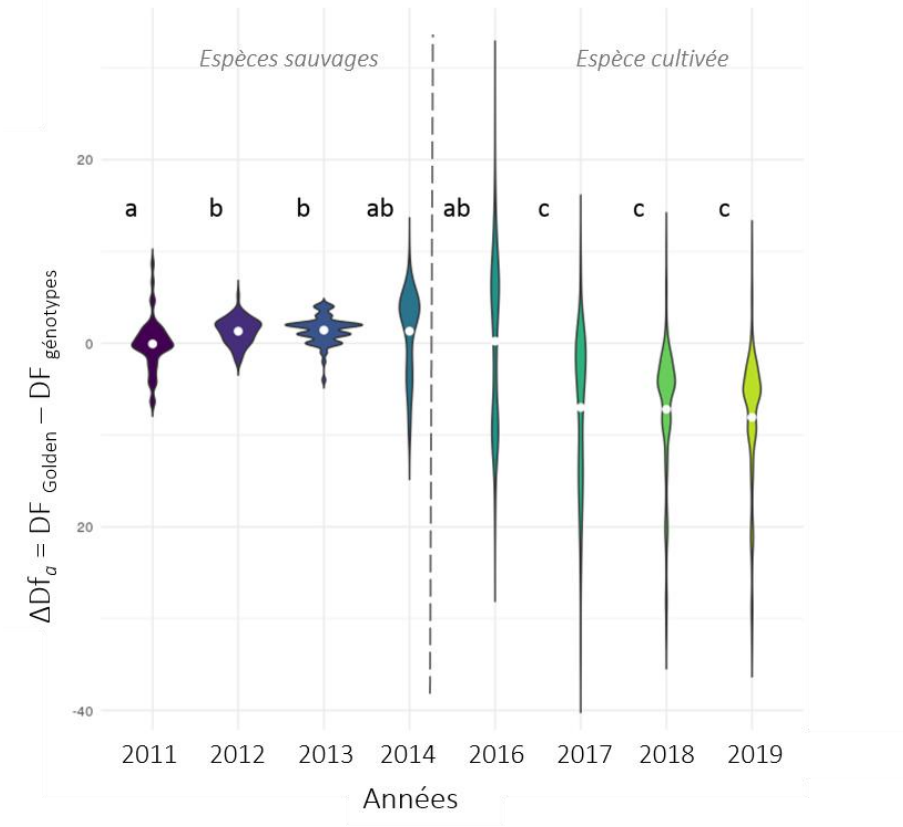


Figure 2 : Comparaison des $\Delta DF_{ajusté}$: différence entre les dates de pleine floraison de la variété référence Golden et des quatre espèces de pommiers sauvages (N=67) et du pommier cultivé (N=235), $\Delta DF_{ajustés}$ a été ajustée selon l'effet année (modélisation par un modèle linéaire mixte). Test de comparaison de moyenne deux à deux de Wilcoxon avec ajustement de Bonferroni des p-value, les différentes lettres (a, b, ab, c) indique que les valeurs du groupe sont significativement (p-value < 0.05) différentes d'un autre groupe.

En effet, concernant les trois paires de *M. domestica,* il n'existe pas dans la littérature de résultats similaires (Muranty et al. 2020). La probabilité que ces individus soient des clones est très faible. C'est notamment le cas pour la paire Anna - Gala, Anna est un hybride d'anciennes variétés israéliennes et Gala une variété française (Trainin et al. 2016). Une analyse par marqueurs microsatellites (marqueurs des répétitions de séquences simples "*SSR"*), d'un des deux parents sauvages des paires *M. baccata* révèlent des différences entre les individus (résultat non publié), mais nous ne disposons pas de données suffisantes pour conclure sur les paires des espèces sauvages. Ces individus ont été donc considérés comme clones et l'un des deux a ensuite été retiré pour les analyses présentées ci-après (Tableau 3, Figure 1 : étape A).

Tableau 3 Identité par descendance entre les individus les plus proches génétiquement dans le jeu de données "complet" (domestiqués et sauvages, N=342). Pour chaque paire, un individu a été retiré pour les analyses suivantes.

	Individu 1	**Individu 2**	**PI_HAT**	**Hypothèse**
M.domestica	Reinette-Clochard	X2361 (Reinette-Clochard)	0.9842	doublon
M. domestica	Anna	Gala	0.9803	Erreur ou contamination
M. domestica	Belrenne	X2640 (Reine des reinettes)	0.9803	Erreur ou contamination
M. domestica	X0048 (Borowitsky)	X8245 (Baguette violette)	0.9758	Erreur ou contamination
M. orientalis	MAL0786	MAL0786b	0.9751	Erreur ou contamination
M. domestica	Patte-de-Loup	X0710 (Patte de loup)	0.9662	doublon
M. baccata	MAL0023	MAL0779	0.9535	Erreur ou contamination
M. baccata	MAL0055	MAL0467	0.9171	Erreur ou contamination

b- Structuration des populations et définition des groupes

i) Core collection et espèces sauvages

Suivant le nombre de groupes (K), le jeu de données était structuré de la manière suivante. Pour *K*=2, seul *M. baccata* se distinguent des trois autres espèces. Pour K=3; *M. baccata, M. sylvestris* forment deux groupes différents, *M. sieversii* et *M. orientalis* sont regroupés ensemble. *Malus domestica* montre lui des signatures d'admixture avec *M. sylvestris* et *M. sieversii - M. orientalis* (résultats non montrés dans le rapport). Pour K=4 le logiciel FastStructure a révélé quatre groupes génétiques distincts séparant *M. domestica, M. baccata, M. sylvestris, et* regroupant toujours *M. sieversii - M. orientalis* (Figure 3A, Figure 1 : étape B). Ces deux dernières espèces *M. sieversii*/ et *M. orientalis*, restent indissociables même pour des valeurs de K plus élevées. Pour K>4, une légère sous structure a été détectée uniquement au sein de *M. domestica,* mais la différenciation génétique entre ces groupes reste très faible (analyses non montrées dans le rapport). Les résultats mis en évidence par FastStructure sont confirmés par l'ACP : les individus *M. sieversii* et *M. orientalis* apparaissent très proches génétiquement (Figure 3B, Figure 1 : étape C), et *M. domestica* présente une variation génétique plus élevée que les autres espèces, formant cependant un seul un groupe. L'analyse FastStructure permet de visualiser les individus admixtes entre les différentes espèces et de définir des groupes panmictiques qui suivent l'EHW. Ainsi le groupe extérieur de la *core collection M. floribunda* se différenciait génétiquement du groupe bleu majoritairement composé de *M. domestica* (flèche orange dans la Figure 3A). Certains individus ont été assignés à plus de 90% à un groupe différent de celui de leur espèce

taxonomique, ils sont indiqués par une flèche rouge dans la figure 3A. Une erreur d'identification sur le terrain expliquerait ces mauvaises assignations, ces individus ont été retirés du jeu de données pour la suite des analyses (Figure 1, étape D). Un seuil d'appartenance à un groupe est défini à 80%, les individus appartenant au groupe bleu sont définis comme *M. domestica*, au groupe orange *M. baccata*, au groupe vert *M. sylvestris* et au groupe violet : *M. sieversii - M. orientalis*. Les individus n'étant pas assignés à 80% à l'un des groupes sont définis comme *hybrides ou formes admixées* pour la suite des analyses. Ce seuil a été défini à partir de l'étude des distributions des assignations moyennes de chaque sous-espèces aux différents groupes (Annexe Figure S2). On observe en effet que plus de 80% des individus d'une sous-espèce appartiennent au même groupe, et cela pour K=4 et K=5. Au total, nous avons donc identifié quatre groupes génétiques distincts au sein du jeu de données complet (Tableau 4). Les indices de fixation (*Fst*) et les distances génétiques Dxy (Tableau 5, Figure 1: étape D) ont été calculées entre les quatre groupes détectés précédemment. *Malus sieversii* et *M. orientalis* on été séparés lors de ces analyses afin de vérifier la proximité de ces deux espèces (seuil d'assignation à l'espèces est de 80%). Le Fst entre ces deux espèces est très faible (0.029) et la distance Dxy la plus faible (223.22), les assigner à un même groupe est donc cohérent. Un arbre SplitsTree, réalisé à partir des distances Dxy, permet de visualiser les distances et relations génétiques entre les groupes (Figure 3C).

Tableau 4 : Composition des quatre groupes génétiques identifiés à l'aide du logiciel FastStructure chez les pommiers sauvages et le pommier cultivé (N=285, excluant 43 hybrides : individus assignés à moins de 80% dans un groupe donné) à l'aide de 5 154 SNP non liés.

