Université de Lille – Science et Technologies

Ecole Doctorale – 104 -Sciences de la Matière, du Rayonnement et de l'Environnement

Thèse de Doctorat pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE LILLE – SCIENCES ET TECHNOLOGIES

Discipline

Sciences agronomiques et écologiques

Sous-discipline

Biologie de l'environnement, des populations, écologie

Mécanismes et conséquences évolutives de la dominance au locus d'auto-incompatibilité chez *Arabidopsis*

Soutenue le 26 Juin 2018 par

Nicolas BURGHGRAEVE

Membres du Jury :

KARINE ALIX	*Maitre de conférences* - AgroParisTech	Rapportrice
SYLVAIN GLEMIN	*Directeur de recherche* – ISEM, CC64. Université Montpellier II	Rapporteur
PASCAL TOUZET	*Professeur* – EEP – UMR8198 CNRS / Université de Lille	Examinateur
PIERRE BOURSOT	*Directeur de Recherche* – ISEM. Université de Montpellier	Examinateur
VINCENT CASTRIC	*Directeur de recherche* – EEP – UMR8198 CNRS / Université de Lille	Directeur de Thèse

Table des matières

Introduction générale

Les bases génétiques et l'évolution de la dominance : un débat fondateur de la génétique

La dominance génétique est une des propriétés de base dans les mécanismes d'hérédité, présente dès les travaux de Mendel sur les pois, et détermine quels traits seront exprimés au niveau phénotypique dans une descendance hétérozygote. Elle se traduit par le fait que le phénotype d'un des deux allèles est masqué à l'état hétérozygote. C'est également un aspect important de la dynamique d'adaptation car les probabilités de fixation des allèles dominants et récessifs peuvent différer de façon majeure (Charlesworth & Charlesworth, 2010). Toutefois, les bases génétiques et l'évolution de la dominance restent toujours une question débattue. Ce sujet a été l'objet d'un grand débat en génétique évolutive au début du 20^{e} siècle qui vit s'opposer les deux pères fondateurs de la génétique des populations, Sir Ronald A. Fisher et Sewall Wright (revue dans Billiard & Castric, 2011). Fisher défendait l'idée selon laquelle les relations de dominances observées puissent être la résultante de l'évolution de modificateurs de dominance, c'est-à-dire d'éléments génétiques contrôlant les interactions de dominance entre allèles d'autres gènes. Wright doutait de cette hypothèse car selon lui, la faible fréquence des génotypes hétérozygotes pour la majorité des mutations (en particulier délétères) fait que la sélection sur ces modificateurs est peu efficace. La mise en évidence expérimentale d'effet de mutations délétères partiellement récessives qui affectent la viabilité, et ce même à l'état hétérozygote, chez *Drosophila* (Charlesworth, 1979), autrement dit, le fait que *h* (l'effet hétérozygote, ou la dominance) soit négativement corrélé avec *s* (l'effet de l'allèle récessif) pour des mutations affectant la viabilité chez des individus hétérozygotes suggère que le phénotype de dominance de l'allèle sauvage est purement biochimique, avec un effet de l'allèle délétère mesurable à l'état hétérozygote, phénomène comparable à un effet dose. (Figure 1). Par ailleurs, le fait que les allèles sauvages chez *Chlamydomonas* tendent à être dominants sur les mutations obtenues au laboratoire alors que cette espèce passe la majeure partie de son cycle de vie à l'était haploïde et donc que la sélection sur les génotypes hétérozygotes semble par conséquence peu importante, sont

autant d'arguments en défaveur de l'existence de modificateurs de dominance (Orr, 1991). Peu de temps après la publication de la théorie de Fisher, Wright (Wright, 1934) et Haldane (Haldane, 1930) ont proposé une approche alternative basée sur l'activité enzymatique : la dominance ne serait qu'une conséquence de la courbe de saturation représentant la relation entre l'activité du gène et le phénotype (Figure1).

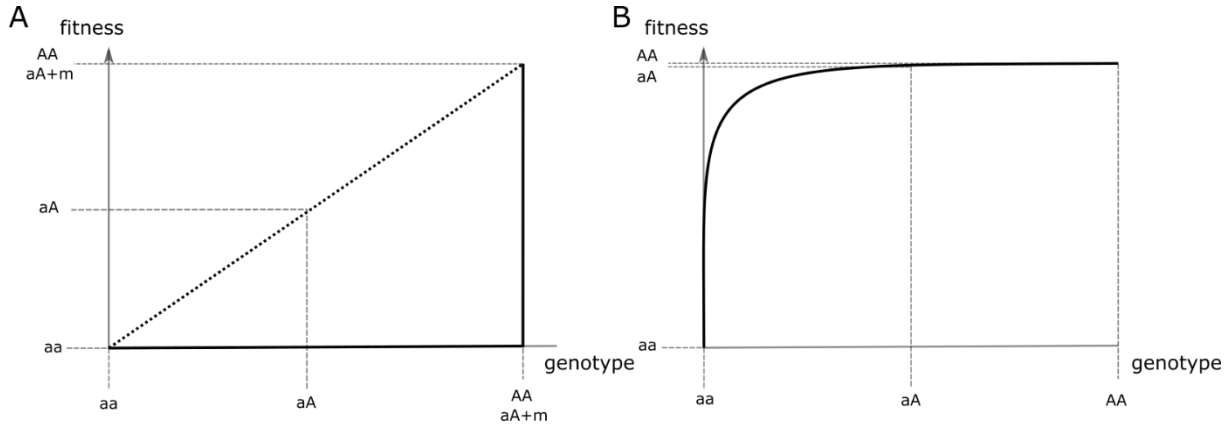


Figure 1 : Représentation des deux théories de dominance. En A, la théorie de Fisher sur les modificateurs de dominance. L'allèle muté délétère (a) le même niveau de dominance que l'allèle sauvage (A) et se traduit par un niveau intermédiaire de fitness qui est très vite contre sélectionné (droite en petits point noirs). Lorsqu'un modificateur de dominance (m) réprime l'expression de a, l'hétérozygote Aa présente la même valeur de fitness que l'homozygote AA. En B, théorie enzymatique de Wright, considérant la forme de la courbe semblable à celle d'une activité enzymatique, la fitness de l'hétérozygote Aa est très similaire à celle de l'homozygote AA.

Cette seconde théorie a été communément admise, car elle permet d'expliquer assez bien la dominance et son évolution sans devoir invoquer des modificateurs de dominance, et s'intègre bien dans un contexte où la plupart des mutations délétères sont récessives. De plus certains travaux sur les levures ont confirmé certaines prédictions de cette théorie comme la corrélation négative entre effet de la sélection (*s*) et dominance (*h*) des mutations (Phadnis & Fry, 2005; Agrawal & Whitlock, 2011). Au cours des années, cette théorie, présentée sous une forme moderne dans le cadre formalisé de la théorie enzymatique (Kacser & Burns, 1981), a obtenu le statut de paradigme (Cornish-Bowden and Nanjundiah. 2006), la possibilité même de l'existence de modificateurs de dominance ayant été largement rejetée (Billiard & Castric, 2011).

Les objections de Wright sur la théorie de dominance de Fisher étaient basées principalement sur la rareté des hétérozygotes à l'équilibre mutation-selection, ce qui ne s'applique pas, comme Wright l'a d'ailleurs signalé (Wright, 1929), à la sélection balancée. Or, les travaux d'Otto & Bourguet en 1999 ont montré que grâce au fort niveau d'hétérozygotie attendu sous sélection balancée, la sélection pour un modificateur de dominance peut être efficace dans certains cas (Otto & Bourguet, 1999). Leurs prédictions ont montré qu'il était théoriquement envisageable que des modificateurs de dominance puissent émerger en lien avec des locus soumis à des processus de sélection balancée (Peischl & Bürger, 2008), mais leur existence n'est longtemps restée qu'une possibilité théorique. En 2010, une étude a mis en évidence pour la première fois, ce que l'on peut considérer comme un modificateur de dominance (Tarutani *et al.*, 2010), sous la forme de petits ARNs non-codants, entre allèles du système d'auto-incompatibilité chez *Brassica*.

Système d'auto-incompatibilité sporophytique et gamétophytique

Le système d'auto-incompatibilité chez les plantes est un mécanisme génétique basé sur la reconnaissance et le rejet de l'auto -pollen, ce qui empêche l'autofécondation au profit de l'allofécondation afin d'éviter la dépression de consanguinité (De Nettancourt, 2001). On distingue deux types de systèmes d'auto-incompatibilité, dits gamétophytique et sporophytique. Chez le système gamétophytique (GSI), que l'on retrouve notamment chez les *Solanaceae*, les *Rosaceae*, les *Scrofulariaceae* ou encore les *Papaveraceae*, le phénotype de reconnaissance du pollen est déterminé par l'expression de son propre génotype haploïde. Ainsi, chez des individus diploïdes, chaque grain de pollen n'exprime qu'un seul allèle, tandis que les individus sont tous hétérozygotes au locus S. A l'inverse, dans le système sporophytique (SSI), le phénotype du pollen est déterminé par le génotype diploïde du parent mâle.

Au sein d'une population, ces systèmes impliquent un type de sélection, dite fréquence dépendante négative, qui favorise les allèles rares, et maintient sur le long terme une grande diversité d'allèles S (Figure 2, Castric & Vekemans, 2004). Le rejet de l'auto-pollen et la grande diversité d'allèles S en populations naturelles conduisent à un excès en hétérozygote au locus S mais aussi aux régions associées à ce locus (Kamau *et al.*, 2007).

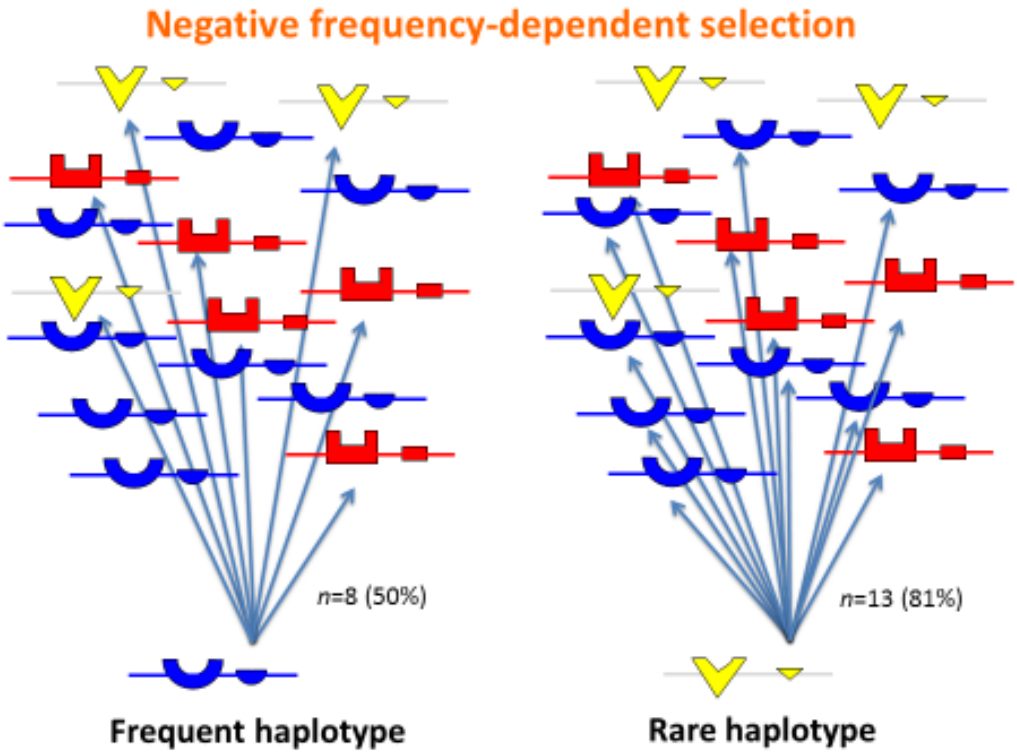


Figure 2 : Illustration de la sélection fréquence dépendante négative. Dans une population donnée de 16 individus, l'haplotype fréquent (bleu) n'est compatible qu'avec n=8 individus (50% de la population), tandis que l'haplotype rare (jaune) est compatible avec n=13 autres individus (81% de la population).

L'auto-incompatibilité sporophytique chez les Brassicaceae

Chez les *Brassicaceae*, ce mécanisme est contrôlé par deux gènes étroitement liés, *SCR* (S-locus cysteine-rich) et *SRK* (S-locus receptor kinase recognition protein), codant pour la fonction pollen et pistil respectivement, présents dans une région variable en taille et peu recombinante (Goubet *et al.*, 2012), que l'on appelle « locus S ». Dans le cas d'une pollinisation entre partenaires présentant un même haplotype au locus S (donc en particulier d'une auto-pollinisation), les deux protéines forment un complexe ligand-récepteur (Ma *et al.*, 2016) qui entraine le rejet de l'auto-pollen.

Sélection sur les modificateurs de dominance au locus S

Les systèmes d'auto-incompatibilité sporophytiques se caractérisent par la possibilité de relations de dominance/récessivité entre les allèles du locus S, et les croisements contrôlés ont permis d'en révéler l'existence chez de nombreuses espèces (Bateman, 1952; Thompson & Taylor, 1966; Kowyama *et al.*, 1994; Hatakeyama *et al.*, 1998). Des travaux théoriques ont montré que l'auto-incompatibilité correspondait bien à une situation dans laquelle la sélection sur des modificateurs de dominance pouvait être efficace. En effet, en situation de co-dominance, le pollen produit par un individu hétérozygote est reconnu et rejeté par l'ensemble des individus dont le pistil exprime l'un ou l'autre de ses deux allèles S. A l'inverse, l'établissement d'une interaction de dominance, telle qu'un modificateur lié au locus S peut la conférer, permet de masquer phénotypiquement l'un des deux allèles et ainsi d'étendre la gamme des partenaires de reproduction possibles, sans pour autant rompre l'auto-incompatibilité (Llaurens *et al.*, 2008; Schoen & Busch, 2009).

Le contrôle de la dominance chez *Brassica*

Des travaux récents chez Brassica ont contribué à révéler les bases moléculaires de ces hypothétiques modificateurs. En particulier, ils ont montré que le phénotype de dominance de la spécificité du pollen est contrôlée au niveau des ARNm et résulte d'une expression mono-allélique du gène *SCR* (Kusaba *et al.*, 2002). Cette expression mono-allélique résulte elle-même d'une répression transcriptionnelle des allèles récessifs présents à l'état hétérozygote en présence d'un allèle plus dominant (Figure 3a, Kakizaki *et al.*, 2003).

La répression de l'allèle récessif est contrôlée par les états de méthylation de la région promotrice des gènes codant pour la spécificité pollen des allèles récessifs (Figure 3b-e, Shiba *et al.*, 2002). Cette méthylation bloque la production d'ARN messager de l'allèle récessif, mais pas celle de l'allèle dominant. Il en résulte que seules les protéines des allèles dominants sont déposées sur la surface du grain de pollen (Kusaba *et al.*, 2002), causant le phénotype de dominance.