	Nombre d'individus	Sous-espèces
Groupe 4	219	*M. domestica*
Groupe 2	14	*M. baccata*
Groupe 1	20	*M. sylvestris*
Groupe 3	32	*M. sieversii - M.orientalis*
Hybrides (exclus pour les analyses suivantes)	43	5 *M. baccata* , 4 *M. sieversii*, 4 *M. orientalis* - 30 *M. domestica*

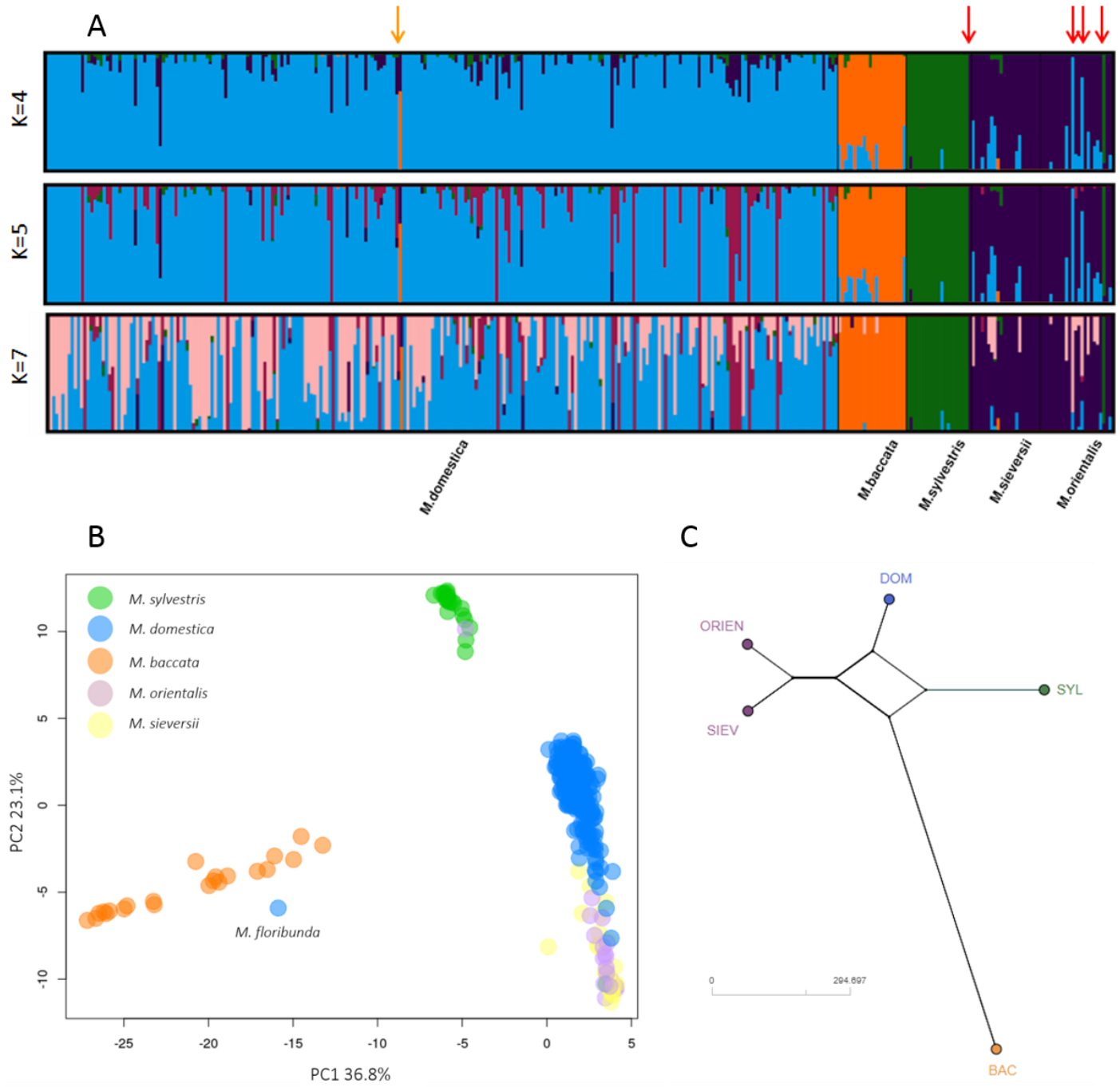


Figure 3 : A) Structuration génétique et admixture observées chez les quatre espèces sauvages (N=88, *Malus sylvestris*, *Malus sieversii* - *Malus orientalis*, *Malus baccata*) et les 254 variétés de pommier cultivé (*Malus domestica*), inférées avec FastStructure avec 5 154 SNP non liés. Les flèches rouges indiquent les individus mal assignés ou trop admixés, la flèche orange désigne *M. floribunda*, l'individu extra-groupe de la *core collection*. B) Analyse en composantes principales incluant les espèces sauvages et les variétés cultivées, avec les couleurs des groupes génétiques inférés avec FastStructure et 5 154 SNP. Les cercles représentent les génotypes, les cercles oranges appartiennent à l'espèce *M. baccata*, les verts : *M. sylvestris*, les bleus : *M. domestica*, les violets : *M.orientalis*, les jaunes : *M. sieversii*. C) Relations phylogénétiques à partir des distances génétique Dxy, entre les cinq groupes génétiques (détectées avec FastStructure avec 5 154 SNP). *M. dom*, *M. syl. M. bac* forment des groupes distincts. Le groupe génétique *sieversii*/*orientalis* forment un même groupe mais ont été séparés sur la figure.

Tableau 5 : Différenciation génétique (indice Fst et distance génétique) Dxy entre les quatre groupes génétiques identifiés à l'aide du logiciel FastStructure, noté que *M. sieversii* et *M. orientalis*, inférés sur 5 154 SNP.

FST | Dxy	*M. baccata*	*M. sylvestris*	*M. sieversii*	*M. orientalis*	*M. domestica*
M. baccata	-	1016.36	1016.24	1019.55	1019.17
M. sylvestris	0.58	-	652.73	654.44	473.38
M. sieversii	0.60	0.48	-	223.22	394.54
M. orientalis	0.61	0.491	0.029	-	393.80
M. domestica	0.53	0.23	0.17	0.17	-

i) Core collection

Les analyses suivantes ont pour but d'explorer une potentielle structure dans la *core collection* (*core collection* construite par (Lassois et al. 2016)) et entre variétés aux phénotypes contrastés (N=254). Les premiers résultats obtenus avec FastStructure suggèrent une structuration génétique dans la *core collection* (Figure 4A, Figure 1: étape B). Les Fst entre les groupes détectés sont relativement faibles et tous inférieurs à 0.1 (p-value < 2.2e-16): pour K=2, le Fst moyen est de 0.02; pour K=3, les Fst varient de 0.01 (*cluster* 1 - 3) à 0.08 (*cluster* 1 vs 2) et pour K=4, les valeurs de Fst sont comprises entre 0.02 (*cluster* 3 vs 4) et 0.09 (*cluster* 1 vs 3) (Annexe Tableau S1). L'analyse du pourcentage d'assignation de chaque individu aux groupes révèle que la majorité des individus sont fortement introgressés et très peu assignés à un groupe en particulier (Annexe Figure S3).

Les résultats de l'ACP (Figure 4B, Figure 1 : étape C) ne montrent aucune structure au sein du jeu de données. L'ensemble de ces résultats suggèrent que la *core collection* est très peu structurée génétiquement. Ces résultats sont concordants avec les analyses sur le jeu de données complet (sauvages+cultivé) suggérant une sous-structure très faible, qui peut ainsi être considérée comme négligeable dans les analyses suivantes.