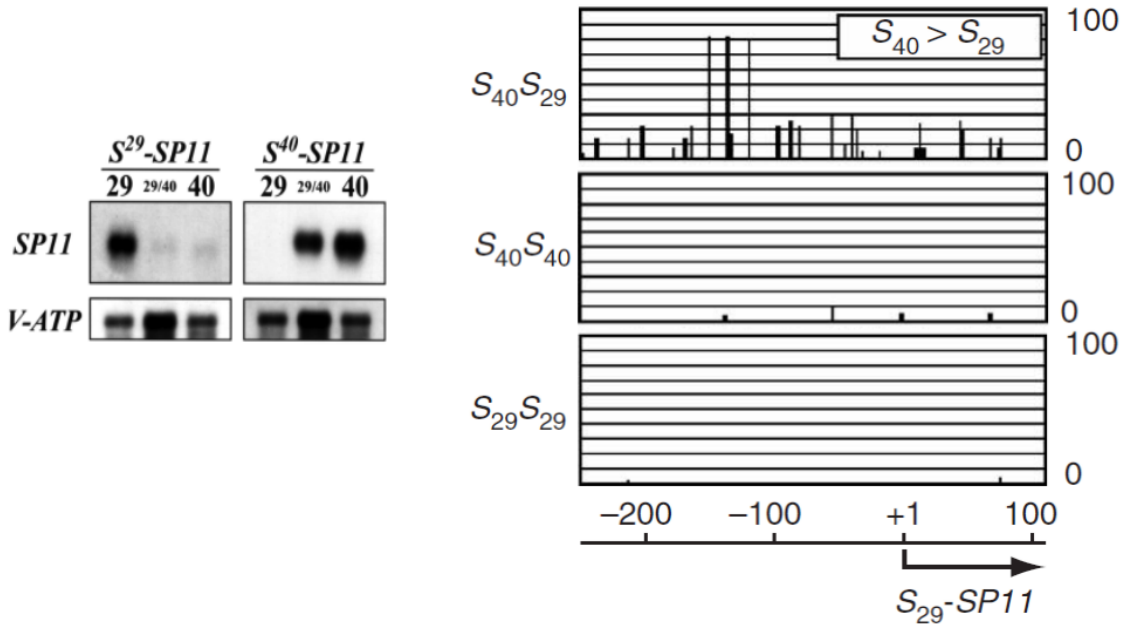


Figure 3: Expression mono-allélique des allèles pollen chez Brassica. A gauche, mesure d'expression à partir d'ARN isolé d'anthères pour chaque plantes (*Brassica*). 29: homozygote S29, 40: homozygote S40; 29/40: hétérozygote S29S40. A droite, profils de méthylation de la région promotrice pour un hétérozygote S40S29, un homozygote S29S29 et un homozygote S40S40 De –250 à 110 nucléotides ont été analysés; le pourcentage de méthylation pour chaque résidu de cytosine (49 pour S29-SP11, 43 pour S40-SP11) sont représentés par un histogramme. Adaptée de (Kakizaki *et al.*, 2003; Shiba *et al.*, 2006).

Le processus de méthylation et donc de la dominance entre allèle de classe I et de classe II (les allèles de classes I étant tous dominants sur les allèles de classe II) est sous le contrôle d'un petit ARN non codant, appelé *Smi* (Tarutani *et al.*, 2010). Par ailleurs, les relations (linéaires) de dominance entre allèles de classe II sont expliqués par un autre petit ARN non codant et polymorphe : *Smi2* (Figure 4, Yasuda *et al.*, 2016). Collectivement, l'ensemble des relations de dominance entre allèles S chez Brassica sont donc expliquées par la combinaison des effets de répression transcriptonels de deux petits ARNs non codants, Smi et Smi2 (Figure 4).

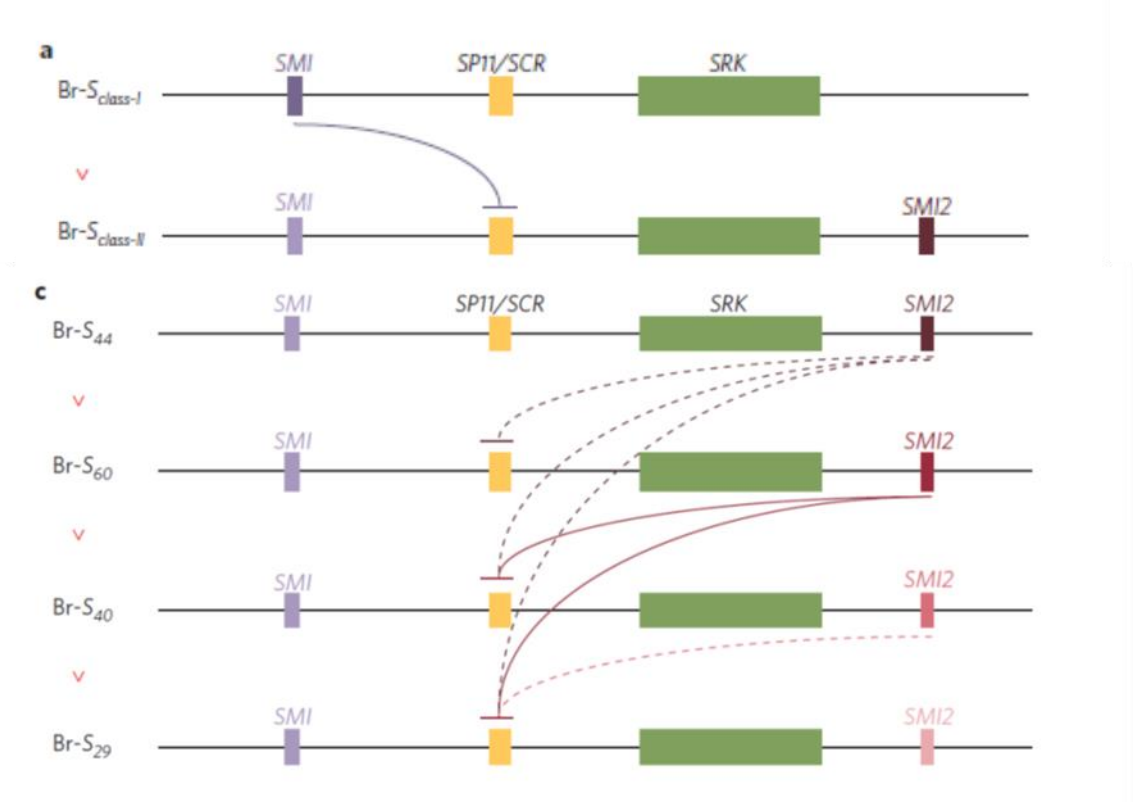


Figure 4 : Action proposée pour l'action des modificateurs de dominance chez *Brassica rapa*. **a**. Le classe-I *Smi* locus supprime l'expression du gène SP11/SCR récessif de classe-II. **b.** La hiérarchie de dominance des allèles de classe-II est sous contrôle du seul *Smi2*. La hiérarchie de dominance est donnée sur la gauche. Les lignes en pointillées représente les interactions prédites entre un miR et une cible, tandis que les lignes pleines représentent les interactions expérimentalement validées. Adaptée d'après (Goring, 2016).

Le système d'auto-incompatibilité et contrôle de la dominance chez *A. halleri*

Arabidopsis halleri est une espèce appartenant à la famille des Brassicaceae, proche de la plante modèle en génétique : *A. thaliana*. Une des différences majeure, d'un point de vue système de reproduction, est qu'*A. thaliana* est auto-compatible, tandis qu'*A. halleri* est auto-incompatible. Le système de régulation de la dominance chez *A. halleri* est plus complexe que chez *Brassica rapa*. En effet, les relations de dominance entre allèles chez Brassica forment un réseau de dominance composé d'une part d'allèles dominants (dits de classe 1, co-dominants entre eux), tous dominants sur des allèles plus récessifs (dits de classe 2, formant une hiérarchie de dominance stricte, (Thompson & Taylor, 1966; Nasrallah & Nasrallah, 1993;

Hatakeyama *et al.*, 1998). Ces deux classes d'allèles sont associées à une structure phylogénétique forte, les allèles dominants et récessifs formant deux clades phylogénétiques distincts. Cette situation est différente chez Arabidopsis, où les allèles forment une hiérarchie de dominance stricte entre de nombreuses lignées alléliques (regroupées en au moins 4 « classes » phylogénétiques dont la définition est moins claire que chez Brassica, Prigoda et al. 2005), sans cas de co-dominance (mis à part Ah20 et Ah13, tous deux très dominants (Llaurens *et al.*, 2008; Durand et al. 2014). Chez cette espèce, le réseau de régulation implique un plus grand nombre d'allèles (au moins 50), et au moins huit familles de sRNAs (Figure 5, Durand *et al.*, 2014).

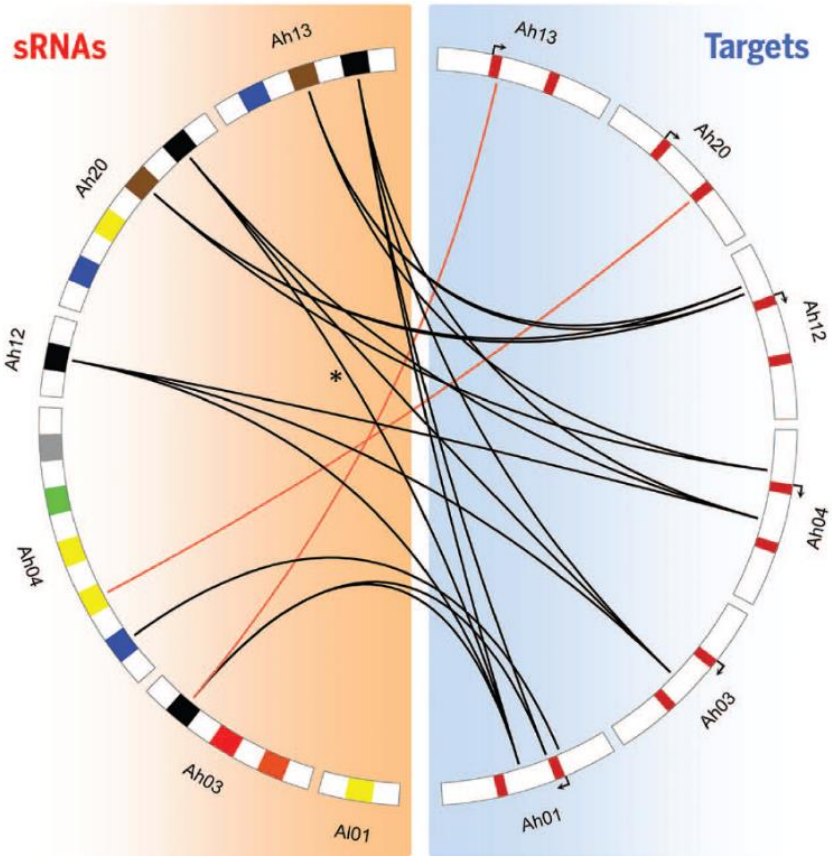


Figure 5 : Répertoire des précurseurs de sRNAs et de leurs cibles respectives. Les sRNAs portées par les allèles sont représentés à gauche, tandis que les cibles (*SCR* +-1kb, les deux exons étant représentés par les boites rouges) le sont à droite. Chaque famille possède un code couleur. Les allèles sont organisés par position dans la hiérarchie de dominance, avec les allèles les plus dominants en haut, et les plus récessifs en bas. Chaque ligne reliant un précurseur à une cible indique qu'un précurseur produit un sRNA pour la cible prédite. Les lignes en rouge représentent les prédictions qui ne sont pas en accord avec le phénotype de dominance.

Des questions mécanistique et évolutives non-résolues

Si une des interactions a été expérimentalement validée (Figure 6), il reste plusieurs incertitudes concernant le réseau de régulation. Tout d'abord, certains phénotypes documentés ne sont pas prédits par le réseau de régulation (par exemple la façon dont Ah04 est dominant sur Ah03). Ensuite, certaines prédictions moléculaires sont en conflit avec les phénotypes documentés (Ah03 n'est pas dominant sur Ah20 malgré la présence d'une interaction prédite entre un miR produit par Ah03 et une séquence cible putative chez Ah20, voir les lignes rouges, Figure 5). La prédiction des interactions miRs/cibles n'est donc pas suffisante pour expliquer tous les phénotypes de dominance observés (figure 5, les traits rouges). La nature de ces miRs est par ailleurs ambiguë, car s'ils ressemblent structurellement et sont produits à la manière des microARNs via une structure tige-boucle et dépendant de la protéine Dicer-like 1 (DCL1), ils agissent en réprimant l'expression de leur gène cible (*SCR*) par méthylation, comme des siRNAs (Voinnet, 2009; Carthew & Sontheimer, 2009). Par ailleurs, la diversité des positions des cibles (Figure 5), que l'on retrouve parfois dans le promoteur du gène *SCR*, ce qui correspond à ce qui a été décrit chez *Brassica* (Tarutani *et al.*, 2010) mais aussi dans les exons du gène, ce qui correspond plus à des cibles canoniques de microARN, en lien à une répression post-transcriptionnelle par clivage des ARNm, laissent supposer la coexistence de plusieurs voies de régulation de l'expression du gène récessif.

Ces observations posent dans un premier temps la question de la généralisation du contrôle transcriptionnel de l'expression des allèles d'auto-incompatibilité chez *A. halleri,* et dans un second temps nous interroge sur la nature des miRs impliqués dans ce réseau de régulation, ainsi que le critère qui permet de distinguer si une interaction prédite entre un miR et une cible représente une réalité biologique ou pas. Par ailleurs, cette diversité de régulateurs et de cibles pose la question des contraintes fonctionnelles qu'ils subissent, une question qui trouve un écho particulier dans le cadre de la controverse historique sur les modificateurs de dominance.

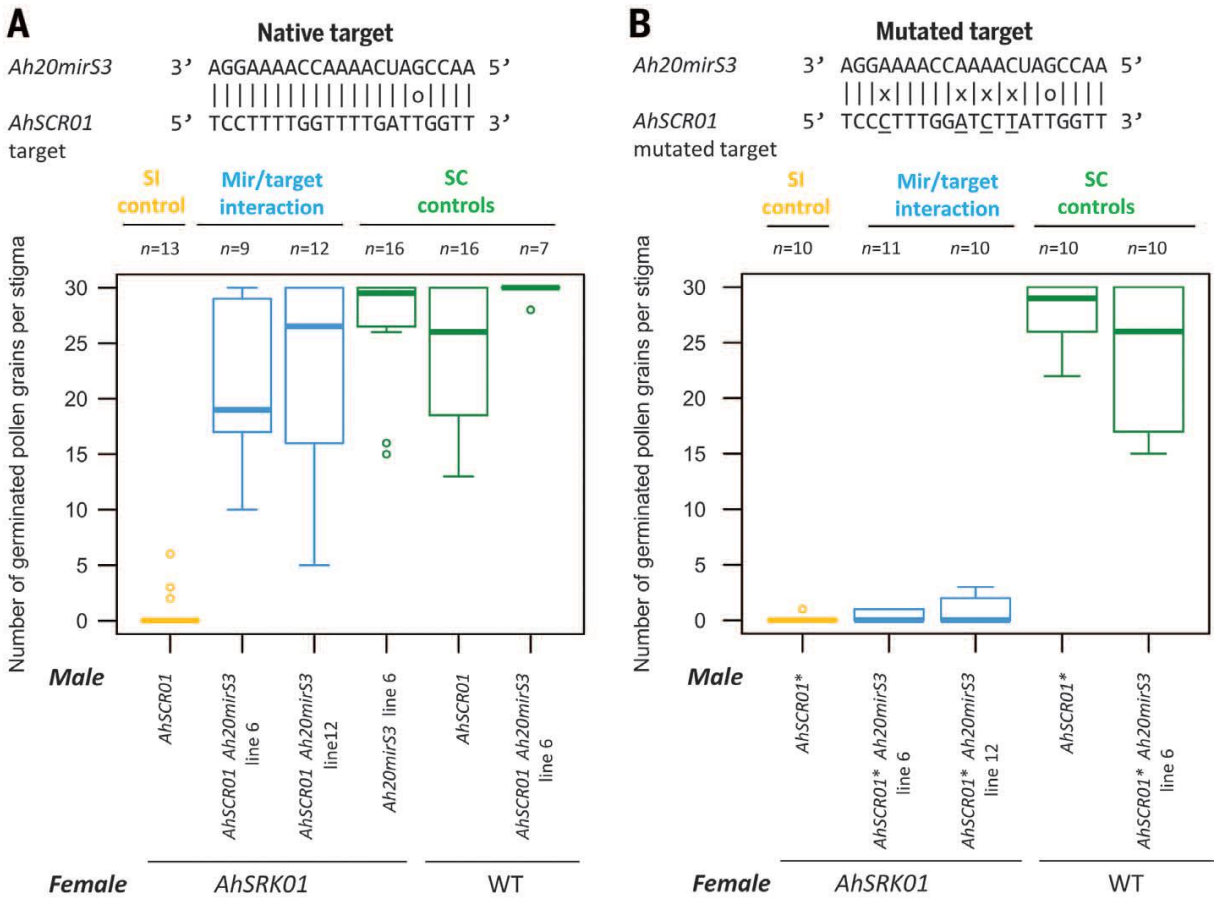


Figure 6 : Validation fonctionnelle *in planta* du contrôle de l'expression phénotypique de *SCR01* par Ah20 *via* le mirS3. **A** : l'insertion de l'Ah20mirS3 dans une plante exprimant AhSCR01 est suffisant pour y abolir la réponse auto-incompatible. **B** : la cible mutée rend AhSCR01* insensible à l'Ah20mirS3, ce qui restaure le phénotype auto-incompatible. La compatibilité du pollen a été estimée en comptant le nombre de grains germés par stigmate. Les croisements incompatibles (SI) sont en orange, les contrôles compatibles (SC) sont en vert, et les tests d'interaction miR/cible en bleu. Les barres en gras représentent les médianes, les boites la distribution interquartile, et les moustaches s'étendent jusqu'à 1.5 fois la distribution interquartile, les points individuels indiquent les outliers. Le nombre de réplicats (*n*) sont indiqués au-dessus de chaque croisement. Les mismatches des séquences sont représentés par un « x », et les bulles G:U par un « o ». Les bases mutées sont soulignées (d'après Durand *et al.*, 2014).

Conséquence de la dominance sur les lignées alléliques

Au-delà des mécanismes moléculaires par lesquels les allèles S acquièrent leur position le long de la hiérarchie de dominance, être dominant ou être récessif joue un rôle déterminant sur plusieurs aspects de leur évolution, que ce soit en termes de la forme et de la force de la sélection à laquelle les allèles sont exposés, ou de fréquence alléliques qu'il peuvent atteindre dans les populations.

Dominance et fréquence à l'équilibre

Pour rappel, dans le cadre d'un SSI, le grain de pollen est recouvert des protéines *SCR* produites dans un tissu parental diploïde, au sein duquel un seul des deux allèles est généralement exprimé et produit les protéines. L'allèle transmis dans le grain de pollen est découplé de l'allèle exprimé chez le parent, de sorte que l'allèle récessif se transmet de manière passive dans un grain de pollen possédant à sa surface la protéine de l'allèle dominant. Ainsi, outre la sélection fréquence dépendante négative, qui, comme son nom l'indique, ne tient compte que de la fréquence des allèles, s'ajoute cette transmission cachée des allèles issus des relations de dominance entre allèle si l'on veut prédire les fréquences alléliques au sein d'une population. Plusieurs modèles déterministes ont montré que les allèles récessifs devraient atteindre une fréquence plus élevée à l'équilibre que les allèles dominants. C'est ce qui est connu sous le nom de « l'effet récessif » (Bateman, 1952; Sampson, 1974). En effet, la sélection fréquence dépendante négative aura tendance à homogénéiser les fréquence des classe phénotypique (« isopléthie »), mais à cause de la dominance, les allèles récessifs peuvent être présent dans plus de classes phénotypiques que les allèles dominants, et par conséquent, atteindre une fréquence plus élevée (Figure7, Cope, 1962).

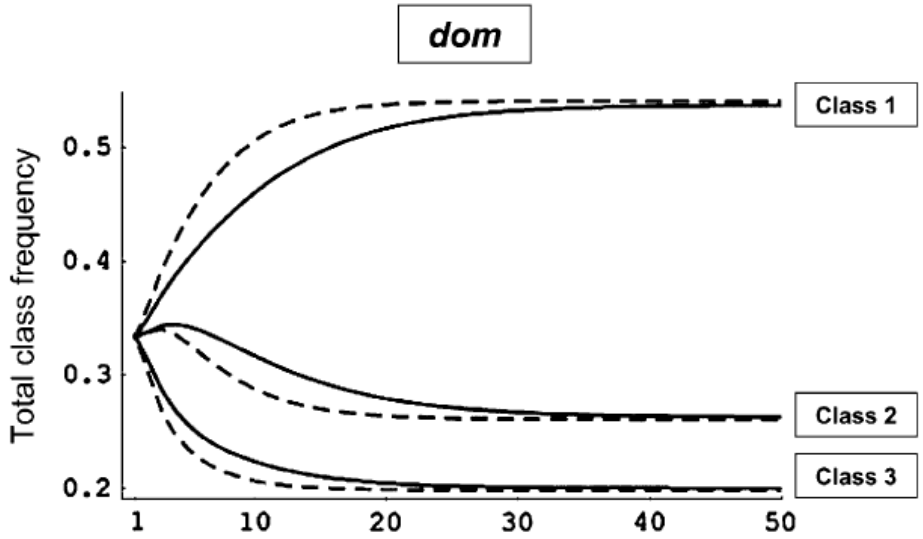


Figure 7 : Dynamique des fréquences des classes de dominance pour trois classes de chacune deux allèles (classe 1<classe2<classe3). L'axe des x représentent le nombre de générations et l'axe des y représente la fréquence respective de chaque classe. Les pointillés représentent les résultats sous un modèle de sélection fréquence-dépendante via les structures reproductives mâles et femelles. Les lignes pleines sous un modèle de sélection fréquence dépendante via les structures reproductives mâles (d'après Billiard *et al.*, 2007).

Asymétrie de l'intensité de sélection

Le système d'auto-incompatibilité sporophytique est plus compliqué à étudier d'un point de vue empirique et théorique que le système d'auto-incompatibilité gamétophytique. D'après Wright (Wright, 1939), pour un système GSI, la force évolutive principale qui agit sur le système d'auto-incompatibilité est la sélection fréquence dépendante négative. Ainsi, la sélection est symétrique entre allèles, et seule sa fréquence dans la population est à considérer. Chez le SSI, les relations de dominance provoquent une asymétrie dans la sélection entre allèles, qui est plus forte sur les allèles dominants que récessifs, qui fait qu'ils sont plus efficacement maintenus proches de leur fréquence d'équilibre (Figure 8, Bateman, 1952; Schierup *et al.*, 1997; Billiard *et al.*, 2007).

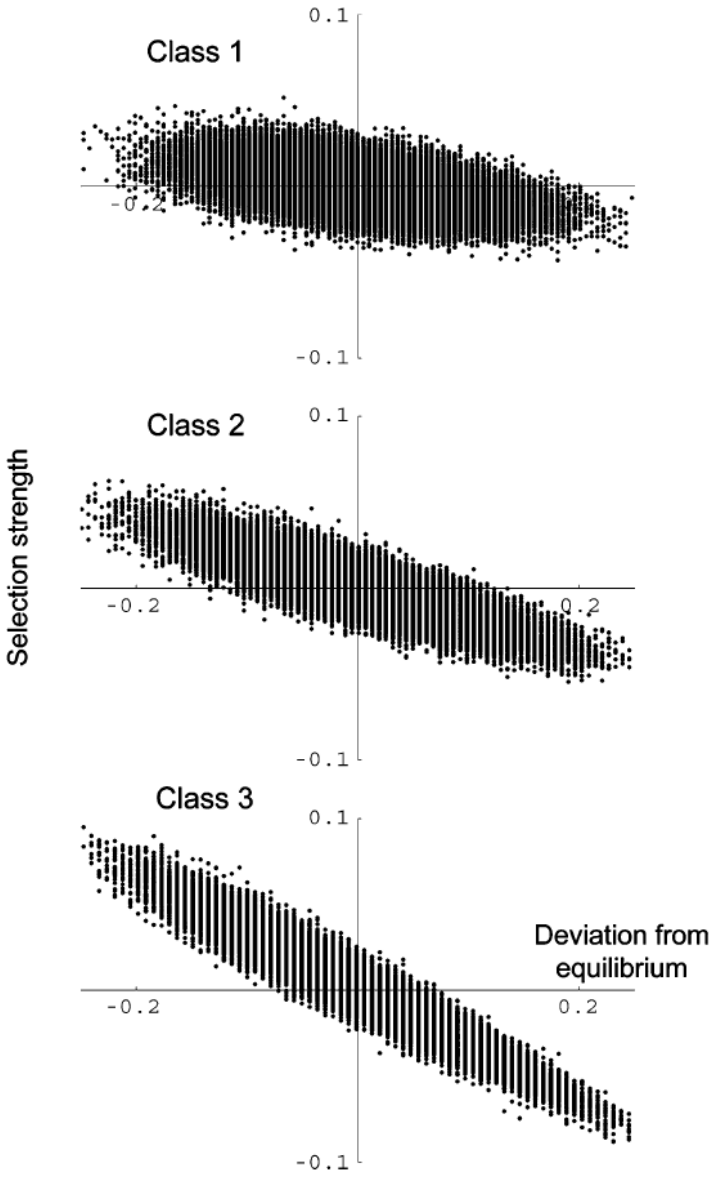


Figure 8 : Force de la sélection représentée sous la forme de déviation par rapport à l'équilibre pour trois classes alléliques distinctes. La pente représente la vitesse à laquelle un allèle revient à sa fréquence d'équilibre quand on le force à s'en écarter. On constate que la pente pour la classe 3 (allèles dominants) est plus forte (la vitesse est plus grande) que pour les allèles de classe 2 (allèles intermédiaires), elle-même plus forte que pour ceux de la classe 1 (récessifs). Issu de (Billiard *et al.*, 2007).

Impact sur les temps de coalescence

Les gènes sous sélection balancée se distinguent par une dynamique évolutive qui varie entre deux échelles de temps différentes. D'une part le maintien sur des échelles de temps très importantes de nombreuses lignées alléliques distinctes (dites « balancées »), pouvant aller jusqu'à plusieurs millions d'années (~60 Ma pour HLA, Klein *et al.*, 2007) traduit par un polymorphisme transpécifique important. Dans le cas des allèles S, cette propriété se traduit par le partage de la quasi-totalité du répertoire allélique à l'échelle des Brassicaceae (~37 Ma, Huang *et al.*, 2016). D'autre part à l'inverse, on s'attend à une très faible profondeur des généalogies au sein de chacune des lignées alléliques (Vekemans et Slatkin 1994). A chacune de ces deux échelles, la forme générale du coalescent est identique à celle d'un coalescent neutre, mais les échelles de temps considérées sont soit très largement supérieures (entre lignées) soit très largement inférieures (au sein des lignées) à celle d'un coalescent neutre. Toutes ces prédictions théoriques sont cependant issues d'un modèle d'auto-incompatibilité gamétophytique où toutes les lignées alléliques sont identiques. Or la présence d'une hiérarchie de dominance dans les systèmes sporophytiques fait que les allèles récessifs ségrégent sous un plus grand nombre de copies. On s'attend donc à ce que le temps de coalescences des allèles récessifs soit plus élevé que pour les allèles dominants (Figure 9, Castric et al. 2010).

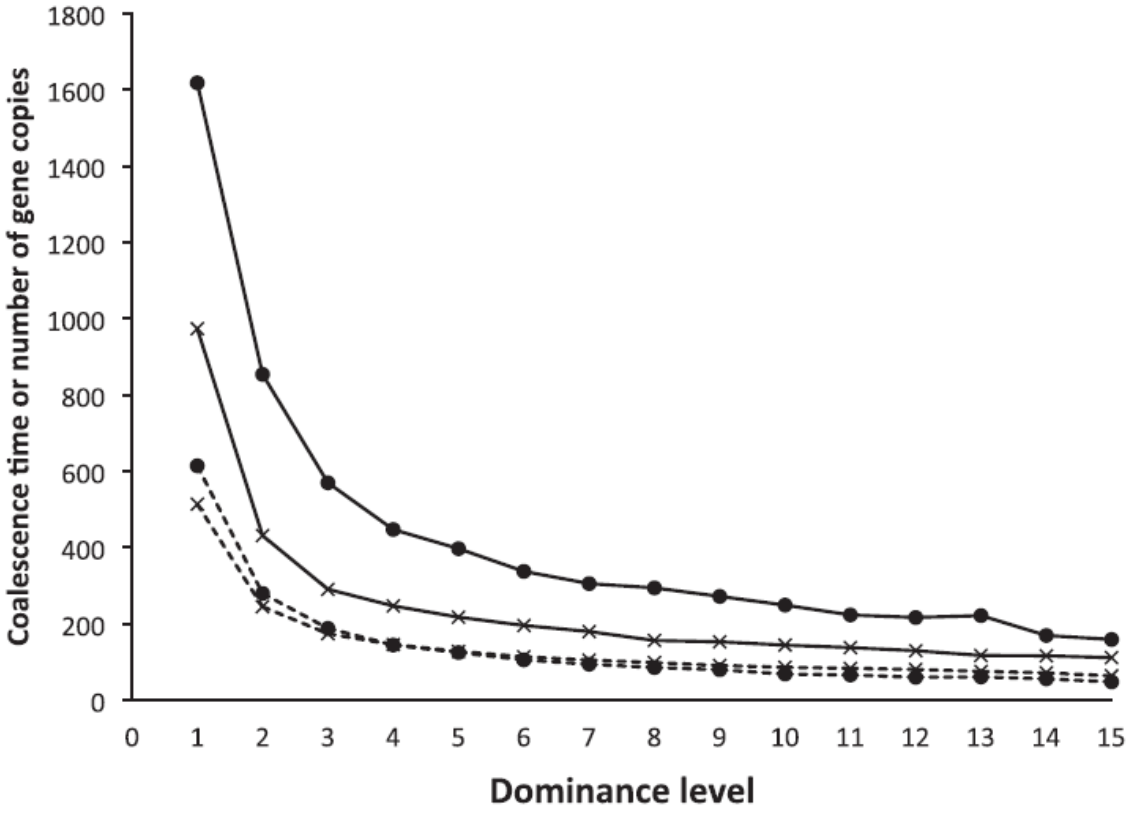


Figure 9 : Nombre attendu de copie de gène au sein des lignées alléliques (pointillés) et temps de coalescence des copies de gènes au sein des lignées alléliques (lignes pleines) en fonction du niveau de dominance dans une population panmictique (**x**) ou dans une population subdivisée en 10 dèmes, avec un taux de migration de 0.6 (•). D'après (Castric *et al.*, 2010).

Conséquence sur le niveau de polymorphisme

Le maintien des lignées alléliques sur de longues périodes de temps, la différence de fréquence à l'équilibre et de l'intensité de la sélection en lien avec la dominance sont autant de facteurs qui peuvent influencer le niveau de polymorphisme que l'on va retrouver au sein du locus S. Par exemple, les allèles récessifs présents en plus grand nombre, coalescent plus profondément au sein des lignées alléliques, et subissent une intensité de sélection plus faible que pour les allèles dominants. On s'attend à ce que le polymorphisme mesuré y soit plus élevé que pour les allèles dominants.

Conséquence de l'auto-incompatibilité sur les régions flanquantes.

Comme nous l'avons évoqué plus haut, le système d'auto-incompatibilité est sous régime de sélection fréquence-dépendante négative, provoquant un processus de sélection balancée qui peut être détectable via ses effets sur le polymorphisme des sites neutres proches.

Augmentation du polymorphisme aux sites neutres liés

Le maintien sur le long terme évoqué ci-dessus, se traduit par l'augmentation de la profondeur des généalogies. Si différents types d'allèles fonctionnels au locus persistent longtemps, chaque classe d'allèle peut acquérir son propre cortège de mutations neutres, à moins que la recombinaison ne vienne casser cette association (Charlesworth *et al.*, 2003). Les régions autour de ces allèles peuvent donc être différentes, de sorte que le polymorphisme mesuré y soit plus élevé qu'à des régions non-liées, sur une distance dépendante du taux de recombinaison (Figure 10). Plusieurs études ont essayé de déterminer l'ampleur de ce pic de polymorphisme neutre attendu aux abords du locus S, chez *A. lyrata* et *A. halleri* (Kamau & Charlesworth, 2005; Roux *et al.*, 2013) et ont montré que la distance à laquelle ce pic est mesurable est faible (de l'ordre de quelque kb). Cependant, ces études ne se basaient que sur des séquences partielles de gènes relativement éloignés les uns des autres (parfois d'une vingtaine de kb), et issues d'un échantillonnage limité. De plus, aucune des deux études n'a cherché à identifier de lien entre dominance et accumulation de polymorphisme.