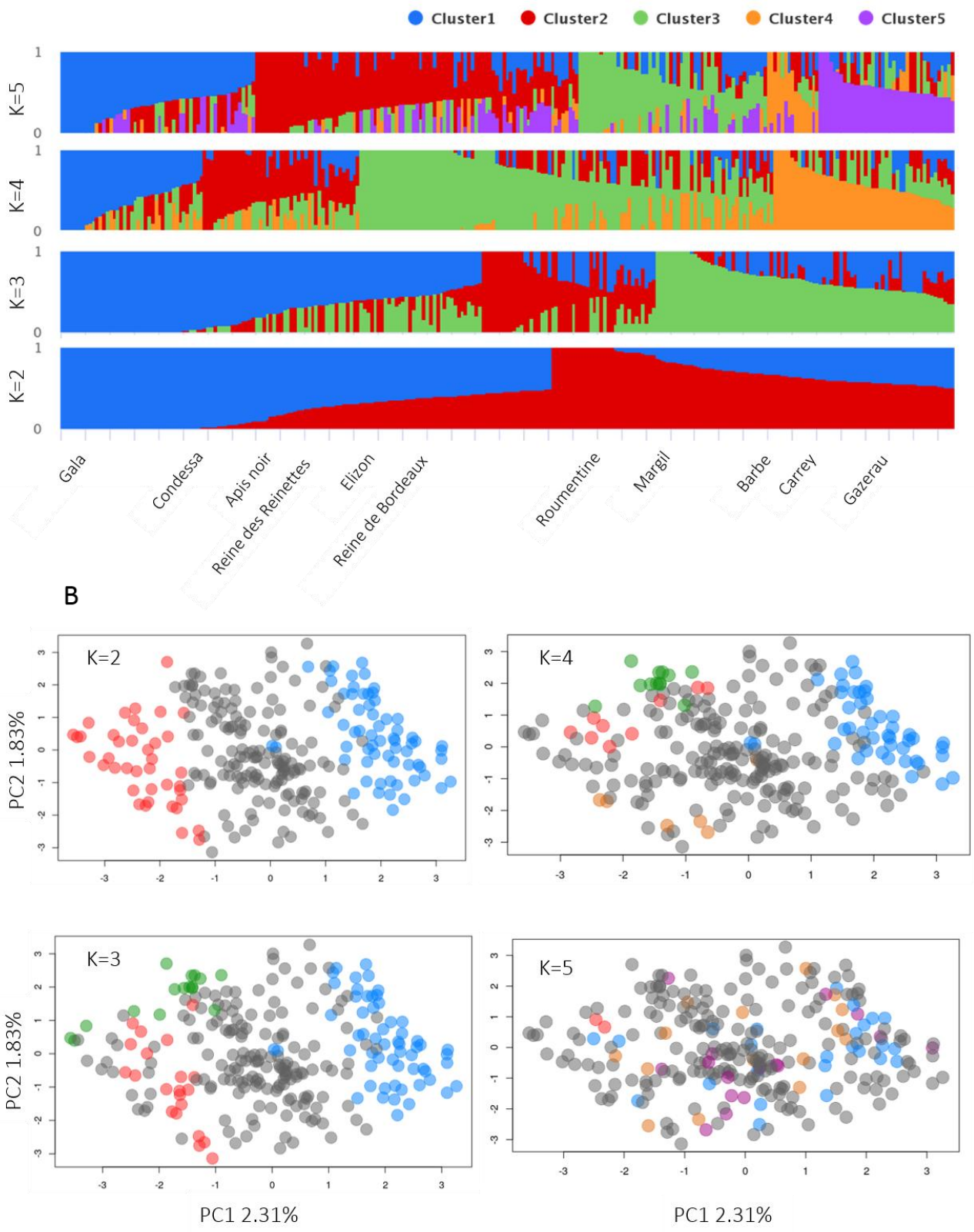


Figure 4 : A) Proportion d'admixture au sein de l'ensemble des variétés de pommier cultivé (*M. domestica*), inféré sur 796 SNP. B) Analyses en composantes principales (ACP) chez l'ensemble des variétés de pommier cultivé (*M. domestica*), inférées avec 796 SNP, selon les groupes inférés par l'analyse FastStructure (4A) et seuil d'appartenance à une groupe est de 80%, les individus gris sont hybrides (assignation inférieur à 80% à l'un des groupes génétiques).

1- Traces de sélection chez les pommiers sauvages et au cours de la domestication

Les analyses réalisées avec le logiciel PCAdapt nous ont permis de détecter des SNP potentiellement impliqués dans l'adaptation locale. Ces SNP sont ceux qui sont anormalement liés aux axes de l'ACP pris en compte dans l'analyse. Pour s'affranchir des effets confondants liés à la structuration des populations, nous n'avons pas pris en compte les SNP corrélés aux axes qui expliquent la structure génétique (Duforet-Frebourg et al. 2016) (Figure 5). Ce raisonnement a été appliqué pour les différents groupes analysés : pommier cultivé (N=248), espèces sauvages et espèce cultivée (N=327), espèces sauvages seules (N=79), *M. domestica – M. orientalis – M. sieversii* (N=288), *M. domestica – M. orientalis – M. sieversii – M. sylvestris* (N=308), les détails des ACP sont disponibles à cette adresse : https://gitlab.southgreen.fr/sidibebocs/domestication_selection_trace/wikis/pcadapt_results

Un total de 119 SNP ont été détectés sur le jeu de données complet incluant les espèces sauvages et le pommier cultivé. Il y a donc des SNP outlier dans les gènes associés à la floraison.

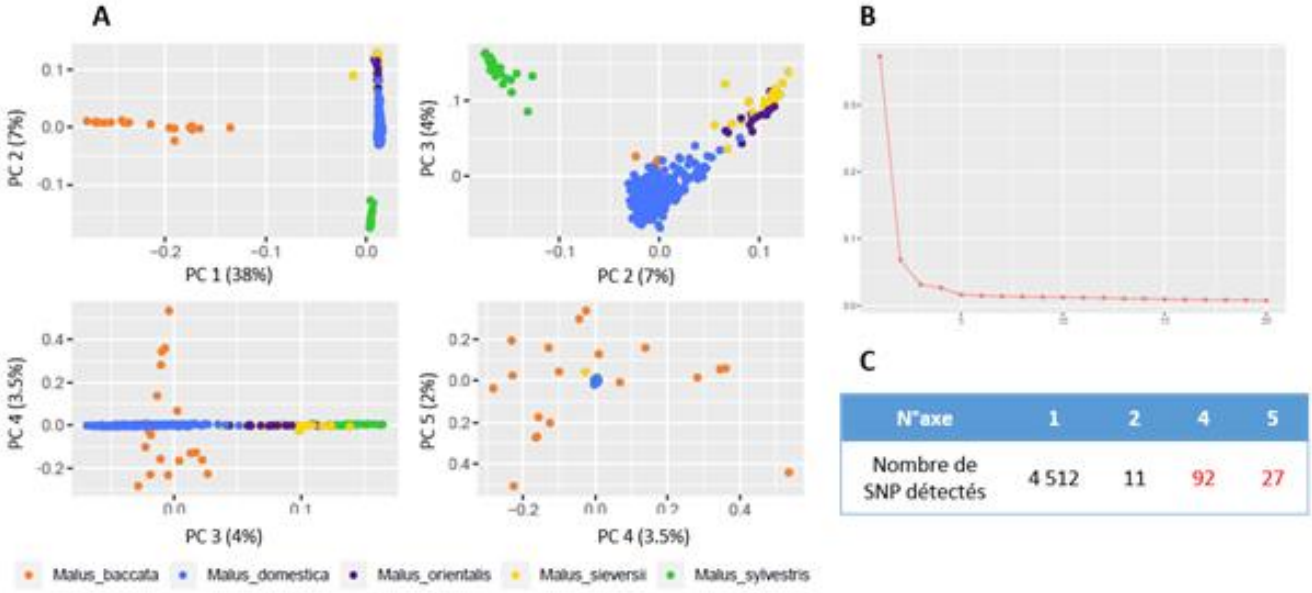


Figure 5 : Détection des SNP liés à l'adaptation locale avec le logiciel PCAdapt, analyse réalisée sur l'ensemble des espèces sauvages et cultivés: *Malus baccata, Malus sieversii, Malus orientalis, Malus sylvestris, Malus domestica* (N=342) sur 18 663 SNP. A) Analyse en composante principale réalisée sur 5 axes. B) Valeurs propres de chacun des axes de l'ACP. C) Nombre de SNP détectés, seuls, ceux corrélés aux axes 4-5 sont gardés, les axes 1 et 2 étant liés à la structure entre les espèces sont éliminés

Le tableau 6 regroupe le nombre des SNP liés à l'adaptation en fonction des groupes génétiques..

Tableau 6 : Nombre de SNP *outlier* détectés avec le logiciel PCAdapt (inféré sur 18 663 SNP) pour les groupes d'échantillons analysés chez les pommiers sauvages et le pommier cultivé.

	Espèces sauvages et cultivés (N=327)	**Espèces sauvages (N=79)**	**Espèce cultivée (N=248)**	***M. domestica M. orientalis M. sieversii (N=308)***	***M. domestica M. orientalis M. sieversii M. sylvestris (N=40)***	***M. orientalis M. sieversii*** **(N=40)**	***M. sylvestris (N=20)***	***M. baccata*** **(N=19)**
Nombre de SNP détectés	119	434	168	50	44	50	54	34

PCAdapt détecte les SNP anormalement liés ("outlier") aux axes de l'ACP, le nombre de SNP détecté dépend donc des ces axes et des espèces/échantillons analysées. La somme des SNP détectés par espèce n'est pas égale au nombre de SNP détectés sur l'ensemble de ces espèces car le pourcentage de variation expliqué par les axes des ACP ne sont plus les mêmes. Nous avons comparé les SNP détectés dans les différents groupes pour voir s'ils étaient propres à une espèce, ou à un complexe d'espèces (Figure 6). Entre les variétés cultivées et l'ensemble des espèces sauvages seul 1 SNP outlier est partagé (Figure 6A). Par ailleurs, 217 SNP outliers sont propres aux variétés cultivées.

Puis nous avons comparé les SNP *outliers* trouvé chez les variétés cultivées et les SNP trouvé chez les complexes d'espèces suivants : (1) Dom_Siev_Ori : *M. domestica, M. sieversii, M. orientalis* (N=288) et (2) Dom_Siev_Ori_Syl : *M. domestica, M. sieversii, M. orientalis, M. sylvestris* (N=308). Ces complexes d'espèces contiennent le pommier cultivé et ses plus proches parents sauvages (Figure 6B). Nous retrouvons **42** SNP outlier en commun entre les variétés cultivées et le complexe (1) (*M. sieversii* et *M. orientalis* sont les espèces sauvages les plus proches génétiquement des variétés cultivées, Figure 2). L'analyse du complexe (2) permet de détecter **21** SNP outlier en commun avec les variétés cultivées, ce complexe d'espèces inclut l'ancêtre du pommier sauvage (*M. sieversii*) et les espèces sauvages qui ont introgressé le pommier cultivé au cours de sa domestication (*M. orientalis, M. sylvestris*). Aucun de ces complexes d'espèces ne partage des SNPs sous sélection positive avec *M. baccata*, le parent sauvage le plus éloigné (Figure 2). Aucune des espèces sauvages

(prise individuellement) ne partagent de SNP avec les variétés cultivées hormis *M. sylvestris* qui en partage **11**.