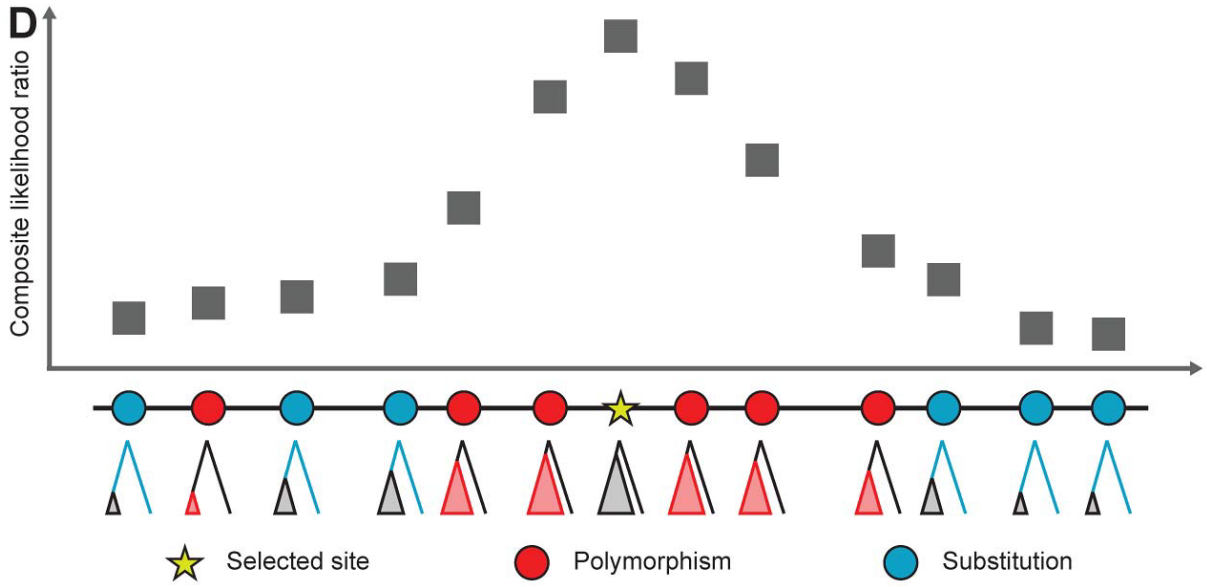


Figure 10 : Aspect des généalogies en lien avec la proximité à un site sous sélection balancée et comment la forme de ces généalogies influe sur le patron de polymorphisme autour de ce site. Le « composite likehood ratio » que l'on assimile ici au niveau de polymorphisme attendu, est plus élevé aux sites proches du site sous sélection, et diminue au fur et à mesure qu'on s'en éloigne. D'après (DeGiorgio *et al.*, 2014).

Dominance et fardeau lié

La présence d'un locus sous sélection balancée semble aussi modifier la manière dont les régions génomiques proches accumulent des mutations délétères, phénomène que l'on nomme fardeau abrité (« sheltered load », Uyenoyama, 1997; Oosterhout, 2009). Le locus d'auto-incompatibilité, et les régions liées seraient plus susceptibles de posséder un tel fardeau pour plusieurs raisons. D'abord, la sélection balancée peut interférer avec la sélection purifiante portant sur les gènes liés, empêchant par exemple l'élimination de variants délétères s'ils sont liés à des lignées alléliques balancées en phase d'augmentation, ou à l'inverse en empêchant la fixation de mutations avantageuses, le polymorphisme étant « protégé » de la fixation par le mécanisme de sélection balancée. Par ailleurs, on s'attend à y trouver un excès en hétérozygotie (Kamau *et al.*, 2007), permettant aux mutations délétères récessives de s'accumuler sans s'exprimer tant que la région n'est pas « forcée » à former des combinaisons homozygotes. Cette accumulation en mutation délétères peut avoir des conséquences importantes, et a été avancée comme une des explications derrière la longueur

des branches terminales dans la généalogie des allèles S par rapport aux attendus (Richman, 2000), et pourrait contribuer de façon substantielle à la dépression de consanguinité, surtout dans les petites populations (Glémin *et al.*, 2001). Les conséquences d'un tel fardeau lié ont été mises en évidence chez une espèce ayant un système d'auto-incompatibilité gamétophytique, *Solanum carolinense* (Stone, 2004, Figure 11).

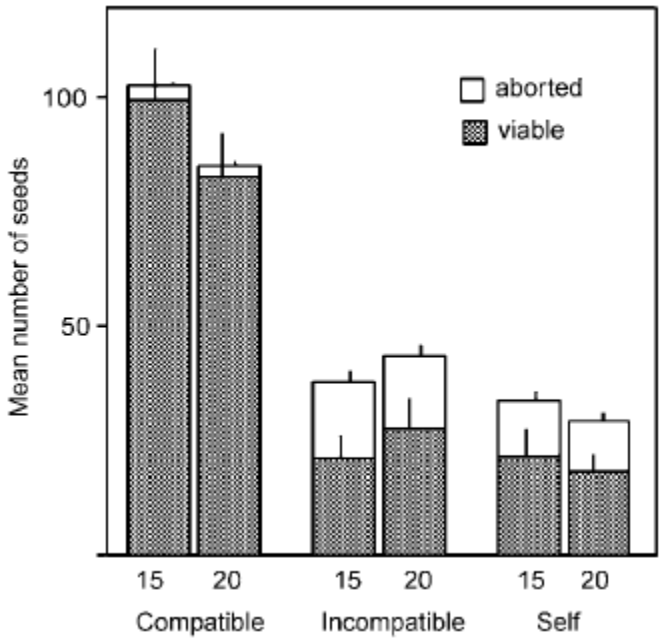


Figure 11 : Nombre total de graines, réparti en viables et avortées, après fécondation forcée de clones de *S. carolinense* de génotype S4S6). « Compatible » : croisement de contrôle entre deux plantes compatibles ; « Incompatible » : croisement forcée entre deux plantes provenant de deux lignées différents possédant les mêmes allèles S ; « Self « : croisement forcée entre deux clones de même lignée. D'après (Stone, 2004).

Chez les espèces possédant un système d'auto-incompatibilité sporophytique, la présence d'une hiérarchie de dominance entre allèles, et donc des différences en terme de pression de sélection, de fréquence et de profondeur de coalescence entre allèles dominants et récessifs peut avoir des conséquences importantes sur l'accumulation du fardeau lié. En effet, les allèles récessifs peuvent être trouvés sous forme homozygote en population naturelle, contrairement aux allèles dominants (Schierup *et al.*, 1997), et donc peuvent potentiellement purger plus rapidement des mutations fortement délétères, et donc limiter l'effet de ce fardeau lié. On s'attend donc à ce que les allèles dominants possèdent un fardeau lié plus important que les allèles récessifs (Llaurens *et al.*, 2009), et cette prédiction a été validée chez *A. halleri* par Llaurens *et al.*, (2009).

L'existence de ce fardeau lié, et sa variation le long de la hiérarchie de dominance, pose la question de son architecture génétique, c'est à dire de l'identification des mutations délétères qui en sont responsables. Etant donné la faible densité en gènes du locus S lui-même, qui ne contient que les gènes responsables des spécificités pistil (*SRK*) et pollen (*SCR*) (Goubet *et al.*, 2012) et les petits ARNs non-codants contrôlant la dominance (Durand et al. 2014), il est peu probable que les mutations responsables du fardeau s'y trouvent et doivent donc toucher les gènes des régions flanquantes. A ce jour cependant, les données de polymorphisme de ces gènes sont très fragmentaires (Kamau *et al.*, 2007; Ruggiero *et al.*, 2008; Roux *et al.*, 2013) et ne nous permettent pas de mesurer directement le niveau d'accumulation de mutations non-synonymes et ce en fonction de la distance au locus S, afin d'évaluer l'architecture génétique de ce fardeau.

Le locus S, une région difficile d'accès

Obtenir des données de polymorphisme de séquences du locus S est un travail ardu, et ce pour plusieurs raisons. En effet, la présence de nombreux allèles, très divergents les uns des autres et de taille variable (Entani *et al.*, 2003; Shiba *et al.*, 2003; Goubet *et al.*, 2012), présentant des réarrangements structuraux importants (Kusaba *et al.*, 2001; Boggs *et al.*, 2009; Guo *et al.*, 2011) avec des régions inter-géniques non alignables, une densité en gène très faible, présentant de nombreux éléments transposables (Tomita *et al.*, 2004; Goubet *et al.*, 2012; Wheeler *et al.*, 2018) et plus généralement de nombreuses séquences répétées rendent le travail de séquençage très difficile. De telles données n'ont pu être produites que

par clonage BAC (Goubet *et al.*, 2012), mais cette approche ne permet pas de multiplier le nombre de copies par allèle, et donc d'étudier le polymorphisme du locus S à l'échelle d'une population.

Objectifs de cette thèse

Le système d'auto-incompatibilité est un cas d'école pour l'étude de la sélection balancée mais reste malgré tout incomplètement décrit, principalement car ce système est présent au sein d'une région fortement perturbée par les conséquences de cette forme particulière de sélection, ce qui rend l'accès aux données génétiques difficile. Dans le cas du SSI, nous savons maintenant quelles sont les bases moléculaires de la dominance entre allèles, mais la question n'est pas complètement élucidée, car les modèles mécanistiques n'arrivent pas à capturer toute la complexité du réseau d'interaction miR/cible. En particulier, il n'est pas clair si le modèle de contrôle transcriptionnel de la dominance présent chez Brassica s'applique au réseau plus complexe de contrôle de la dominance chez Arabidopsis. Au cours du premier chapitre de ma thèse, j'ai donc réalisé une validation expérimentale de ce modèle en comparant via une approche par qPCR les niveaux d'expression des allèles du gène *SCR* selon leur statut de dominance au sein d'un grand nombre de combinaisons hétérozygotes. Cela m'a en particulier permis de mieux définir les critères d'appariement permettant aux miRs modificateurs de dominance d'accomplir leur fonction. Au-delà des questions sur les causes de la dominance, je me suis ensuite intéressé à différentes facettes des conséquences de la dominance sur le polymorphisme du locus S. Dans le deuxième chapitre, je me suis ainsi intéressé spécifiquement au polymorphisme des éléments de la machinerie de contrôle de la dominance elle-même, dans l'objectif d'évaluer l'intensité de la contrainte fonctionnelle qui pèse sur ces modificateurs de dominance (petits ARNs et leurs séquences cibles). J'ai pour cela produit des séquences complètes d'un ensemble d'allèles du locus S issus de populations naturelles grâce au développement d'une approche originale par capture de séquences. Ces données nous ont permis dans un second temps de confirmer l'absence de lien entre dominance et polymorphisme intra-allélique, en accord avec des données partielles précédemment publiées (Castric et *al.* 2010) mais en contradiction avec les attendus théoriques. Enfin, nous avons étudié dans le cadre du troisième chapitre de cette thèse les conséquences de la sélection balancée au locus S sur la diversité des régions associées, avec

comme objectif principal de décrire l'étendue de la région génomique dont le polymorphisme est affecté et de déterminer si des traces de diminution de l'efficacité de la sélection purifiante sont perceptibles au sein de cette région, qui pourraient être responsables de l'accumulation du fardeau lié. Ce chapitre combine l'analyse des données de capture de séquences à celle d'un ensemble de clones BACs, qui permettent de décrire avec une précision inégalée la structure haplotypique de cette région génomique.

Structure de la thèse

Le premier chapitre se présente sous la forme d'un article en préparation pour la revue « New Phytologist, » et est rédigé en anglais. Le deuxième et le troisième chapitre sont rédigés sous forme de projets d'articles mais sont à ce stade rédigés en français, et correspondent à des études préliminaires qui devront être complétées avant d'envisager leur valorisation.

Bibliographie

Agrawal AF, Whitlock MC. **2011**. Inferences about the distribution of dominance drawn from yeast gene knockout data. *Genetics* **187**: 553–566.

Bateman AJ. **1952**. Self-incompatibility systems in angiosperms: I. Theory. *Heredity* **6**: 285–310.

Billiard S, Castric V. **2011**. Evidence for Fisher's dominance theory: How many 'special cases'? *Trends in Genetics* **27**: 441–445.

Billiard S, Castric V, Vekemans X. **2007**. A general model to explore complex dominance patterns in plant sporophytic self-incompatibility systems. *Genetics* **175**: 1351–1369.

Boggs NA, Dwyer KG, Shah P, Mcculloch AA, Bechsgaard J, Schierup MH, Nasrallah ME, Nasrallah JB. **2009**. Expression of distinct self-incompatibility specificities in Arabidopsis thaliana. *Genetics* **182**: 1313–1321.

Carthew RW, Sontheimer EJ. **2009**. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**: 642–655.

Castric V, Bechsgaard JS, Grenier S, Noureddine R, Schierup MH, Vekemans X. **2010**. Molecular evolution within and between self-incompatibility specificities. *Molecular Biology and Evolution* **27**: 11–20.

Charlesworth B. **1979**. Evidence against Fisher's theory of dominance. *Nature* **278**: 848–849.

Charlesworth B, Charlesworth D. **2010**. *Elements of Evolutionary Genetics*. Roberts and Company.

Charlesworth B, Charlesworth D, Barton H. **2003**. The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution, and Systematics* **34**: 99–125.

Cope FW. **1962**. The effects of incompatibility and compatibility on genotype proportions in populations of Theobroma cacao L. *Heredity* **17**: 183–195.

DeGiorgio M, Lohmueller KE, Nielsen R. **2014**. A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genetics* **10**: e1004561.

Durand E, Méheust R, Soucaze M, Goubet PM, Gallina S, Poux C, Fobis-loisy I, Guillon E, Gaude T, Sarazin A, *et al.* 2014. Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. *Science* **346**: 1200–1205.

Entani T, Iwano M, Shiba H, Che FS, Isogai A, Takayama S. **2003**. Comparative analysis of the self- incompatibility (S-) locusregion of Prunus mume: identification of a pollen- expressed F-box gene with allelic diversity. *Genes to Cells* **8**: 203–213.

Glémin S, Bataillon T, Ronfort J, Mignot A, Olivieri I. **2001**. Inbreeding depression in small populations of self-incompatible plants. *Genetics* **159**: 1217–1229.

Goring DR. **2016**. Dominance modifier: Expanding mate options. *Nature Plants* **3**: 16210.

Goubet PM, Bergès H, Bellec A, Prat E, Helmstetter N, Mangenot S, Gallina S, Holl A-C, Fobis-Loisy I, Vekemans X, *et al.* 2012. Contrasted pattern of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. *PLoS genetics* **8**.

Guo Y-L, Zhao X, Lanz C, Weigel D. **2011**. Evolution of the S-Locus Region in Arabidopsis Relatives. *PLANT PHYSIOLOGY* **157**: 937–946.

Haldane JBS. **1930**. A Note on Fisher's Theory of the Origin of Dominance, and on a Correlation between Dominance and Linkage. *The American Naturalist* **64**: 87–90.

Hatakeyama K, Watanabe M, Takasaki T, Ojima K, Hinata K. **1998**. Dominance relationships between S-alleles in self-incompatible Brassica campestris L. *Heredity* **80**: 241–247.

Huang CH, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-Shehbaz I, Edger PP, *et al.* 2016. Resolution of brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution* **33**: 394–412.

Kakizaki T, Takada Y, Ito A, Suzuki G, Shiba H, Takayama S, Isogai A, Watanabe M. **2003**. Linear dominance relationship among four class-II S haplotypes in pollen is determined by the expression of SP11 in Brassica self-incompatibility. *Plant & cell physiology* **44**: 70–75.

Kamau E, Charlesworth D. **2005**. Balancing selection and low recombination affect diversity near the self-incompatibility loci of the plant Arabidopsis lyrata. *Current Biology* **15**: 1773–1778.

Kamau E, Charlesworth B, Charlesworth D. **2007**. Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of Arabidopsis lyrata. *Genetics* **176**: 2357–2369.

Klein J, Sato A, Nikolaidis N. **2007**. MHC, TSP, and the Origin of Species: From Immunogenetics to Evolutionary Genetics. *Annual Review of Genetics* **41**: 281–304.