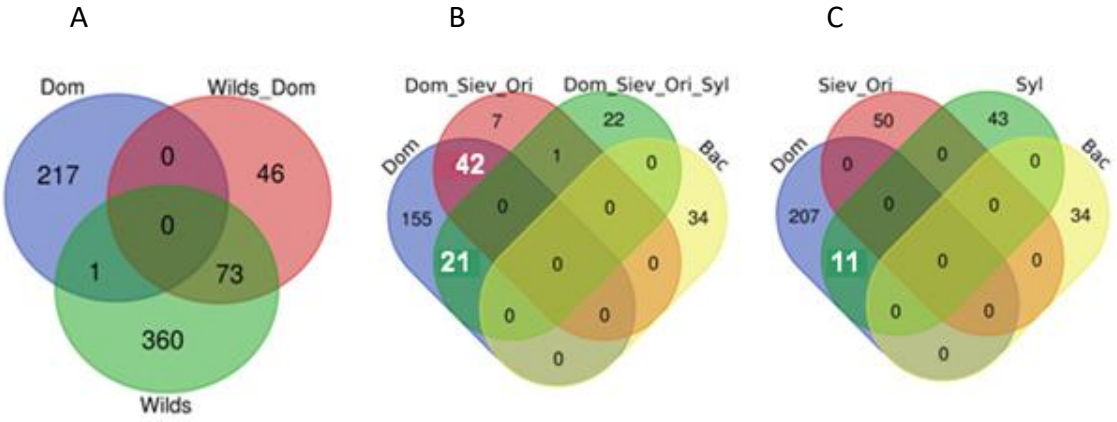


Figure 6 : Nombre de SNP outlier liés à l'adaptation locale communs entre différents groupes d'espèces détecté avec le logiciel PCAdapt en utilisant 18 663 SNP. **A**. Dom : *M. domestica* l'espèce cultivée (N=248), Wilds : toutes les espèces sauvages (*M. orientalis, M. sieversii, M. sylvestris, M. baccata)* (N=79), Wilds_Dom : espèces sauvages et cultivés. **B** Dom_Siev_Ori : *M. domestica, M. sieversii, M. orientalis* (N=288), Dom_Siev_Ori_Syl : *M. domestica, M. sieversii, M. orientalis, M. sylvestris* (N=308), Bac : *M. baccata* (N=19). **C** Siev_Ori : *M. sieversii, M. orientalis* (N=40), Syl : *M. sylvestris* (N=20).

L'outil de visualisation IGV nous a permis de comparer les séquences génétiques entre les variétés cultivés et les complexes d'espèces. Si le SNP peut être commun aux variétés cultivées et aux complexes sauvages, ce n'est pas forcément le même allèle qui a été sélectionné. Parmi les 42 SNP communs à *M. domestica* et au complexe (1), 35 SNP appartiennent au même gène du chromosome 8 : **MD08G1159600**, une histone méthyltransférase. Or, ce ne sont pas les mêmes allèles qui sont retrouvés chez les variétés cultivés et chez les deux espèces sauvages (Figure 7A). Les deux espèces sauvages sont homozygotes du même allèle que le génome de référence (pommier cultivé variété Golden Delicious GDDH1.3) (SNP en gris), 40% des individus des variétés cultivées sont homozygotes identiques à l'allèle du génome de référence GDDH1.3 (SNP en gris), 40% des individus sont hétérozygotes (SNP en bleus foncés) et à 20% sont homozygotes d'un autre allèle que celui du génome de référence (SNP en bleus clairs). Il y a donc eu fixation d'allèles différents à un même SNP. Les 42 SNP communs aux variétés cultivées et au complexe Dom_Siev_Ori appartiennent à deux gènes et comme montré dans la Figure 7A, la majorité des individus des variétés cultivées sont hétérozygotes ou homozygotes à ces SNP mais l'allèle est différent de celui du complexe sauvage (Annexe Tableau S2). En revanche pour les SNP communs aux variétés cultivées et au complexe Dom_Siev_Ori_Syl, les individus sont presque tous homozygotes du même allèle que celui du génome de référence pour toutes les

espèces (Annexe Tableau S2). Pour les 11 SNP communs aux variétés cultivées et à *M. sylvestris*, les individus des variétés cultivées sont majoritairement homozygotes de l'allèle du génome de référence mais ce n'est pas le cas des individus de *M. sylvestris* qui sont homozygotes d'un autre allèle, comme illustré Figure 7B avec un SNP se trouvant dans le gène **MD04G1214500** qui code une des protéine *Frigida* (Annexe Tableau S2).

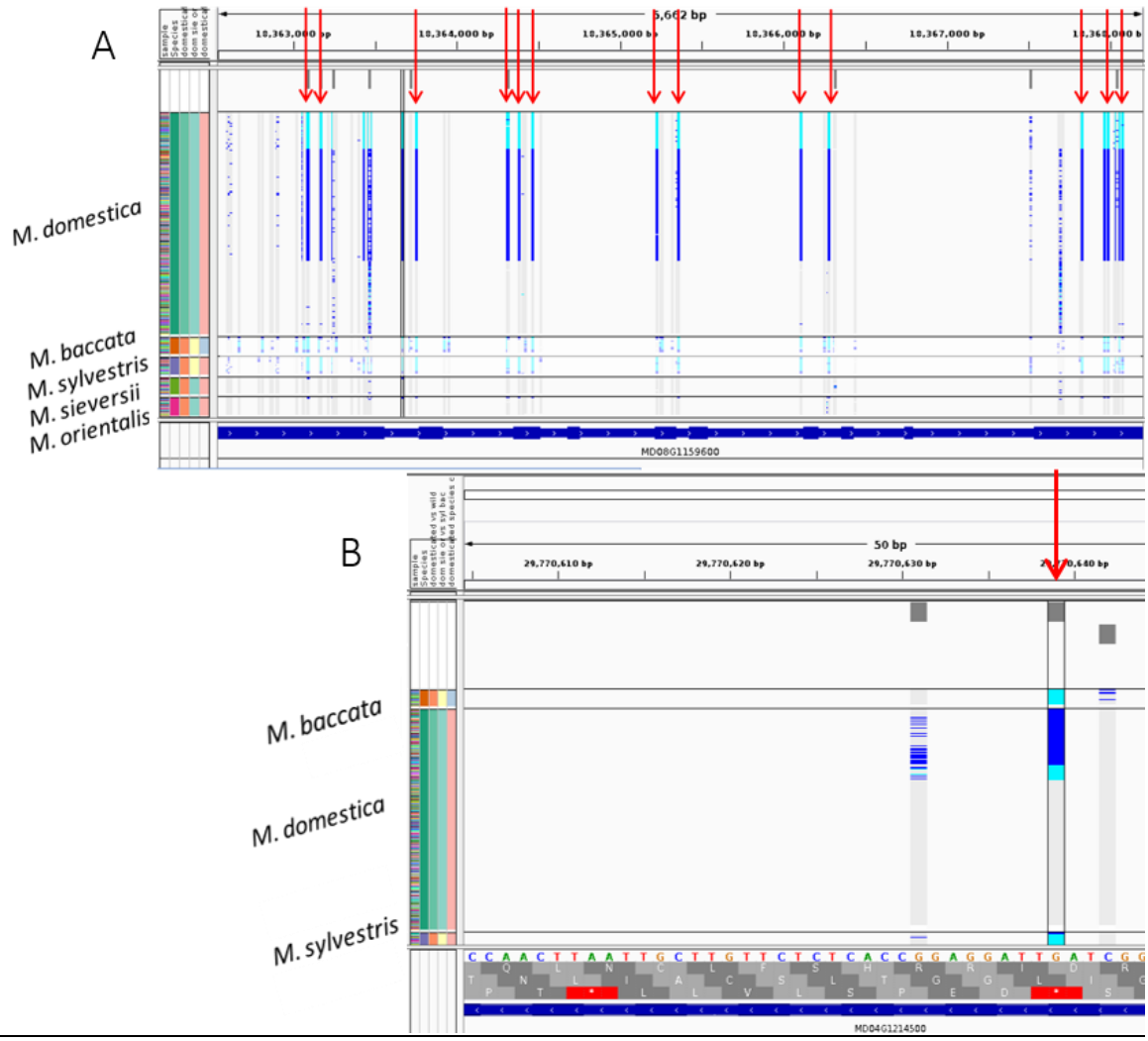


Figure 7 : Visualisation avec le logiciel Integrative Genomics Viewer (IGV) de SNP outlier détectés avec le logiciel PCAdapt chez les variétés de pommier cultivé et le complexe d'espèces(*Malus sieversii, Malus orientalis* et *Malus domestica*) à l'aide de 18 663 SNP. Les SNP en gris clair sont homozygotes pour l'allèle de référence, les SNP en bleu foncé sont hétérozygotes et les SNP en cyan sont homozygotes pour l'autre allèle. A) Gène MD08G1159600 du chromosome 8 (une histone méthyltransférase) dans lequel se trouve des SNP communs aux variétés cultivées et au complexe *M. sieversii, M. orientalis et M. domestic*a. B) Gène MD04G124500 du chromosome 4 (code la protéine Frigida), dans lequel se trouve un SNP outlier commun aux variétés cultivées et à *M. sylvestris.*

Discussion

Une floraison plus tardive pour les espèces cultivées ?