Kowyama Y, Takahasi H, Muraoka K, Tani T, Hara K, Shiotani I. **1994**. Number, frequency & dominance relationships of S-alleles in diploid Ipomoea trifida. *Heredity* **73**: 275–283.

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME. **2001**. Self-incompatibility in the genus Arabidopsis: characterization of the S locus in the outcrossing A. lyrata and its autogamous relative A. thaliana. *The Plant cell* **13**: 627–643.

Kusaba M, Tung C-W, Nasrallah ME, Nasrallah JB. **2002**. Monoallelic expression and dominance interactions in anthers of self-incompatible *Arabidopsis lyrata*. *Plant physiology* **128**: 17–20.

Llaurens V, Billiard S, Leducq JB, Castric V, Klein EK, Vekemans X. **2008**. Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in *Arabidopsis halleri*? *Evolution* **62**: 2545–2557.

Llaurens V, Gonthier L, Billiard S. **2009**. The sheltered genetic load linked to the S locus in plants: New insights from theoretical and empirical approaches in sporophytic self-incompatibility. *Genetics* **183**: 1105–1118.

Nasrallah JB, Nasrallah ME. **1993**. Pollen Stigma Signaling in the Sporophytic Self-Incompatibility Response. *The Plant Cell* **5**: 1325–1335.

van Oosterhout C. **2009**. A new theory of MHC evolution: beyond selection on the immune genes. *Proceedings of the Royal Society B: Biological Sciences* **276**: 657–665.

Orr HA. **1991**. A test of Fisher's theory of dominance. *Proceedings of the National Academy of Sciences* **88**: 11413–11415.

Otto SP, Bourguet D. **1999**. Balanced Polymorphisms and the Evolution of Dominance. *The American Naturalist* **153**: 561–574.

Peischl S, Bürger R. **2008**. Evolution of dominance under frequency-dependent intraspecific competition. *Journal of Theoretical Biology* **251**: 210–226.

Phadnis N, Fry JD. **2005**. Widespread correlations between dominance and homozygous effects of mutations: Implications for theories of dominance. *Genetics* **171**: 385–392.

Richman A. **2000**. Evolution of balanced genetic polymorphism. *Molecular Ecology* **9**: 1953–1963.

Roux C, Pauwels M, Ruggiero M-V, Charlesworth D, Castric V, Vekemans X. **2013**. Recent and Ancient Signature of Balancing Selection around the S-Locus in Arabidopsis halleri and A. lyrata. *Molecular Biology and Evolution* **30**: 435–447.

Ruggiero MV, Jacquemin B, Castric V, Vekemans X. **2008**. Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in Arabidopsis halleri. *Genetics Research* **90**.

Sampson DR. **1974**. Equilibrium frequencies of sporophytic self-incompatibility alleles. *Canadian Journal of Genetics and Cytology* **16**: 611–618.

Schierup MH, Vekemans X, Christiansen FB. **1997**. Evolutionary dynamics of sporophytic self-incompatibility alleles in plants. *Genetics* **147**: 835–846.

Schoen DJ, Busch JW. **2009**. The evolution of dominance in sporophytic self-incompatibility systems. ii. mate availability and recombination. *Evolution* **63**: 2099–2113.

Shiba H, Kakizaki T, Iwano M, Tarutani Y, Watanabe M, Isogai A, Takayama S. **2006**. Dominance relationships between self-incompatibility alleles controlled by DNA methylation. *Nature Genetics* **38**: 297–9.

Shiba H, Kenmochi M, Sugihara M, Iwano M, Kawasaki S, Suzuki G, Watanabe M, Isogai A, Takayama S. **2003**. Genomic Organization of the S -Locus Region of Brassica. *Bioscience, Biotechnology, and Biochemistry* **67**: 622–626.

Stone JL. **2004**. Sheltered load associated with S-alleles in Solanum carolinense. *Heredity* **92**: 335–342.

Tarutani Y, Shiba H, Iwano M, Kakizaki T, Suzuki G, Watanabe M, Isogai A, Takayama S. **2010**. Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. *Nature* **466**: 983–986.

Thompson KF, Taylor JP. **1966**. Non-linear dominance relationships between S alleles. *Heredity* **21**: 345–362.

Tomita RN, Suzuki G, Yoshida K, Yano Y, Tsuchiya T. **2004**. Molecular characterization of a 313-kb genomic region containing the self- incompatibility locus of Ipomoea trifida , a diploid relative of sweet potato. *Breeding Science* **54**: 165–175.

Uyenoyama MK. **1997**. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics Sociev of America* **1400**: 1389–1400.

Voinnet O. **2009**. Origin, Biogenesis, and Activity of Plant MicroRNAs. *Cell* **136**: 669–687.

Wheeler MJ, Armstrong SA, Franklin FCH. **2018**. Genomic organization of the *Papaver rhoeas* self- incompatibility S1 locus. *Journal of Experimental Botany* **54**: 131–139.

Wright S. **1929**. Fisher's Theory of Dominance. *The American Naturalist* **63**: 274–279.

Wright S. **1934**. Physiological and Evolutionary Theories of Dominance. *American Naturalist* **68**: 25.

Wright S. **1939**. The Distribution of Self-Sterility Alleles in Populations. *Genetics* **24**: 538–552.

Yasuda S, Wada Y, Kakizaki T, Tarutani Y, Miura-uno E, Murase K, Fujii S, Hioki T, Shimoda T, Takada Y, *et al.* 2016. A complex dominance hierarchy is controlled by polymorphism of small RNAs and their targets. *Nature Plants* **16206**: 1–6.

Chapitre 1

Base-pairing requirements for small RNA-mediated gene silencing of recessive self-incompatibility alleles in *Arabidopsis halleri*

Authors

N. Burghgraeve[1], S. Simon[1], S. Barral[1], I. Fobis-Loisy[2], A-C Holl[1], C. Poniztki[1], E. Schmitt[1], X. Vekemans[1], V. Castric[1*]

Affiliations

1. CNRS, Univ. Lille, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France
2. Reproduction et Développement des Plantes, Institut Fédératif de Recherche 128, Centre National de la Recherche Scientifique, Institut National de la Recherche Agronomique, Université Claude Bernard Lyon I, Ecole Normale Supérieure de Lyon, Lyon, France

* Corresponding author

Summary

- Small RNAs (sRNA) regulate gene expression in various ways, yet identifying the molecular alphabet of sRNA-target interactions has remained a major challenge in the field. Here, we take advantage of the system of highly diversified sRNA-target interactions controlling the pollen dominance hierarchy among self-incompatibility alleles in *Arabidopsis halleri* to evaluate the base-pairing requirements for effective transcriptional silencing by sRNAs.
- We used RT-qPCR to follow expression of the pollen (*SCR*) and pistil (*SRK*) determinants of self-incompatibility in *Arabidopsis halleri* in a large number of heterozygous combinations across four developmental stages.
- *SCR* and *SRK* were expressed at their highest level in early and late stages of flower development, respectively. Recessive *SCR* alleles were transcriptionally silenced in all heterozygote combinations examined, bringing levels of *SCR* transcripts below detection limits in most cases regardless of the position of the sRNA target along the *SCR* sequence. A simple threshold model of base-pairing for the sRNA-target interaction captures most of the variation in *SCR* transcript levels. In sharp contrast, both *SRK* alleles were expressed at similar levels in all heterozygote genotypes.
- We show that the base-pairing requirements for effective transcriptional silencing by these sRNAs are broadly similar to those of canonical microRNAs, even though they are believed to function in sharply different ways. We discuss the generality of this observation and its implications on the evolutionary processes associated with the origin and maintenance of the dominance hierarchy among self-incompatibility alleles.

Key Words

Dominance/recessivity. Sporophytic self-incompatibility. RT-qPCR. heterochomatic siRNA. microRNA. *Arabidopsis halleri*.

Introduction

Small non-coding RNAs are short RNA molecules (20-25nt) with a range of regulatory functions whose central importance has constituted a major discovery in the last 20 years (Vazquez *et al.*, 2010; Aalto & Pasquinelli, 2012). The best-known members of this class of molecules are microRNAs, which are typically involved in post-transcriptional gene silencing and regulate the activity of their target gene in *trans* by either mRNA cleavage (quickly followed by degradation) or by blocking translation (Li *et al.*, 2014). In some cases, the action of microRNAs leads to the production of secondary phased short interfering RNAs (pha-siRNAs) by their target coding or non-coding sequence, which in turn can regulate downstream targets (Fei *et al.*, 2013). Another major set of small RNAs is heterochromatic short interfering RNAs (hc-siRNAs) which are mediating transcriptional silencing of repeat sequences in the genome through epigenetic modification by the RNA-dependent DNA methylation pathway (RdDM, Matzke *et al.*, 2009).

Both microRNAs and siRNAs guide their effector molecules (members of the ARGONAUTE gene family: AGO1 and AGO4, respectively) to their target sites by sequence similarity through base-pairing. For plant microRNAs, sequence similarity with the target sequence is typically very high and appears to be a shared feature of all functionally verified interactions (Wang *et al.*, 2015). Total base-pairing complementarity, however, is not the sole determinant of target specificity, and the position of the mismatches along the microRNA:target duplex is also important. Indeed, expression assays showed that while individual mismatches typically have limited functional consequences, they can also entirely inactivate the interaction when they hit specific positions such as e.g. the 10^{th} and 11^{th} nucleotide, corresponding to the site of cleavage (Jones-Rhoades *et al.*, 2006). Furthermore, the position of mismatches along the microRNA:target duplex also seems to be crucial, with a greater tolerance in the 3' than the 5' region of the microRNA (up to four mismatches generally have limited functional consequences in the 3' region, while only two mismatches in the 5' region seem sufficient to abolish the target recognition capability (Liu *et al.*, 2014, Mallory *et al.*, 2004; Parizotto *et al.*, 2004; Schwab *et al.*, 2005). These observations have led to the formulation of general "rules" for microRNA targeting (Axtell & Meyers, 2018), but in the same time they also revealed a large number of exceptions. As a result, *in silico* prediction of microRNA target sites currently

remains a difficult challenge (Ding *et al.*, 2012; Axtell & Meyers, 2018). For other types of small RNAs (pha-siRNAs and hc-siRNAs), even less is known about the base-pairing requirements for targeting, mostly because of the absence of experimentally confirmed examples of discrete, single siRNA target sites either in *cis* or in *trans* (Wang *et al.*, 2015). In this context, the recent discovery by Tarutani *et al.* (2010), Durand *et al.* (2014) and Yasuda *et al.*, (2016) of a highly diversified set of small non-coding RNAs at the gene cluster controlling self-incompatibility (SI) in Brassicaceae, provides an experimentally tractable model to evaluate the base-pairing requirements for silencing by a set of sRNAs that are regulating expression of a single gene.

Sporophytic SI is a genetic system that evolved in several hermaphroditic plant lineages to enforce outcrossing by preventing self-fertilization, hence avoiding inbreeding depression (De Nettancourt, 2001). In the Brassicaceae family, SI is controlled by a single genomic region called the "S-locus", which contains two tightly linked genes, namely *SCR* and *SRK,* that encode the pollen S-locus cysteine-rich and the stigma S-locus receptor kinase recognition proteins, respectively. Following self-pollination, these two proteins form a ligand-receptor complex (Ma *et al.*, 2016) that triggers self-pollen rejection. This system involves a polymorphism in which multiple alleles are maintained, and accordingly a large number of S-alleles is typically found in natural populations of self-incompatible species (Castric & Vekemans, 2004). With such a large allelic diversity and the very process of self-rejection, most individual plants are heterozygotes at the S-locus. Yet in most cases, only one of the two S-alleles in a heterozygous genotype is expressed at the phenotypic level in either pollen or pistil, as can be revealed by controlled pollination assays on pollen or pistil tester lines (Llaurens *et al.*, 2008; Durand *et al.*, 2014). Which of the two alleles is expressed is determined by their relative position along a dominance hierarchy, whose molecular basis for the pollen phenotype has been initially studied in the genus Brassica. In this genus, dominance is controlled at the transcriptional level in pollen, such that the phenotypically dominant S-alleles are transcribed in both homozygous and heterozygous contexts, while the recessive alleles are only expressed in homozygous context (Schopfer 1999, Kakizaki et al. 2003). Transcriptional silencing of recessive alleles by dominant alleles is caused by 24nt-long trans-acting small RNAs produced by dominant S-alleles and capable of targeting a DNA sequence in the promoter sequence of the *SCR* gene of recessive S-alleles, provoking DNA methylation (Shiba et al. 2006). Details of how these sRNAs achieve their silencing function remains incompletely understood (Finnegan *et al.*, 2011), but

it is clear that their biogenesis is similar to that of microRNAs (*i.e.,* they are produced by a short hairpin structure), while their mode of action is rather reminiscent of that of siRNAs (*i.e.,* the transcriptional gene silencing functions through recruitment of the methylation machinery). Strikingly, the full dominance hierarchy in the Brassica genus seems to be controlled by just two small RNAs called *Smi* and *Smi2* (Tarutani *et al.*, 2010, Yasuda *et al.* 2016). *Smi* and *Smi2* target distinct DNA sequences, but both are located in the promoter region of *SCR*, and both involve DNA methylation and 24-nt active RNA molecules.

The dominance hierarchy in Brassica is however peculiar in that only two ancestral allelic lineages segregate in that genus (the class I and class II alleles referred to above, see *e.g.* Leducq *et al.*, 2014), whereas self-incompatible species in Brassicaceae typically retain dozens of highly divergent ancestral allelic lineages (Castric & Vekemans, 2004). A recent study showed that in *Arabidopsis halleri*, a Brassicaceae species with multiple allelic lineages at the S-locus, the dominance-recessivity hierarchy among S-alleles in pollen is controlled by not just two but as many as eight different sRNA precursor families and their target sites, whose interactions collectively determine the position of alleles along the hierarchy (Durand *et al.*, 2014). In that genus, much less is known about the mechanisms by which the predicted sRNA-target interactions translate into the dominance phenotypes. First, the expression dynamics of the *SCR* gene across flower development stages is poorly known. Indeed, Kusaba et al. (2002) measured expression of *SCR* alleles in *A. lyrata,* but focused on only two S-alleles (*SCRa* and *SCRb*, also known as *AlSCR13* and *AlSCR20*, respectively, in Mable et al. 2003) that showed striking differences in their expression dynamics in anthers. Namely, *SCRa* was only expressed 3 days before flower opening, while *SCR*b was expressed at all stages, including within microspores. Hence, the developmental stage at which the transcriptional control of dominance in pollen should be tested is not clear. Second, while they did confirm monoallelic expression, consistent with the observed dominance relationship between the two alleles (*SCR*b > *SCR*a, Kusaba et al. 2002), the fact that only a single heterozygote combination was measured among the myriad possible combinations given the large number of S-alleles segregating in that species (at least 38 S-alleles: Castric *et al.*, 2008) prevents generalization at this step. Hence, a proper experimental validation of the transcriptional control of dominance in Arabidopsis is still lacking. Third, Durand *et al.*, (2014) observed rare sRNA-target interaction predictions that did not agree with the observed dominance phenotype. In

particular, they identified cases where no sRNA produced by a dominant allele was predicted to target the *SCR* gene of a recessive allele, while the dominance phenotype had been well established phenotypically by controlled crosses (*e.g.* Ah04>Ah03) suggesting the possibility that mechanisms other than transcriptional control may be acting. Conversely, in other rare cases, sRNAs produced by a recessive S-allele were predicted to target the *SCR* gene of a more dominant allele, suggesting exceptions to the set of base-pairing rules used to predict target sites. Fourth, although the target sites for the two sRNAs in Brassica were both located in the promoter sequence, and can thus be reasonably expected to prevent transcription initiation through recruitment of the methylation machinery, many of the predicted sRNA target sites in *A. halleri* are rather mapped to the *SCR* intron or the intron-exon boundary (beside some in the promoter as well), which suggests that distinct silencing pathways might be acting (Cuerda-Gil & Slotkin, 2016). It thus remains to be determined whether transcriptional control is also valid when the targets are at other locations along the *SCR* gene. Finally, the dominance hierarchy at the female determinant *SRK* differs from that at *SCR*, co-dominance being more frequent than on the pollen side both in Brassica (Hatakeyama *et al.*, 2001) and in *A. halleri* (Llaurens *et al.*, 2008). Limited transcriptional analysis in Brassica and Arabidopsis suggests that dominance in pistils is not associated with *SRK* expression differences, but again the number of interactions tested has remained limited (Suzuki et al. 1999; Kusaba et al. 2002).

Here, we take advantage of the fact that dominance interactions in Arabidopsis SI are controlled in pollen by a diversity of sRNAs and the diversity of their target sites to determine the base-pairing requirements for successful small-RNA mediated transcriptional silencing of recessive *SCR* alleles in plants with controlled S-locus genotypes. We first developed and validated a protocol for qPCR expression analysis of a set of *SCR* and *SRK* alleles in *A. halleri*. We then analysed the expression dynamics across four flower developmental stages of nine *SCR* and five *SRK* alleles and tested the transcriptional control of dominance for both genes in many heterozygote combinations. We quantified the strength of silencing of recessive *SCR* alleles and propose a quantitative threshold model for how sequence identity between the small non-coding RNAs and their target sites results in silencing. We discuss the implications of this model on the evolutionary processes associated with the origin and maintenance of the S-locus dominance hierarchy in Brassicaceae.

Material & Methods

Plant material

We used a collection of 88 *A. halleri* plants containing nine different S-alleles (S1, S2, S3, S4, S10, S12, S13, S20, and S29) in a total of 37 of all 45 possible homozygous and heterozygous combinations. Each plant was genotyped at the S-locus using the PCR-based protocol described in Llaurens et al. (2008). The plants were obtained by controlled crosses (Llaurens *et al.*, 2008; Durand *et al.*, 2014; Leducq *et al.*, 2014) and in a few instances were cloned by cuttings. Hence, a given S-locus genotype can be either represented in the collection by different clones (identical genetic background) or by offspring from crosses of distinct parental origins (different genetic backgrounds). Below we refer to these two levels of experimental replicates as "clone replicates" and "biological replicates", respectively. On average, the collection comprises *n*= 2.05 biological replicates per S-locus genotype and clone replicates were available for three different S-locus genotypes, Table S1). The pairwise dominance interactions between these alleles as determined by pollen and pistil compatibility phenotypes of heterozygote plants are reported in Table S3.

RNA extraction and reverse transcription

On each plant, we collected flower buds at four developmental stages: 1) five highly immature inflorescences (more than 2.5 days before opening, buds below 0.5mm, stages 1-10 in *A. thaliana* according to Smyth *et al.*, 1990), 2) ten immature buds (2.5 days before opening, between 0.5 and 1mm, approximately stage 11), 3) ten mature buds (one day before opening, longer than 1mm, approximately stage 12), and 4) ten open flowers (approximately stages 13-15). These stages were characterized by establishing the size distribution within each stage and measuring the time to flower opening based on ten buds. Samples collected were flash-frozen in liquid nitrogen, then stored at - 80°C before RNA extraction. Tissues were finely ground with a FastPrep-24 5G Benchtop Homogenizer (MP Biomedicals, Model #6004-500) equipped with Coolprep 24 x 2mL adapter (6002-528) and FastPrep Lysis Beads & Matrix tube D. Total RNAs were extracted with the Arcturus "Picopure RNA isolation" kit from Life Science (PN: KIT0204) according to the manufacturer's protocol, including a step of incubation with DNAse to remove gDNA contamination. We normalized samples by using 1 mg of total RNA to

perform reverse-transcription (RT) using the RevertAid Fermentas enzyme following the manufacturer's instructions.

Primer design

A major challenge to study expression of multiple S-alleles is the very high levels of nucleotide sequence divergence among them, precluding the possibility of designing qPCR primers that would amplify all alleles of the allelic series (both for *SRK* and *SCR*). Hence, we rather designed qPCR primers specifically targeted towards each of the *SCR* and *SRK* alleles, and for each heterozygote genotype we independently measured expression of both alleles of each gene. Primers were designed based on genomic sequences from BAC clones (Goubet et al 2012; Durand et al. 2014; Novikova et al. 2017), with a length of ~20 nucleotides, a GC content around 50% and a target amplicon size of 200nt. Whenever possible, we placed primers on either side of the *SCR* intron to identify and discard amplification from residual gDNA. However, because the coding sequence of the *SCR* gene is short, the number of possible primers was limited and this was not always possible. In two cases (*SCR01* and *SCR20*), both primers were thus located in the same exon. For *SRK* alleles, the primers were designed on either side of the first intron. To obtain relative expression levels across samples, we used *actin 8* as a housekeeping gene for standardization after we verified that the *A. thaliana* and *A. halleri* sequences are identical at the primer positions (An et al. 1996). Primer sequences and relative positions along the gene sequences are reported in Table S1.

Quantitative real-time PCR

On each cDNA sample, at least three qPCR reactions (refer to below as "technical" replicates) were performed for *actin 8* and for each of the S-alleles contained in the genotype (one S-allele for homozygotes, two S-alleles for heterozygotes). The runs were made on a LightCycler480 (Roche) with iTaq Universal SYBR Green Supermix (Bio-rad, ref 172-5121). Amplified cDNA was quantified by the number of cycles at which the fluorescence signal was greater than a defined threshold during the logarithmic phase of amplification using the LightCycler 480 software release 1.5.0 SP3. The relative transcript levels are shown after normalisation with actin amplification through the comparative $2^{-\Delta Ct}$ method (Livak &

Schmittgen, 2001). The Ct_{SCR} and Ct_{SRK} values of each technical replicate were normalized relative to the average Ct_{actin} measure across three replicates.

Validation of qPCR primers at the dilution limits

Given the very large nucleotide divergence between alleles of either *SCR* or *SRK*, cross-amplification is unlikely. However, to formally exclude that possibility, we first performed cross-amplification experiments by using each pair of *SCR* primers on a set of cDNA samples that did not contain that target *SCR* allele but instead contained each of the other *SCR* alleles present in our experiment.

To evaluate our ability to measure expression of *SCR* alleles in biological situations where they are expected be transcriptionally silenced, we then used a series of limit dilutions to explore the loss of linearity of the relationship between Ct and the dilution factor. We used six to eight replicates per dilution level to evaluate the linearity of the amplification curve. Then we looked at the shape of the melting curves to determine whether our measures at this limit dilution reflected proper PCR amplification or the formation of primer dimers. Finally, we used water in place of cDNA to evaluate the formation of primer dimers in complete absence of the target template DNA.

Expression dynamics and the effect of dominance

We used generalized linear mixed models (lme4 package in *R*; Bates *et al.*, 2014) with Ct values normalized by the actin control as the dependant variable (the variable to explain) and six independent (or explanatory) variables: biological, clone, and technical replicates -reflecting the hierarchical structure of our dataset-, developmental stages, dominance phenotype and allelic identity (Table S5). Because expression of the different *SCR* (and *SRK*) alleles was quantified by different primer pairs with inevitably different amplification efficiencies, Ct values cannot be directly compared across alleles. Most analyses were thus performed by comparing expression levels of a given focal allele in different contexts (e.g. different genotypic contexts, different developmental stages) and accordingly we considered the identity of *SCR* or *SRK* alleles as nuisance parameters in our statistical model by including them as random effects. We visually examined normality of the residuals of the model under different distributions of $2^{-\Delta Ct}$, including Gaussian, Gamma and Gaussian with logarithmic

transformations. We then tested the effect of developmental stages and dominance on *SCR* and *SRK* expression by considering them as fixed effects. Phenotypic pairwise dominance relationships were obtained from Llaurens *et al.*, (2008) and Durand *et al.*, (2014), and a set of additional controlled crosses performed following the same protocol (Table S3). Pollen and pistil dominance relationships were used to assess the effect of dominance on *SCR* and *SRK*, respectively. The existence of an interaction between the "S-allele" and "stage" effects was used to test whether the different S-alleles have distinct expression profiles across developmental stages, as suggested by Kusaba *et al.* (2002) in *A. lyrata*.

Target features and silencing effect

We then sought to determine how the expression of *SCR* alleles was affected by specific features of the small RNA-target interactions between alleles within heterozygote genotypes. We first used the small RNA sequencing data in Durand *et al.* (2014) and Novikova *et al.* (2017) to identify the populations of 18-26nt small RNA molecules produced by the small RNA precursors carried at the S-locus by each of the nine S-alleles. For each heterozygote combination, we then predicted the presence of putative target sites of the small RNAs produced by one S-allele on the genomic sequence of the *SCR* gene of the other S-allele including 2kb of nucleotide sequence upstream and downstream of *SCR* using the dedicated alignment algorithm and scoring matrix described in Durand *et al.* (2014). The reciprocal analysis was also performed regardless of the dominance relationship. Briefly, alignment quality was assessed by a scoring system based on the addition of positive or negative values for properly paired nucleotides (+1), mismatches and gaps (-1), taking into account the non-canonical G:U interaction (-0.5). For each pair of alleles considered, only the sRNA/target combination with the highest score was selected for further analysis (Table S4). We used Akaike Information Criteria (AIC) to compare how well different base-pairing scores for target site identification predicted the level of *SCR* expression (and hence the silencing phenomenon), varying the threshold from 14 to 22 (Table S5c). Lower values of AIC are associated with a best fit of the model. We then added a new fixed effect in our model to test whether targets in the promoter or in the intron of the *SCR* gene were associated with different strengths of silencing. For this analysis, we included only targets above the threshold identified (score >= 18).

To determine whether the base-pair requirement for silencing were identical between Brassica and Arabidopsis, we calculated the alignment score with our method between *Smi* & *Smi2* sRNAs and their targets sites in the class II alleles in *Brassica rapa* (Tarutani *et al.*, 2010, Yasuda *et al.*, 2016). Finally, we used the phylogeny in (Durand *et al.*, 2014) to classify sRNA/target interactions into "ancient" and "recent". Based on this classification, by compared the mean alignment score and a linear regression, for recent and ancient sRNA/target interactions in order to test the hypothesis that interactions with base-pairing scores above the threshold at which silencing was apparently already complete correspond to recently emerged interactions that did not yet have time to accumulate mismatches.

Results

Validation of the qPCR protocol and the allele-specific primers

Melting curves confirmed proper amplification and low primer dimers formation unless template DNA concentration was very low (data not shown). The specificity test confirmed the absence of cross-amplification between alleles, as the Ct measures for water control and cross amplification were comparably high (around Ct=34) and both were higher than the positive controls (median Ct=22, Figure S1). For each allele tested, we then evaluated the linearity of Ct values through serial dilutions of the template cDNA. Overall, the range of variation of Ct values spanned by a given allele across the different developmental stages or dominance status was generally well within the range over which Ct varied mostly linearly with template cDNA concentration, suggesting high power to detect these effects. For *SCR*, linearity was good throughout most of the dilution range, but was lost as expected at very low concentration (in particular for alleles *SCR01*, *SCR02*, *SCR04*, *SCR13* and *SCR20*, Figure S2a). For *SRK*, poor linearity was observed for *SRK12* (Figure S2b), so this allele was excluded from further analyses. We note that comparing levels of expression for a given allele between different recessive contexts (*e.g.* when silenced by different sRNAs) should be more challenging, especially for the above-mentioned alleles.

SCR and *SRK* expression dynamics across flower development stages

In total, we performed 344 RNA extractions and RT-PCR from the 37 different S-locus genotypes sampled at four developmental stages and measured 1,838 Ct_{SCR}/Ct_{actin} expression ratios, resulting in an average of 26.9 measures of each S-allele for each diploid genotype when combining clone, biological, and technical replicates and 480 Ct_{SRK}/Ct_{actin} (Table S1, Table S2). Distribution of the residues of the generalized mixed linear model was closest to normality after log-transformation of the ratios (Figure S4). As expected, measured expression levels were more highly repeatable across clones than across biological replicates for a given S-locus genotype, but these sources of variation were minor as compared to the technical error and the allele's expression dynamic in our experiment (deviance estimates of 0.40, 1.07, 6.08 and 4.56 respectively, Table S5a) after taking allele identity, developmental stage and dominance status into account. To determine the expression dynamics of the different *SCR* alleles, we focused on genotypes in which a given focal allele was known to be dominant at the phenotypic level (Figure 1a). Overall, we observed a consistent pattern of variation among stages (F-value: 13.805; p-value: 1.107e-05, Table S5c) with a very high expression in buds at early developmental stages (<0.5 to 1mm), and low level of expression in late buds right before opening and in open flowers, consistent with degeneration in these stages of the anther tapetum where *SCR* is expected to be expressed. Accordingly to Kusaba *et al.*, (2002), we found evidence that the expression dynamics varied across alleles (Chi^2: 217.32, p-value < 2.2e-16, Table S5c). The *SRK* alleles had sharply distinct dynamics of expression, with consistently increasing expression in the course of flower development (Chi^2: 6.9103, p-value 0.00857, Table S5g), with lowest expression in immature buds (<0.5mm) and highest expression in open flowers (Figure 1b).

Transcriptional control

Based on these results, we compared expression of *SCR* alleles across genotypes by averaging $2^{-\Delta Ct}$ values across <0.5mm to 1mm stages. Beside a few exceptions (see below), our expression data were largely consistent with the hypothesis of transcriptional control of the dominance hierarchy in pollen (31 of 37 genotypic combinations, Figure 2). In the four S-alleles

for which homozygote genotypes were available (S1, S2, S3 and S20), *SCR* alleles had substantial expression in homozygotes and this was the only case where expression of the most recessive allele (*SCR*01) could be detected. One of the two biological replicates for the S1S1 homozygote genotype had consistently low expression across two clone replicates (Figure S3), so we confirmed homozygosity of these two samples by analysing segregation after crossing to plants that did not carry S1 (all of 58 tested progenies indeed carried *SCR*01). Climbing up the dominance hierarchy from most recessive to most dominant, the S-alleles measured were expressed in an increasing number of heterozygous combinations. At the top of the dominance hierarchy, the two most dominant alleles, *SCR*13 and *SCR*20, were expressed in all heterozygous contexts, including when they formed a heterozygote combination with one another, as expected given the codominance observed between them at the phenotypic level (Durand *et al.*, 2014). This general rule had a few exceptions however (Figure 2). For instance, we detected some expression for both *SCR02* and *SCR*29 in heterozygote combination even though phenotypic data indicate that S2>S29 in pollen (Table S3). We also observed low expression for *SCR*10 and *SCR*12 when they were in heterozygote combination with *SCR*01 and the absence of expression for both *SCR*10 and *SCR*12 in the heterozygote combination they formed together, which is not consistent with the documented phenotypic dominance of these two alleles over *SCR*01 and between them (*SCR*12>*SCR*10; see Table S3). We confirmed proper phenotypic expression of S12 in pollen produced by the S10S12 genotype, as five replicate pollinations on a S1S12 plant produced no silique.

Overall, in spite of these six exceptions, we observed a striking contrast in transcript levels for a given allele according to its relative phenotypic dominance status in the genotype, with at least an overall 145-fold increase in transcript abundance in genotypes where a given focal allele was phenotypically dominant as compared to genotypes in which the same focal allele was recessive at stages when *SCR* is expressed, according to our result (F-value: 38.582; p-value: < 2.2e-16, Table S5c). In most cases, the recessive allele came close to or even below the detection limits of our method as determined by the break of linearity of the dilution experiment (Figure S1), so this fold-change value is probably under-estimated. In contrast, we found a slight effect only of dominance in pistils on *SRK* expression (F-value: 4.4107, p-value: 0.006, Figure 3, Table S5h).