La date floraison est un trait phénologique important dans le cycle de vie d'une plante, il est impliqué dans la valeur adaptative et l'adaptation locale (Price et al. 2020). Contrairement à des traits tels que la taille des fruits et leurs qualités nutritionnelles, la floraison a été peu étudiée au cours de la domestication du pommier cultivé, *M. domestica*. Une première étape pour étudier la floraison au cours de la domestication a été de comparer les dates de floraison des quatre espèces sauvages *M. sieversii, M. orientalis, M. sylvestris* et *M. baccata* à l'espèce cultivée *Malus domestica*. Malgré l'hétérogénéité spatiale (les espèces sauvages sont plantées dans l'est de l'Allemagne et les variétés cultivées dans le sud de la France) et temporelle (date de floraison de 2011 à 2014 pour les espèces sauvages et 2016 à 2019 pour les pommiers cultivés), nous avons pu montrer que la variabilité des dates de floraison est assez importante pour toutes les espèces sauvages et le pommier cultivé (modèles (2) et (4), Tableau 1 et 2). La variation intra-espèce est même plus importante que la variation inter espèces chez les espèces sauvages (modèle (2)), la date de floraison varierait principalement avec le climat. Nos résultats montrent que la floraison des quatre espèces sauvages est plus précoce que la variété référence Golden, alors que l'espèce cultivée est plus tardive. La différence entre les dates de floraison relatives à Golden des espèces sauvages et celles des cultivées est significative (hormis pour l'année 2016) (Figure 2). Cette différence sauvage-cultivé suggère que la floraison tardive est le résultat d'un processus de sélection au cours de la domestication du pommier. Cependant au vue de l'hétérogénéité du jeu de données, ces résultats seraient à confirmer. Une nouvelle collaboration avec l'équipe allemande en charge de la collection de pommier sauvages va nous permettre d'avoir les dates de floraison des espèces sauvages en 2020, en même temps que celles des variétés cultivées et sur le même site. Nous pourrons alors mieux appréhender les effets des différents sites et des différentes années et ainsi mieux les modéliser pour en tenir compte.

Cinq espèces structurées en quatre groupes génétiques ?

Connaître la structuration des populations du jeu de données est une étape indispensable aux études de l'adaptation des populations. La structuration génétique des populations, résultant de la balance dérive-migration-mutation, peut être source de facteurs confondant avec la détection des faux positifs dans les études d'adaptation (Sul, Martin, et Eskin 2018; Tiffin et Ross-Ibarra 2014)*)*. Tout d'abord nous avons regardé la structure génétique de l'ensemble du

jeu de données : espèces sauvages et cultivés (N=342) sur 5 154 SNP. Nous avons détecté quatre groupes génétiques distincts : *M. baccata, M.sylvestris, M. domestica* et un groupe composé des deux espèces *M. orientalis et M. sieversii* (Figure 3). Ce résultat n'est pas inhabituel, le pommier sauvage d'Asie Centrale, *M. sieversii* et, le pommier sauvage caucasien, *M. orientalis,* sont souvent difficilement distinguables, que cela soit avec des données SSR (Cornille et al. 2012) ou de séquençage génomique (Duan et al. 2017). L'espèce *M. baccata* est très distincte des autres espèces sauvages et cultivés, ce qui est là aussi cohérent avec les précédents travaux sur la domestication du pommier cultivé (Cornille et al. 2012). En revanche, aucune trace d'introgressions de *M. sylvestris* n'est détectée avec les deux méthodes (FastStructure et ACP). Lors d'analyses précédentes basée sur des marqueurs microsatellites, *M. sylvestris* apparaissait comme très proche génétiquement du pommier cultivé, même plus proche que son ancêtre *M. sieversii* (Cornille et al. 2012). L'absence de trace d'introgressions de *M. sylvestris* dans le génome de *M. domestica* peut être dû à un échantillonnage insuffisant de la diversité génétique de *M. sylvestris*. Les individus présents dans le jeu de données ne seraient représentatifs que d'une partie de la diversité génétique de l'espèce, et ne permettraient pas de détecter les introgressions. Une autre explication serait que les introgressions du génome de *M. sylvestris* n'aient pas eu lieu dans les gènes liés à la floraison. La core collection incluant 242 génotypes à dessert représente 90% de la diversité allélique totale connue chez le pommier cultivé (Lassois et al. 2016). Elle a été construite de manière à ne pas présenter de structuration au sein des différents types de variétés (pomme à dessert, pomme à cidre). Notre jeu de données est constitué en grande majorité de pomme à dessert (trois variétés de pomme à cidre sont présentes sur 242). Nos résultats confirment l'absence de structure génétique au sein de cette core collection.

Des SNP liés à l'adaptation locale propres aux variétés cultivées ?

Pour détecter des balayages sélectifs liés à l'adaptation locale nous avons utilisé le logiciel PCAdapt. Ce logiciel implémente une méthode hiérarchique bayésìenne permettant de détecter des loci anormalement liés à la structuration de la population inférée par l'ACP. Les SNP *outlier* détectés par cette méthode sont considérés comme impliqués dans l'adaptation locale (Duforet-Frebourg, Bazin, et Blum 2014). Cette méthode de détection de signaux de sélection positive récente nous a permis de prendre en compte la structuration des populations résultant de l'histoire démographique complexe du pommier cultivé et de ses apparentés sauvages. En effet, nous avons pu nous affranchir des effets de structure en sélectionnant les axes des ACP non corrélés à la structuration des populations entre les espèces (Figure 5).

Cette méthode conservative nous a évité de détecter des faux-positifs, des SNP qui pourraient ne pas être liés à l'adaptation locale mais à la structure génétique et l'histoire évolutive du complexe Malus (Duforet-Frebourg, Bazin, et Blum 2014). Lors d'une première analyse sur l'ensemble des espèces cultivées et sauvages, 119 SNP ont été détectés, et sur toutes les espèces sauvages 434 SNP. Ces résultats préliminaires laissent à penser qu'il y aurait une pression de sélection naturelle sur les gènes liés à la floraison pour l'adaptation locale chez les espèces sauvages. Mais nous ne disposons pas d'un jeu de données suffisant (N=79) pour explorer une telle hypothèse.

Nous nous sommes ensuite intéressés aux SNP détectés au sein des variétés cultivées uniquement: 218 SNP ont été détectés (Figure 6.A). Certains de ces variants sont également détectés chez certaines espèces (*M. sylvestris* Figure 6.C) ou complexe d'espèces : (**1**) *M. sieversii - M. orientalis - M. domestica* et (**2**) *M. sieversii - M. orientalis - M. sylvestris - M. domestica*. Ces complexes d'espèces sont constitués des variétés cultivées, de leurs ancêtres (*M. sieversii - M. orientalis)* et des espèces qui les ont introgressées (*M. sylvestris),* 42 SNP outliers sont communs aux variétés cultivées et au complexe (**1**) et 21 SNP outliers concernent le complexe (**2**). Aucun des ces deux complexes ne partagent de SNP avec *M. baccata*, l'espèce sauvage la plus éloignée génétiquement (Figure 6.B). Nos résultats montrent des traces de sélection positive à deux niveaux, le premier au sein des sauvages (*sieversii - orientalis - sylvestris)* avec les 42 et 21 SNPs en commun et le deuxième niveau correspond à des traces de sélection au cours de la domestication. Il y aurait donc eu une sélection naturelle peu avant la domestication sur ces SNP des gènes de la floraison impliqué dans l'adaptation locale mais après la divergence avec l'espèce *M. baccata*. De tels résultats avaient déjà été trouvés sur les gènes impliqués dans la taille des fruits (Yao et al. 2015). En effet, le transposon MIR miRNA 172p impliqué dans la taille des fruits chez le pommier est fixé chez les espèces sauvages *M. sieversii, M. orientalis, M. sylvestris* et *M. domestica*, cette fixation a eu lieu avant la domestication des pommiers.

Les résultats obtenus avec le logiciel IGV nous renseignent un peu plus sur la distribution de la variabilité allélique pour chacun des SNP outliers communs entre variétés cultivées et pommiers sauvages. Concernant les SNP communs aux variétés cultivées et au complexe (2), ce sont les mêmes allèles qui ont été fixés (annexe, Tableau S2), la plus grande majorité des variétés cultivées sont homozygotes pour le même allèle que les trois espèces sauvages. Ce résultat va dans le sens d'une fixation pré-domestication de l'allèle. En revanche, ce n'est pas

le cas pour les SNP communs aux variétés cultivées et au complexe (**1**). Environ la moitié des variétés cultivées sont hétérozygotes ou homozygotes mais pour un allèle différent des espèces sauvages *M. sieversii* et *M. orientalis*, ce qui va dans le sens d'une sélection au cours de la domestication (pour certaines variétés du moins). Le même cas de figure se présente pour les SNP communs aux variétés cultivées et à *M. sylvestris*, les mêmes SNP liés à l'adaptation locale sont détectés mais ce n'est pas le même allèle qui est fixé (Figure 7B et Tableau S2). Une étude plus poussé de ces allèles, notamment sur l'état ancestral, permettrait de conforter l'hypothèse d'une fixation pré-domestication et la sélection au cours de la domestication.

Confronter nos résultats avec une étude en génomique d'association (*EGA*) sur notre jeu de données nous permettrait de faire un lien entre ces variants génétiques et le phénotype des différentes espèces afin de savoir si ces différences alléliques sont impliquées dans les différences phénologiques observées. Une précédente EGA réalisé par l'équipe a déjà été faite sur notre jeu de données mais au stade phénologique "date de débourrement" (Ousmane 2019, non publié). Sept gènes candidat ont été trouvés en commun entre cette EGA et nos analyses PCAdapt. Ces gènes sont les suivants: MD08G1004400 (LSD1-like), MD09G1052600, MD12G1228900 (Frigida-like), MD12G1229100 (Frigida-like), MD15G1003800 (LSD1-like), MD15G1008300 (bZIP), MD16G1272200 (Transducin/WD40). Le gène candidat situé sur le chromosome 15, MD15G1008300, serait l'orthologue du gène *FD* d'Arabidopsis (At4g35900), qui est un facteur de transcription clé dans la floraison (Abe 2005).

Perspectives

Nous avons détecté des variants génétiques lié à l'adaptation dans des gènes de floraison chez les pommiers sauvages et cultivés, caractère clé en agronomie, tout en nous affranchissant de faux positifs induits par des effets de structuration des populations. Ces résultats seront confrontés aux EGA déjà réalisées au laboratoire et restant à faire comme la GWAS au climat afin de les mettre en perspectives des futurs programmes d'amélioration variétale des pommiers. Il est à noter que notre méthode ne nous a permis de détecter que des signaux de sélection positive forts (un petit nombre de gènes à effet majeur) et non la détection de signaux de sélection polygénique (un grand nombre de gènes à faible effet), or ces derniers peuvent également jouer un rôle dans l'adaptation locale (Csilléry et al. 2018). Une étude de tels signaux de sélection permettrait de compléter et de conforter nos résultats sur la sélection de variants génétiques impliqués dans l'adaptation locale au cours de la domestication du

pommier cultivé. Enfin, en détectant des traces de sélection positive au sein des pommiers sauvages et cultivés, notre étude pose la question de connaitre l'importance des adaptations locales dans les sauvages par rapport à la sélection de ce trait phénologique au cours de la domestication. L'examen de cette question nécessite d'analyser un échantillonnage représentant la diversité des populations sauvages.

Bibliographie

Abe, M. 2005. « FD, a BZIP Protein Mediating Signals from the Floral Pathway Integrator FT at the Shoot Apex ». *Science* 309 (5737): 1052-56. https://doi.org/10.1126/science.1115983.

Allard, Alix, et 9 co-auteurs, 2016. « Detecting QTLs and Putative Candidate Genes Involved in Budbreak and Flowering Time in an Apple Multiparental Population ». *Journal of Experimental Botany* 67 (9): 2875-88. https://doi.org/10.1093/jxb/erw130.

Andrés, Fernando, et George Coupland. 2012. « The Genetic Basis of Flowering Responses to Seasonal Cues ». *Nature Reviews Genetics* 13 (9): 627-39. https://doi.org/10.1038/nrg3291.

Bates, Douglas, Martin Mächler, Ben Bolker, et Steve Walker. 2015. « Fitting Linear Mixed-Effects Models Using Lme4 ». *Journal of Statistical Software* 67 (1): 1-48. https://doi.org/10.18637/jss.v067.i01.

Chuine, Isabelle, et Elisabeth G. Beaubien. 2001. « Phenology Is a Major Determinant of Tree Species Range ». *Ecology Letters* 4 (5): 500-510. https://doi.org/10.1046/j.1461-0248.2001.00261.x.

* Cet article fait le lien entre les traits phénologiques et la valeur sélectives d'une plante, ainsi que l'implication de ces traits dans l'adaptation locale

Cornille, Amandine, Tatiana Giraud, Marinus J.M. Smulders, Isabel Roldán-Ruiz, et Pierre Gladieux. 2014. « The Domestication and Evolutionary Ecology of Apples ». *Trends in Genetics* 30 (2): 57-65. https://doi.org/10.1016/j.tig.2013.10.002.

* Cette review traite de la domestication du pommier cultivé et de son l'histoire évolutive, et des nombreuses introgressions qu'il a subi au cours de la domestication

Cornille, Amandine, Pierre Gladieux, Marinus J. M. Smulders, Isabel Roldán-Ruiz, François Laurens, Bruno Le Cam, Anush Nersesyan, et al. 2012. « New Insight into the History of Domesticated Apple: Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties ». Édité par Rodney Mauricio. *PLoS Genetics* 8 (5): e1002703. https://doi.org/10.1371/journal.pgen.1002703.

Csilléry, Katalin, Alejandra Rodríguez-Verdugo, Christian Rellstab, et Frédéric Guillaume. 2018. « Detecting the Genomic Signal of Polygenic Adaptation and the Role of Epistasis in Evolution ». *Molecular Ecology* 27 (3): 606-12. https://doi.org/10.1111/mec.14499.

Duan, Naibin, Yang Bai, Honghe Sun, Nan Wang, Yumin Ma, Mingjun Li, Xin Wang, et al. 2017. « Genome Re-Sequencing Reveals the History of Apple and Supports a Two-Stage Model for Fruit Enlargement ». *Nature Communications* 8 (1): 249. https://doi.org/10.1038/s41467-017-00336-7.

Duforet-Frebourg, Nicolas, Eric Bazin, et Michael G.B. Blum. 2014. « Genome Scans for Detecting Footprints of Local Adaptation Using a Bayesian Factor Model ». *Molecular Biology and Evolution* 31 (9): 2483-95. https://doi.org/10.1093/molbev/msu182.
*Cet article détaille la méthode implémenté dans le logiciel PCAdapt et les résultats obtenus selon le modèle de sélection (modèle en îles, modèle de divergence)

Duforet-Frebourg, Nicolas, Keurcien Luu, Guillaume Laval, Eric Bazin, et Michael G.B. Blum. 2016. « Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data ». *Molecular Biology and Evolution* 33 (4): 1082-93. https://doi.org/10.1093/molbev/msv334.

Excoffier, L. and H.E. L. Lischer (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 10: 564-567.

Gaut, Brandon S., Concepción M. Díez, et Peter L. Morrell. 2015. « Genomics and the Contrasting Dynamics of Annual and Perennial Domestication ». *Trends in Genetics* 31 (12): 709-19. https://doi.org/10.1016/j.tig.2015.10.002.
*Cette review concerne la domestication des espèces pérennes, traitant du coût de la domestication et des particularités liées aux espèces pérennes.

Gros-Balthazard, Muriel, Guillaume Besnard, Gautier Sarah, Yan Holtz, Julie Leclercq, Sylvain Santoni, Daniel Wegmann, Sylvain Glémin, et Bouchaib Khadari. 2019. « Evolutionary Transcriptomics Reveals the Origins of Olives and the Genomic Changes Associated with Their Domestication ». *The Plant Journal* 100 (1): 143-57. https://doi.org/10.1111/tpj.14435.

Höfer, Monika, Mohamed Ali Mohamed Saad Eldin Ali, Jörg Sellmann, et Andreas Peil. 2014. « Phenotypic Evaluation and Characterization of a Collection of Malus Species ». *Genetic Resources and Crop Evolution* 61 (5): 943-64. https://doi.org/10.1007/s10722-014-0088-3.

D.H. Huson and D. Bryant, Application of Phylogenetic Networks in Evolutionary Studies, Molecular Biology and Evolution, 23(2):254-267, 2006. software available from www. splitstree.org,

Jombart, Thibaut, et Ismaïl Ahmed. 2011. « Adegenet 1.3-1: New Tools for the Analysis of Genome-Wide SNP Data ». *Bioinformatics* 27 (21): 3070-71. https://doi.org/10.1093/bioinformatics/btr521.

Kozarewa, Iwanka, Javier Armisen, Andrew F. Gardner, Barton E. Slatko, et C. L. Hendrickson. 2015. « Overview of Target Enrichment Strategies ». *Current Protocols in Molecular Biology* 112 (1): 7.21.1-7.21.23. https://doi.org/10.1002/0471142727.mb0721s112.

Lassois, Ludivine, Caroline Denancé, Elisa Ravon, Arnaud Guyader, Rémi Guisnel, Laurence Hibrand-Saint-Oyant, Charles Poncet, Pauline Lasserre-Zuber, Laurence Feugey, et Charles-Eric Durel. 2016. « Genetic Diversity, Population Structure, Parentage Analysis, and Construction of Core Collections in the French Apple Germplasm Based on SSR Markers ». *Plant Molecular Biology Reporter* 34 (4): 827-44. https://doi.org/10.1007/s11105-015-0966-7.

Li YL, Liu JX (2018) StructureSelector: A web based software to select and visualize the optimal number of clusters using multiple methods.Molecular Ecology Resources, 18:176–177

Meyer, Rachel S., Ashley E. DuVal, et Helen R. Jensen. 2012. « Patterns and Processes in Crop Domestication: An Historical Review and Quantitative Analysis of 203 Global Food Crops ». *New Phytologist* 196 (1): 29-48. https://doi.org/10.1111/j.1469-8137.2012.04253.x.

Miller, Allison J., et Briana L. Gross. 2011. « From Forest to Field: Perennial Fruit Crop Domestication ». *American Journal of Botany* 98 (9): 1389-1414. https://doi.org/10.3732/ajb.1000522.

Muranty, Hélène, Caroline Denancé, Laurence Feugey, Jean-Luc Crépin, Yves Barbier, Stefano Tartarini, Matthew Ordidge, et al. 2020. « Using Whole-Genome SNP Data to Reconstruct a Large Multi-Generation Pedigree in Apple Germplasm ». *BMC Plant Biology* 20 (1): 2. https://doi.org/10.1186/s12870-019-2171-6.

Pfeifer B, Wittelsbuerger U, Ramos-Onsins SE, Lercher MJ (2014). "PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R." *Molecular Biology and Evolution*, **31**, 1929-1936. doi: 10.1093/molbev/msu136.

Price, Nicholas, Lua Lopez, Adrian E. Platts, et Jesse R. Lasky. 2020. « In the Presence of Population Structure: From Genomics to Candidate Genes Underlying Local Adaptation ». *Ecology and Evolution* 10 (4): 1889-1904. https://doi.org/10.1002/ece3.6002.