Target features and silencing effect

Levels of *SCR* expression of any given focal allele varied sharply with the alignment score of the "best" target available for the repertoire of sRNAs produced by the other allele present in the genotype (Figure 4a). Specifically, we observed high and variable expression of *SCR* when the score of its best predicted target was low, but consistently low *SCR* expression when the score of the best target was high (Figure 4a, Table S5d). Strikingly, the transition between high expression and low expression was very abrupt (around an alignment score of 18), suggesting a threshold effect rather than a quantitative model for transcriptional silencing. In three cases, the presence of a target with a high score within the *SCR* gene of the dominant allele was associated with high relative *SCR* expression (in agreement with the dominant phenotype), suggesting the absence of silencing in spite of the presence of a target with high sequence similarity to the sRNA produced by the recessive allele (sRNA from Ah03 on *SCR29*, score =18.5; sRNA from Ah04 on *SCR20*, score=20; and sRNA from Ah10 on *SCR20*, score =21; Figure 5a). Examining these targets in detail did not reveal mismatches at the 10-11th nucleotide position, suggesting that mismatches at other positions have rendered these sRNA-target interactions inactive (Figure 5a). Another exception concerns the observed low score (15.5) for the best match between a sRNA from the dominant allele Ah04 and a target at *SCR* from the recessive Ah03 allele (Figure 5b). Whether Ah04 silences *SCR* from Ah03 through this unusual target or through another elusive mechanism remains to be discovered. The alignment score obtained in Brassica for *Smi* & *Smi2* show that a threshold also exist in this family, fixed at 16.5 according to our score calculation. In spite of the generally very low expression of all recessive alleles, we found some evidence that the strength of silencing experienced by a given *SCR* allele may vary across genotypic combinations for a given allele (F-value=2.2717, p-value = 0.07558, Table S5i). However, we found no evidence that the position of the target site on the measured allele (promoter; intron; intron-exon boundary *vs.* upstream/downstream) could explain this variation (F-value=1.4432, n.s, TableS5e). Finally, we found neither evidence of being recent or old on the mean alignment score (mean= 20.41 and 20.22 respectively, F-value: 0.0362; ns; TableS5j)

Discussion

Our main objective was to evaluate the base-pairing requirement of the sRNA-target interactions controlling dominance/recessivity interactions between alleles of the allelic series controlling SI in *A. halleri*. Determining the base-pairing requirement for sRNA silencing in plants has remained challenging because the "rules" used for target prediction have typically been deduced from observations that conflate distinct microRNA genes and their distinct mRNA targets over different genes. Moreover, detailed evaluations of the functional consequences of mismatches have relied on heterologous reporter systems (typically GFP in transient tobacco assays), hence limiting the scope of the phenotypic consequences that can be studied. Here, we used a genetic system (plant self-incompatibility) where multiple sRNAs regulate target sites on a single gene (*SCR*), and in which we are able to make a direct link between the sRNA-target interactions, the level of *SCR* transcript and the encoded phenotype (dominance/recessivity interaction).

The first step was to clarify several aspects of the expression pattern of the genes controlling SI in *A. halleri,* as earlier accounts had suggested that alleles of the allelic series may differ from one another in their expression profile (Kusaba *et al.*, 2002). In line with Kakizaki *et al.*, (2003), Suzuki *et al.*, (1999); Schopfer *et al.*, (1999); Takayama *et al.*, (2000); Shiba *et al.*, (2002),we found maximal expression of *SCR* in early buds but low or no expression at the open flower stage. This expression pattern is consistent with *in situ* hybridation experiments showing that *SCR* transcripts are localized in the tapetum, a specialized layer of cells involved in pollen grains coating (Iwano *et al.*, 2003), which undergoes apoptosis and is quickly degraded as the development of pollen grains inside the anther progresses (Murphy & Ross, 1998; Takayama *et al.*, 2000). We confirmed that differences exist in the temporal dynamics expression among alleles, as suggested by Kusaba *et al.* (2002) in *A. lyrata*, possibly as the result of strong sequence divergence of the promotor sequence of the different *SCR* alleles. Finally, we confirmed that *SCR* and *SRK* have sharply distinct expression dynamics throughout flower development. Indeed, transcript levels of *SRK* increased steadily along development and were very low in early buds, consistent with the observation that SI can be experimentally overcome to obtain selfed progenies by "bud-pollinisation" (Llaurens et al. 2009).

Based on this clarified transcriptional dynamics, we confirmed the generality of the transcriptional control of dominance for *SCR*. In particular, we observed that even in the few heterozygote genotypes where no sRNA produced by the phenotypically dominant allele was predicted to target the sequence of the phenotypically recessive *SCR* allele, transcripts from the recessive *SCR* allele were undetected. This suggests either that some functional sRNAs or targets have remained undetected by previous sequencing and/or our *in silico* prediction procedures, or that mechanisms other than sRNAs may cause transcriptional silencing for some S-allele combinations. In contrast, we confirmed the absence of transcriptional control for *SRK*, for which both alleles were consistently expressed at similar levels in all heterozygote genotypes examined, irrespective of the (pistil) dominance phenotype. For *SRK*, other dominance mechanisms must therefore be acting, which are yet to be discovered (*e.g.* Naithani *et al.*, 2007).

An important feature of the silencing phenomenon is that the decrease of transcript levels for recessive *SCR* alleles was very strong in heterozygous genotypes, bringing down transcript levels below the limits of detection in most cases. This is in line with the intensity of transcriptional silencing by heterochromatic siRNAs (typically very strong for transposable element sequences, see Marí-Ordóñez *et al.*, 2013), while post-transcriptional gene silencing by microRNAs is typically more quantitative (Liu *et al.*, 2014). As a result of this strong decrease of transcript levels, the strength of silencing appeared independent from the position of the sRNA target along the *SCR* gene (promoter *vs.* intron), although we note that our power to distinguish among levels of transcripts of recessive alleles, which were all extremely low, is itself fairly low.

Based on the many allelic combinations where we could compare the agnostic prediction of putative target sites with the level of transcriptional silencing, we find that a simple threshold model for base-pairing between sRNAs and their target sites captures most of the variation in *SCR* expression in heterozygotes. This result provides a direct experimental validation of the *ad-hoc* criteria used in Durand *et al.* (2014). However, our results also indicate that this quantitative threshold is not entirely sufficient to capture the complexity of targeting interactions. Indeed, in two cases for which the dominance relationship is known, this simple threshold model would inappropriately predict that sRNAs from recessive alleles should be able to target more dominant *SCR* alleles, yet the dominant *SCR* alleles were expressed at

normal levels with no sign of silencing in these heterozygote genotypes (Figure 5a). The position of the mismatches on these sRNAs (at position 15 and 18 of the sRNA for Ah03 on Ah29, and position 3 and 12 for the others) therefore appear to be sufficient to abolish the function of the targeting interaction. Similarly, a mismatch at position 10 in the *Smi* interaction in Brassica (Tarutani *et al.*, 2010) and in other microRNA-targets interactions (Franco-Zorrilla *et al.*, 2007) was shown to result in loss of function of the interaction (Table S4). Interestingly, quantitative differences may exist between Arabidopsis and Brassica, as the experimentally validated targets in Brassica (Tarutani *et al.*, 2010; Yasuda *et al.*, 2016) correspond to base-pairing threshold below that we find in Arabidopsis (*i.e.* a target score of 16.5 seems sufficient for silencing in Brassica *vs.* 18 in Arabidopsis). For Brassica, both class I and class II alleles have *Smi*, but a mismatch at the 10th position was proposed to explain why the class II *Smi* is not functional. Here, we found that this mismatch drives the alignment score under the 16.5 threshold and could be sufficient to explain the loss of function regardless of its position. Overall, although these small RNAs achieve their function in a way that is sharply different from classical microRNAs (DNA methylation *vs.* mRNA cleavage), our results suggest that the sRNA-target complementarity rules for silencing in both cases are qualitatively consistent (Liu *et al.*, 2014). Better understanding the molecular pathway by which these sRNAs epigenetically silence their target gene (*SCR*) will now be key to determine whether this threshold model can be generalized to more classical siRNAs found across the genome, as evidence is still missing for such classes of sRNAs.

The existence of a threshold model has important implications for how the dominance hierarchy can evolve. In fact, our model suggests that a single SNP can be sufficient to turn a codominance interaction into a dominance interaction (and vice-versa), making this a relatively trivial molecular event. This is actually what Yasuda *et al.*, (2016) observed in *B. rapa*, where the combination of single SNPs at the sRNA *Smi2* and its *SCR* target sequences resulted in a linear dominance hierarchy among the four class II S-alleles found in that species. Strikingly, in some cases, we observed base pairing at sRNA-target interactions with very high alignment scores (up to 22), *i.e.* above the threshold at which transcriptional silencing was already complete (score =18). Under our threshold model, such interactions are not expected since complete silencing is already achieved at the threshold, and no further fitness gain is therefore to be expected by acquiring a more perfect target. A first possibility is that these

interactions reflect the recent emergence of these silencing interactions. In fact, one of the models for the emergence of new microRNAs in plant genomes involves a partial duplication of the target gene, hence entailing perfect complementarity at the time of origin that becomes degraded over time by the accumulation of mutations (Allen *et al.*, 2004). Under this scenario, the higher-than-expected levels of sRNA-target complementarity could reflect the recent origin of these sRNAs but we found no evidence of a difference in mean alignment score for young *vs.* old microRNAs (Table S5g). A second possibility is that selection for developmental robustness is acting to prevent the phenotypic switch from mono- to bi-allelic expression of *SCR* (especially during stress events, Boukhibar & Barkoulas, 2016) that could be devastating for the plant reproductive fitness. Indeed, we do observe strong variation in overall *SCR* expression when the sRNA target score of the companion allele is below the threshold, and it is possible that under stress conditions the epigenetic machinery may be less efficient, hence requiring stronger base-pairing to achieve proper silencing than in the greenhouse conditions under which we observed them in the present study. Finally, a third possibility is that sRNA-target complementarity above the threshold reflects the pleiotropic constraint of having a given sRNA from a dominant allele control silencing of the complete set of target sequences from the multiple recessive alleles segregating, and reciprocally of having a given *SCR* target in a recessive allele maintaining molecular match with a given sRNA distributed among a variety of dominant alleles. Comparing the complementarity score of sRNA/target interactions among sRNAs or targets that contribute to high versus low numbers of dominance/recessive interactions will now require a more complete depiction of the sRNA-target regulatory network among the larger set of S-alleles segregating in natural populations.

Acknowledgments

We thank Sylvain Billiard and Isabelle de Cauwer for statistical advice and discussions, Romuald Rouger and Anne Duputié for help with producing figures and Alexis Sarazin for comments on the manuscript. This work was funded by the European Research Council (NOVEL project, grant #648321). N.B. was supported by a doctoral grant from the president of Université de Lille-Sciences et Technologies and the French ministry of research. The authors also thank the Région Hauts-de-France, and the Ministère de l'Enseignement Supérieur et de la Recherche (CPER Climibio), and the European Fund for Regional Economic Development for their financial support

Author Contribution

NB, SS, SB, ACH performed the molecular biology experiments. CP and ES obtained and took care of the plants. SS, IFL and XV provided advice on the experimental strategy and interpretations. NB performed the statistical analyses. VC supervised the work. NB and VC wrote the manuscript.

References cited

Aalto AP, Pasquinelli AE. **2012**. Small non-coding RNAs mount a silent revolution in gene expression. *Current Opinion in Cell Biology* **24**: 333–340.

Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC. **2004**. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nature Genetics* **36**: 1282–1290.

An YQ, McDowell JM, Huang S, McKinney EC, Chambliss S, Meagher RB. **1996**. Strong, constitutive expression of the Arabidopsis ACT2/ACT8 actin subclass in vegetative tissues. *The Plant journal : for cell and molecular biology* **10**: 107–121.

Axtell MJ, Meyers BC. **2018**. Revisiting criteria for plant miRNA annotation in the era of big data. *The Plant Cell*: tpc.00851.2017.

Bates D, Mächler M, Bolker B, Walker S. **2014**. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software* **67**: 1–48.

Boukhibar LM, Barkoulas M. **2016**. The developmental genetics of biological robustness. *Annals of Botany* **117**: 699–707.

Castric V, Bechsgaard J, Schierup MH, Vekemans X. **2008**. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genetics* **4**.

Castric V, Vekemans X. **2004**. Plant self-incompatibility in natural populations: A critical assessment of recent theoretical and empirical advances. *Molecular Ecology* **13**: 2873–2889.

Cuerda-Gil D, Slotkin RK. **2016**. Non-canonical RNA-directed DNA methylation. *Nature Plants* **2**: 16163.

Ding J, Zhou S, Guan J. **2012**. Finding MicroRNA Targets in Plants: Current Status and

Perspectives. *Genomics, Proteomics and Bioinformatics* **10**: 264–275.

Durand E, Méheust R, Soucaze M, Goubet PM, Gallina S, Poux C, Fobis-loisy I, Guillon E, Gaude T, Sarazin A, *et al.* 2014. Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. *Science* **346**: 1200–1205.

Fei Q, Xia R, Meyers BC. **2013**. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant Cell* **25**: 2400–2415.

Finnegan EJ, Liang D, Wang M. **2011**. Self-incompatibility : *Smi* silences through a novel sRNA pathway. *Trends in Plant Science* **16**: 238–241.

Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, García JA, Paz-Ares J. **2007**. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics* **39**: 1033–1037.

Goubet PM, Bergès H, Bellec A, Prat E, Helmstetter N, Mangenot S, Gallina S, Holl A-C, Fobis-Loisy I, Vekemans X, *et al.* 2012. Contrasted pattern of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. *PLoS genetics* **8**.

Hatakeyama K, Takasaki T, Suzuki G, Nishio T, Watanabe M, Isogai A, Hinata K. **2001**. The S Receptor Kinase gene determines dominance relationships in stigma expression of self-incompatibility in Brassica. *Plant Journal* **26**: 69–76.

Iwano M, Shiba H, Funato M, Shimosato H, Takayama S, Isogai A. **2003**. Immunohistochemical studies on translocation of pollen S-haplotype determinant in self-incompatibility of *Brassica rapa*. *Plant and Cell Physiology* **44**: 428–436.

Jones-Rhoades MW, Bartel DP, Bartel B. **2006**. MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biology* **57**: 19–53.

Kakizaki T, Takada Y, Ito A, Suzuki G, Shiba H, Takayama S, Isogai A, Watanabe M. **2003**. Linear dominance relationship among four class-II S haplotypes in pollen is determined by the expression of SP11 in Brassica self-incompatibility. *Plant & cell physiology* **44**: 70–75.

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME. **2001**. Self-incompatibility in the genus Arabidopsis: characterization of the S locus in the outcrossing A. lyrata and its autogamous relative A. thaliana. *The Plant cell* **13**: 627–643.

Kusaba M, Tung C-W, Nasrallah ME, Nasrallah JB. **2002**. Monoallelic expression and dominance interactions in anthers of self-incompatible *Arabidopsis lyrata*. *Plant physiology* **128**: 17–20.

Leducq J-B, Gosset CC, Gries R, Calin K, Schmitt É, Castric V, Vekemans X. **2014**. Self-incompatibility in Brassicaceae: identification and characterization of *SRK* -like sequences linked to the S-Locus in the tribe Biscutelleae. *Genes|Genomes|Genetics* **4**: 983–992.

Li J, Reichel M, Li Y, Millar AA. **2014**. The functional scope of plant microRNA-mediated silencing. *Trends in Plant Science* **19**: 750–756.

Liu Q, Wang F, Axtell MJ. **2014**. Analysis of complementarity requirements for plant MicroRNA targeting using a Nicotiana benthamiana quantitative transient assay. *The Plant Cell* **26**: 741–753.

Livak KJ, Schmittgen TD. **2001**. Analysis of relative gene expression data using real-time quantitative PCR and. *Methods* **25**: 402–408.

Llaurens V, Billiard S, Leducq JB, Castric V, Klein EK, Vekemans X. **2008**. Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in *Arabidopsis halleri*? *Evolution* **62**: 2545–2557.

Ma R, Han Z, Hu Z, Lin G, Gong X, Zhang H, Nasrallah JB, Chai J. **2016**. Structural basis for specific self-incompatibility response in Brassica. *Nature Publishing Group* **26**: 1320–1329.

Mable BK, Schierup MH, Charlesworth D. **2003**. Estimating the number, frequency, and dominance of S-alleles in a natural population of *Arabidopsis lyrata* (Brassicaceae) with sporophytic control of self-incompatibility. *Heredity* **90**: 422–31.