Privé, Florian, Keurcien Luu, Bjarni J. Vilhjálmsson, et Michael G. B. Blum. s. d. « Performing Highly Efficient Genome Scans for Local Adaptation with R Package Pcadapt Version 4 ». *Molecular Biology and Evolution*. Consulté le 2 mai 2020. https://doi.org/10.1093/molbev/msaa053.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. « PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses ». *American Journal of Human Genetics* 81 (3): 559-75. https://doi.org/10.1086/519795.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical, https://www.R-project.org/

Raj Anil, Matthew Stephens, and Jonathan K. Pritchard. *fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets* , (Genetics) June 2014 197:573-589 https://doi.org/10.1534/genetics.114.164350

Robinson, J., Thorvaldsdóttir, H., Winckler, W. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29,** 24–26 (2011). https://doi.org/10.1038/nbt.1754

Srikanth, Anusha, et Markus Schmid. 2011. « Regulation of Flowering Time: All Roads Lead to Rome ». *Cellular and Molecular Life Sciences* 68 (12): 2013-37. https://doi.org/10.1007/s00018-011-0673-y.

Sul, Jae Hoon, Lana S. Martin, et Eleazar Eskin. 2018. « Population Structure in Genetic Studies: Confounding Factors and Mixed Models ». Édité par Gregory S. Barsh. *PLOS Genetics* 14 (12): e1007309. https://doi.org/10.1371/journal.pgen.1007309.

Tiffin, Peter, et Jeffrey Ross-Ibarra. 2014. « Advances and Limits of Using Population Genetics to Understand Local Adaptation ». *Trends in Ecology & Evolution* 29 (12): 673-80. https://doi.org/10.1016/j.tree.2014.10.004.
*Cet article évoque les limites liés à la structuration des populations dans la détéction de variants génétiques impliqué dans l'adaptation locale

Trainin, Taly, Matat Zohar, Einav Shimoni-Shor, Adi Doron-Faigenboim, Irit Bar-Ya'akov, Kamel Hatib, Noa Sela, Doron Holland, et Tal Isaacson. 2016. « A Unique Haplotype Found in Apple Accessions Exhibiting Early Bud-Break Could Serve as a Marker for Breeding Apples with Low Chilling Requirements ». *Molecular Breeding* 36 (11): 158. https://doi.org/10.1007/s11032-016-0575-7.

Urrestarazu, Jorge, Hélène Muranty, Caroline Denancé, Diane Leforestier, Elisa Ravon, Arnaud Guyader, Rémi Guisnel, et al. 2017. « Genome-Wide Association Mapping of Flowering and Ripening Periods in Apple ». *Frontiers in Plant Science* 8 (novembre): 1923. https://doi.org/10.3389/fpls.2017.01923.

Verhoeven, K. J. F., H. Poorter, E. Nevo, et A. Biere. 2008. « Habitat-Specific Natural Selection at a Flowering-Time QTL Is a Main Driver of Local Adaptation in Two Wild Barley Populations ». *Molecular Ecology*, juin, ???-??? https://doi.org/10.1111/j.1365-294X.2008.03847.x.

Wang, Jing, Jihua Ding, Biyue Tan, Kathryn M. Robinson, Ingrid H. Michelson, Anna Johansson, Björn Nystedt, et al. 2018. « A Major Locus Controls Local Adaptation and Adaptive Life History Variation in a Perennial Plant ». *Genome Biology* 19 (1): 72. https://doi.org/10.1186/s13059-018-1444-y.

Yao, Jia-Long, Juan Xu, Amandine Cornille, Sumathi Tomes, Sakuntala Karunairetnam, Zhiwei Luo, Heather Bassett, et al. 2015. « A *MicroRNA* Allele That Emerged Prior to Apple Domestication May Underlie Fruit Size Evolution ». *The Plant Journal* 84 (2): 417-27. https://doi.org/10.1111/tpj.13021.

ANNEXE

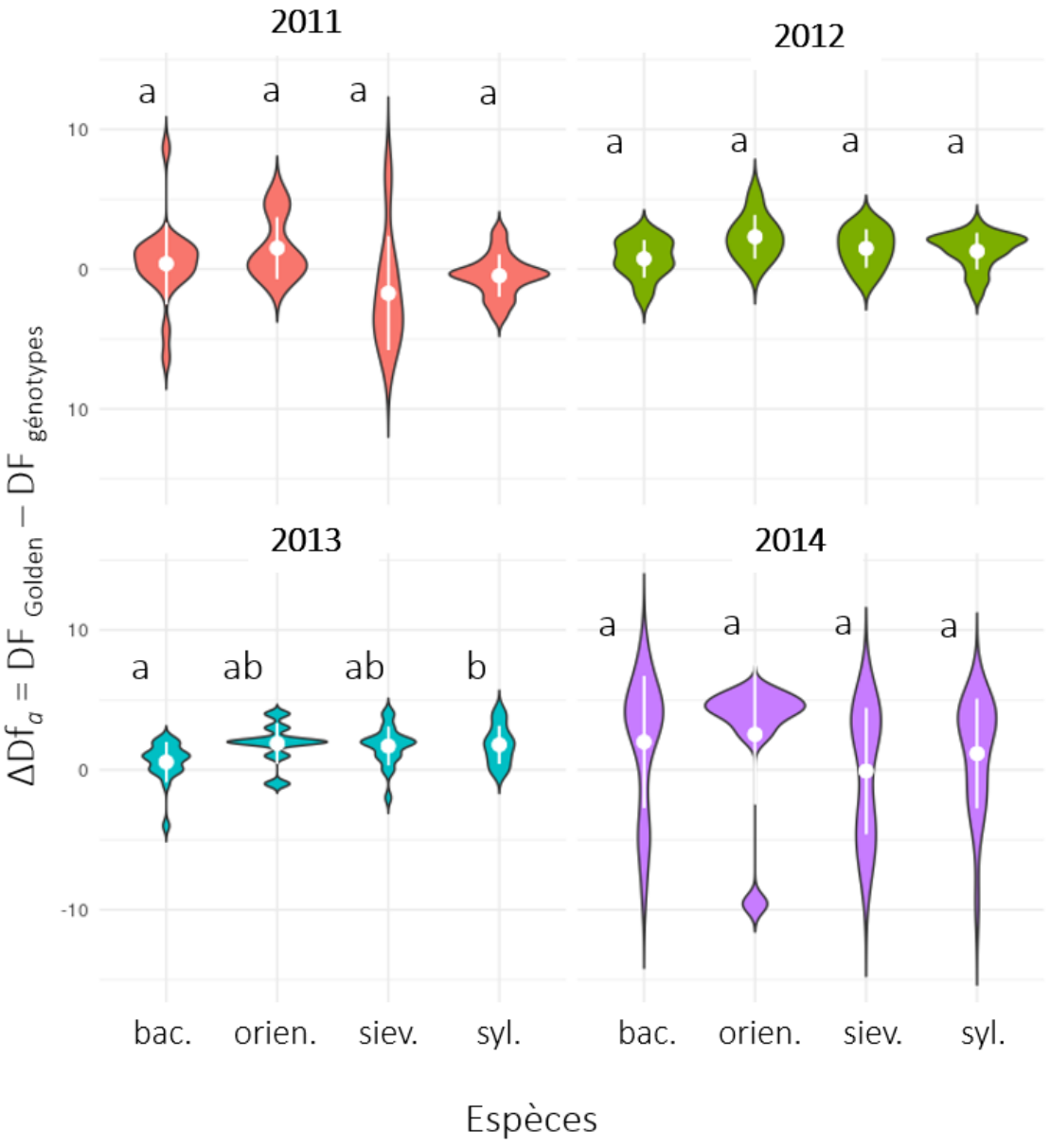


Figure S1: Comparaison des $\Delta DF_{ajusté}$: différence entre les dates de pleine floraison de la variété Golden et des espèces sauvages selon l'année, bac : *Malus baccata*, orien : *Malus orientalis*, siev : *Malus. sieversii*, syl : *Malus sylvestris* (N=67), $\Delta DF_{ajusté}$ a été ajustée selon l'effet année (modélisation par un modèle linéaire mixte). Test de comparaison de moyenne deux à deux de wilcoxon avec ajustement de bonferroni des p-value, les différentes lettres (a, b, ab , c) indique que les valeurs du groupe sont significativement différentes d'un autre groupe (p-value < 0.05).

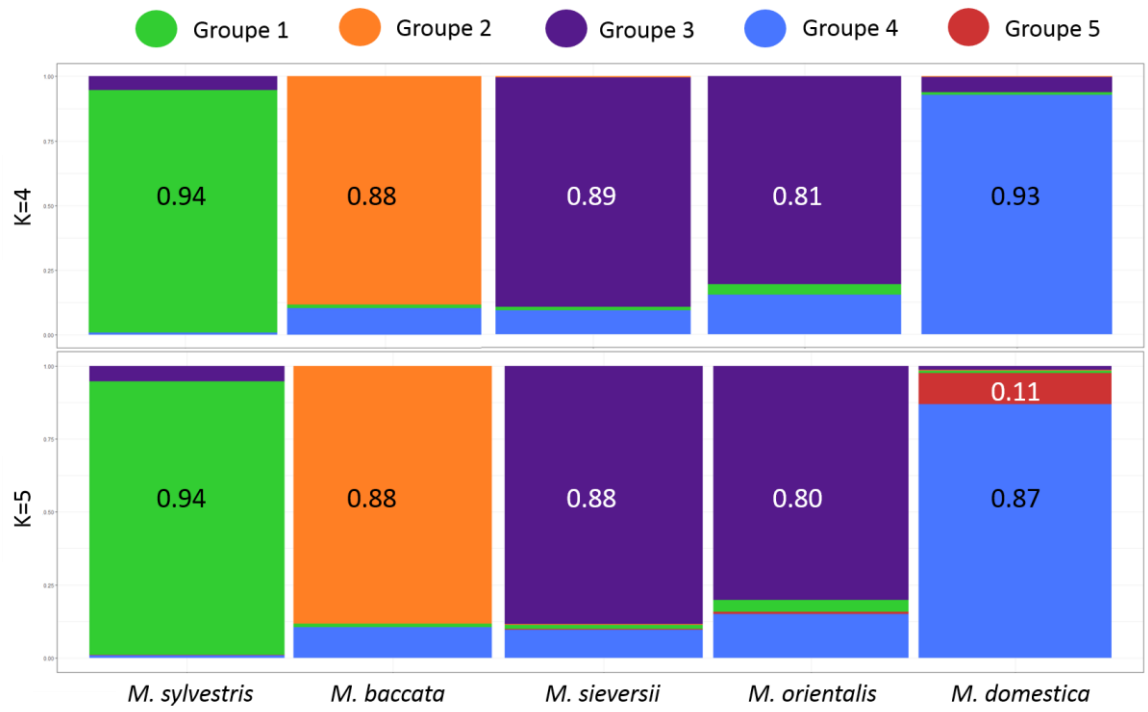


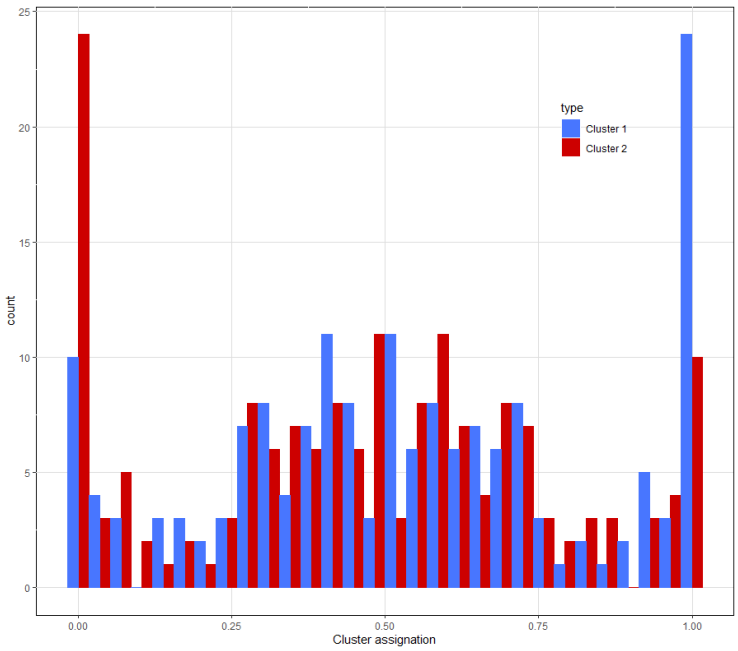
Figure S2: Pourcentage moyen d'assignation des différentes espèces à un groupe inféré par FastStructure avec 5 154 SNP, chez les espèces sauvages et les variétés domestiquées.

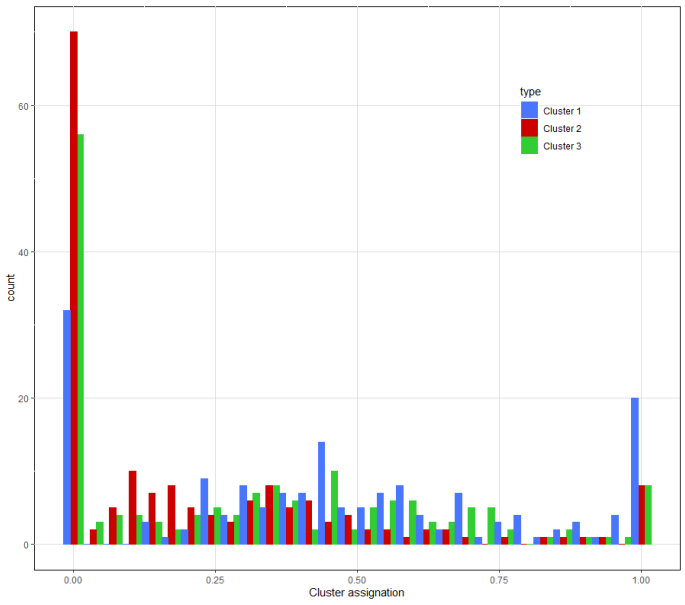
Tableau S1 : Différenciation génétique (Fst) entre groupes de *M. domestica* identifiés à l'aide du logiciel FastStructure, inférés sur 18 663 SNP

K2	Cluster 1
Cluster 2	0.0204624

K3	Cluster 1	Cluster 2
Cluster 1	na	na
Cluster 2	0.0774284	na
Cluster 3	0.0144004	0.0626694

K4	Cluster 1	Cluster 2	Cluster 3
Cluster 1	na	na	na
Cluster 2	0.0795258	na	na
Cluster 3	0.0929888	0.0259035	na
Cluster 4	0.068646	0.0373891	0.0240664





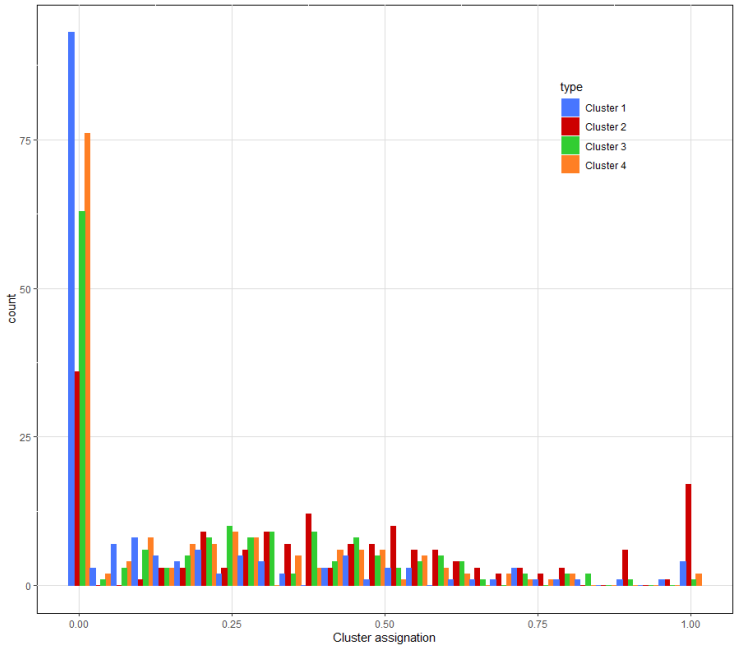


Figure S3 : Distribution de la proportion d'assignation de chaque individus inféré par fast structure pour la core collection de pommier cultivé pour 796 SNP pour a) K=2 b) K=3 and c) K=4 Nombre d'individus selon leur pourcentage admixture à un cluster, chez les variétés française.

Tableau S2: Visualisation avec le logiciel IGV de SNP liés à l'adaptation locale, détectés avec le logiciel PCAdapt chez les variétés cultivées et le complexe d'espèce: *Malus sieversii, Malus orientalis* et *Malus domestica*. PCAdapt a été inféré sur 18 663 SNP. *Ho.* ref homozygote de l'allèle de référence, *He* sont hétérozygotes et*Ho.non ref* sont homozygote de l'autre allèle. A) SNP communs aux variétés cultivées et au complexe *Malus sieversii, Malus orientalis* et *Malus domestica*. B) SNP communs aux variétés cultivées et au complexe *Malus sieversii, Malus orientalis, Malus sylvestris* et *Malus domestica*. C) SNP commun au variétés cultivées et à M. sylvestris.

A.

Nom du gène	nb SNP	*Malus domestica*			*Malus orientalis*			*Malus sieversii*		
		Ho. ref	He	Ho. non ref	Ho. ref	He	Ho. Non ref	Ho. ref	He	Ho. Non ref
MDG08G1159600	35	40%	40%	20%	100%	0%	0%	100%	0%	0%
MD00G1041900	5	30%	50%	20%	100%	0%	0%	100%	0%	0%

B.

Nom du gène	nb SNP	*Malus domestica*			*Malus orientalis – Malus sieversii*			*Malus sylvestris*		
		Ho. ref	He	Ho. non ref	Ho. ref	He	Ho. Non ref	Ho. ref	He	Ho. Non ref
MD16G1002500	1	90%	10%		100%	0%	0%	100%	0%	0%
MD01G1036900	1	95%	5%		100%	0%	0%	100%	0%	0%
MD03G1112700	1	90%	10%		100%	0%	0%	100%	0%	0%
MD03G1134200	1	90%	10%		100%	0%	0%	100%	0%	0%
MD03G1139200	2	90%	10%		100%	0%	0%	100%	0%	0%
MD07G1075300	8	90%	10%		100%	0%	0%	100%	0%	0%
MD07G1288500	1	90%	10%		100%	0%	0%	100%	0%	0%
MD08G1159600	1	40%	50%	10%	100%	0%	0%	100%	0%	0%

C.

Nom du gène	nb SNP	*Malus domestica*			*Malus sylvestris*		
		Ho. ref	He	Ho. non ref	Ho. ref	He	Ho. Non ref
MD04G1199600	3	60%	30%	10%	10%	10%	80%
MD04G1199500	4	60%	30%	10%	10%	10%	80%
MD04G1229800	4	60%	30%	10%	10%	10%	80%