Mallory AC, Reinhart BJ, Jones-Rhoades MW, Tang G, Zamore PD, Barton MK, Bartel DP. **2004**. MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. *The EMBO Journal* **23**: 3356–3364.

Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O. **2013**. Reconstructing *de novo* silencing of an active plant retrotransposon. *Nature genetics* **45**: 1029–1039.

Matzke M, Kanno T, Daxinger L, Huettel B, Matzke AJ. **2009**. RNA-mediated chromatin-based silencing in plants. *Current Opinion in Cell Biology* **21**: 367–376.

Murphy DJ, Ross JH. **1998**. Biosynthesis, targeting and processing of oleosin-like proteins, which are major pollen coat components in *Brassica napus*. *Plant J* **13**: 1–16.

Naithani S, Chookajorn T, Ripoll DR, Nasrallah JB. **2007**. Structural modules for receptor dimerization in the S-locus receptor kinase extracellular domain. *Proceedings of the National Academy of Sciences* **104**: 12211–6.

De Nettancourt D. **2001**. *Incompatibility and Incongruity in Wild and Cultivated Plants.* (BY Springer-Verlag., Ed.).

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM, *et al.* 2016. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics* **48**: 1077–1082.

Palazzo AF, Lee ES. **2015**. Non-coding RNA: What is functional and what is junk? *Frontiers in Genetics* **5**: 1–11.

Parizotto EA, Parizotto EA, Dunoyer P, Dunoyer P, Rahm N, Rahm N, Himber C, Himber C, Voinnet O, Voinnet O. **2004**. In vivo investigation of the transcription, processing, endonucleolytic activity, and functional relevance of the spatial distribution of a plant miRNA. *Genes & Development* **18(18)**: 2237–2242.

Remans T, Smeets K, Opdenakker K, Mathijsen D, Vangronsveld J, Cuypers A. **2008**. Normalisation of real-time RT-PCR gene expression measurements in Arabidopsis thaliana exposed to increased metal concentrations. *Planta* **227**: 1343–1349.

Schopfer CR, Nasrallah ME, Nasrallah JB. **1999**. The male determinant of self-incompatibility in Brassica. *Science* **286**: 1697 LP-1700.

Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, Weigel D. **2005**. Specific effects of microRNAs on the plant transcriptome. *Developmental Cell* **8**: 517–527.

Shiba H, Iwano M, Entani T, Ishimoto K, Shimosato H, Che F-S, Satta Y, Ito A, Takada Y,

Watanabe M, *et al.* 2002. The dominance of alleles controlling self-incompatibility in Brassica pollen is regulated at the RNA level. *The Plant cell* **14**: 491–504.

Shiba H, Kakizaki T, Iwano M, Tarutani Y, Watanabe M, Isogai A, Takayama S. **2006**. Dominance relationships between self-incompatibility alleles controlled by DNA methylation. *Nature Genetics* **38**: 297–9.

Smyth DR, Bowman JL, Meyerowitz EM. **1990**. Early flower development in Arabidopsis. *The Plant Cell* **2**: 755–767.

Suzuki G, Kai N, Hirose T, Fukui K, Nishio T, Takayama S, Isogai A, Watanabe M, Hinata K. **1999**. Genomic organization of the S locus : identification and characterization of genes in SLG / SRK region of S9 haplotype of Brassica campestris (syn . rapa). *Genetics* **153(1)**: 391–400.

Takayama S, Shiba H, Iwano M, Shimosato H, Che FS, Kai N, Watanabe M, Suzuki G, Hinata K, Isogai A. **2000**. The pollen determinant of self-incompatibility in *Brassica campestris*. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 1920–1925.

Tarutani Y, Shiba H, Iwano M, Kakizaki T, Suzuki G, Watanabe M, Isogai A, Takayama S. **2010**. Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. *Nature* **466**: 983–986.

Vaucheret H, Béclin C, Elmayan T, Feuerbach F, Godon C, Morel JB, Mourrain P, Palauqui JC, Vernhettes S. **1998**. Transgene-induced gene silencing in plants. *The Plant journal* **16**: 651–659.

Vazquez F, Legrand S, Windels D. **2010**. The biosynthetic pathways and biological scopes of plant small RNAs. *Trends in Plant Science* **15**: 337–345.

Wang F, Polydore S, Axtell MJ. **2015**. More than meets the eye? Factors that affect target selection by plant miRNAs and heterochromatic siRNAs. *Current Opinion in Plant Biology* **27**: 118–124.

Wickham H. **2009**. *ggplot2: Elegant Graphics for Data Analysis.* New-York: Springer-Verlag.

Yasuda S, Wada Y, Kakizaki T, Tarutani Y, Miura-uno E, Murase K, Fujii S, Hioki T, Shimoda T, Takada Y, *et al.* 2016. A complex dominance hierarchy is controlled by polymorphism of small

RNAs and their targets. *Nature Plants* **16206**: 1–6.

Figure legends

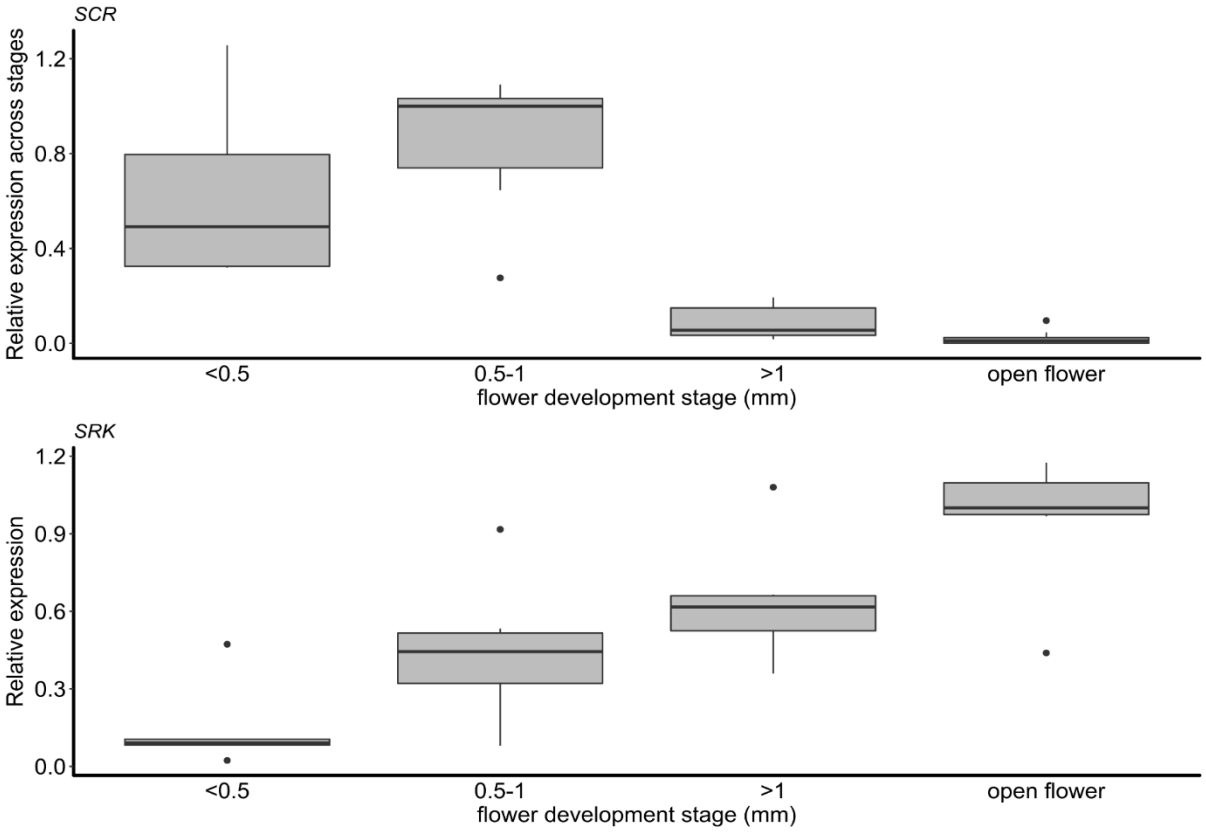


Figure 1: Expression dynamics of **a.** *SCR* and **b.** *SRK* during flower development, from early buds (<0.5mm) to open flowers. For *SCR*, only genotypes in which a given allele was either dominant or co-dominant were included (recessive *SCR* alleles were strongly silenced at all stages and were therefore not informative here). For each allele, $2^{-\Delta Ct}$ values were normalized relative to the developmental stage with the highest expression. For each stage, the thick horizontal line represents the median, the box represents the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * Inter Quartile Range from the hinge (or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge and the black dots represents outlier values.

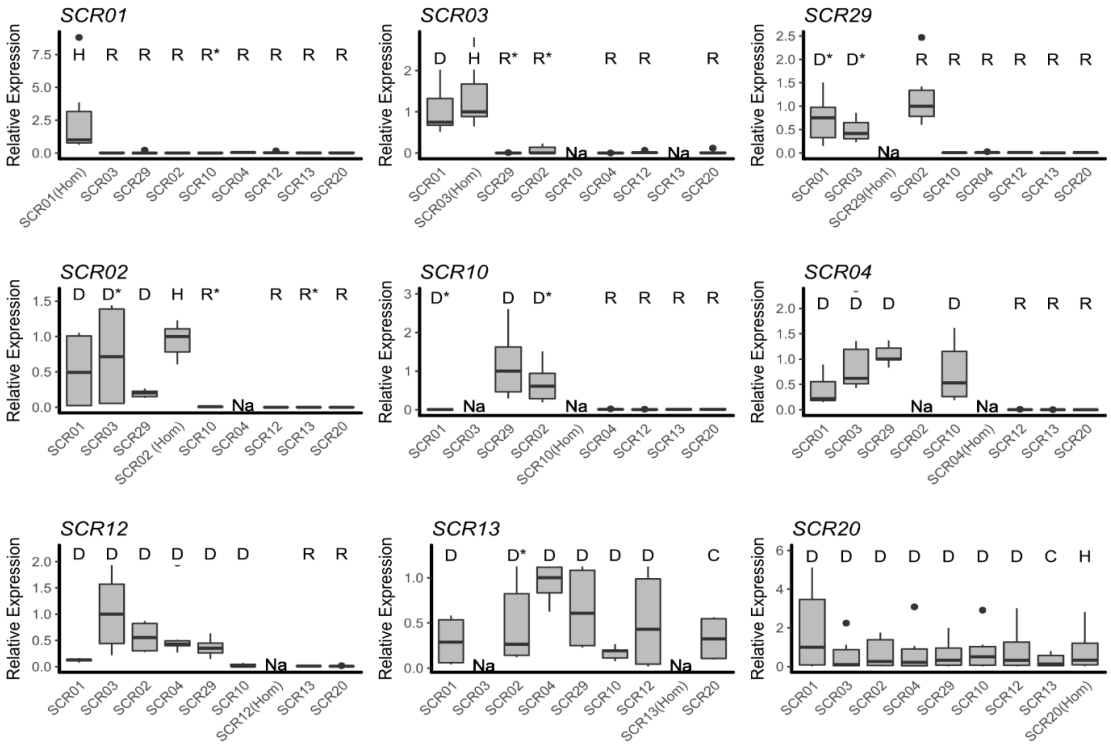


Figure 2: Expression of individual *SCR* alleles in different genotypic contexts. Pollen dominance status of the S-allele whose expression is measured relative to the other allele in the genotype as determined by controlled crosses are represented by different letters (**D**: dominant; **C**: codominant; **R**: recessive; **U**: unknown; **H**: Homozygote, Table S3). In a few instances, relative dominance status of the two alleles had not been resolved phenotypically and were inferred from the phylogeny (marked by asterisks). Thick horizontal bars represent the median of $2^{-\Delta Ct}$ values, 1st and 3rd quartile are indicated by the upper and lower limits of the. The upper whisker extends from the hinge to the largest value no further than 1.5 * Inter Quartile Range from the hinge (or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge and the black dots represents outlier values. We normalized values relative to the highest median across heterozygous combinations within each panel. Alleles are ordered from left to right and from top to bottom according to their position along the dominance hierarchy, with SCR01 the most recessive and SCR13 and SCR20 the most dominant alleles. Under a model of transcriptional control of dominance, high expression is expected when a given allele is either dominant or co-dominant and low expression when it is recessive. Exceptions to this model are marked by black vertical arrows and discussed in the text. “Na” marks homozygote or heterozygote genotypes that were not available.

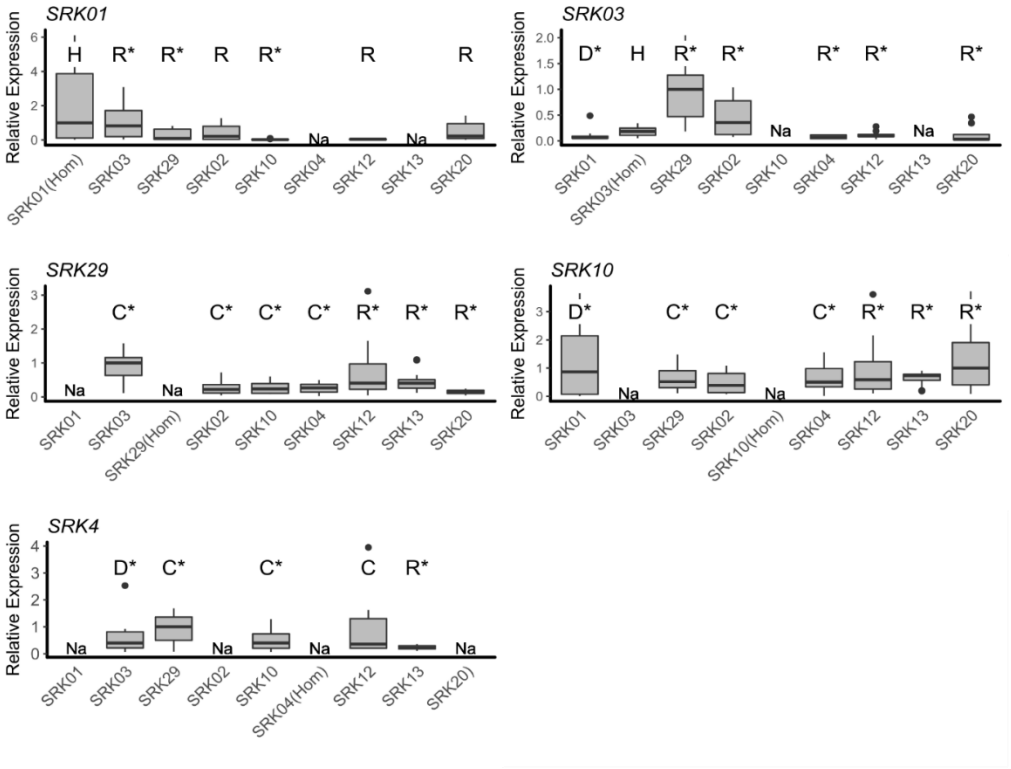


Figure 3: Expression of individual *SRK* alleles in different genotypic contexts. Putative pistil dominance status of the S-allele whose expression is measured relative to the other allele in the genotype is represented by different letters (**D**: dominant; **R**: recessive; **U**: unknown; **H**: Homozygote). Note that the pistil dominance hierarchy of the S-allele have been less precisely determined than the pollen hierarchy, and so many of the pairwise dominance interactions were indirectly inferred from the phylogenetic relationships (and marked by an asterisk) rather than directly measured phenotypically. Thick horizontal bars represent the median of $2^{-\Delta Ct}$ values, 1st and 3rd quartile are indicated by the upper and lower limits of the boxes. The upper whisker extends from the hinge to the largest value no further than 1.5 * Inter Quartile Range from the hinge (or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge and the black dots represents outlier values. We normalized the values for each allele relative to the higher median across heterozygous combination. We normalized values relative to the highest median across heterozygous combinations within each panel. Alleles are ordered from left to right and from top to bottom according to their position in the pistil dominance hierarchy, with SRK01 the most recessive and SRK04 the most dominant allele in our sample, based on the phenotypic determination in Llaurens *et al.* (2008).

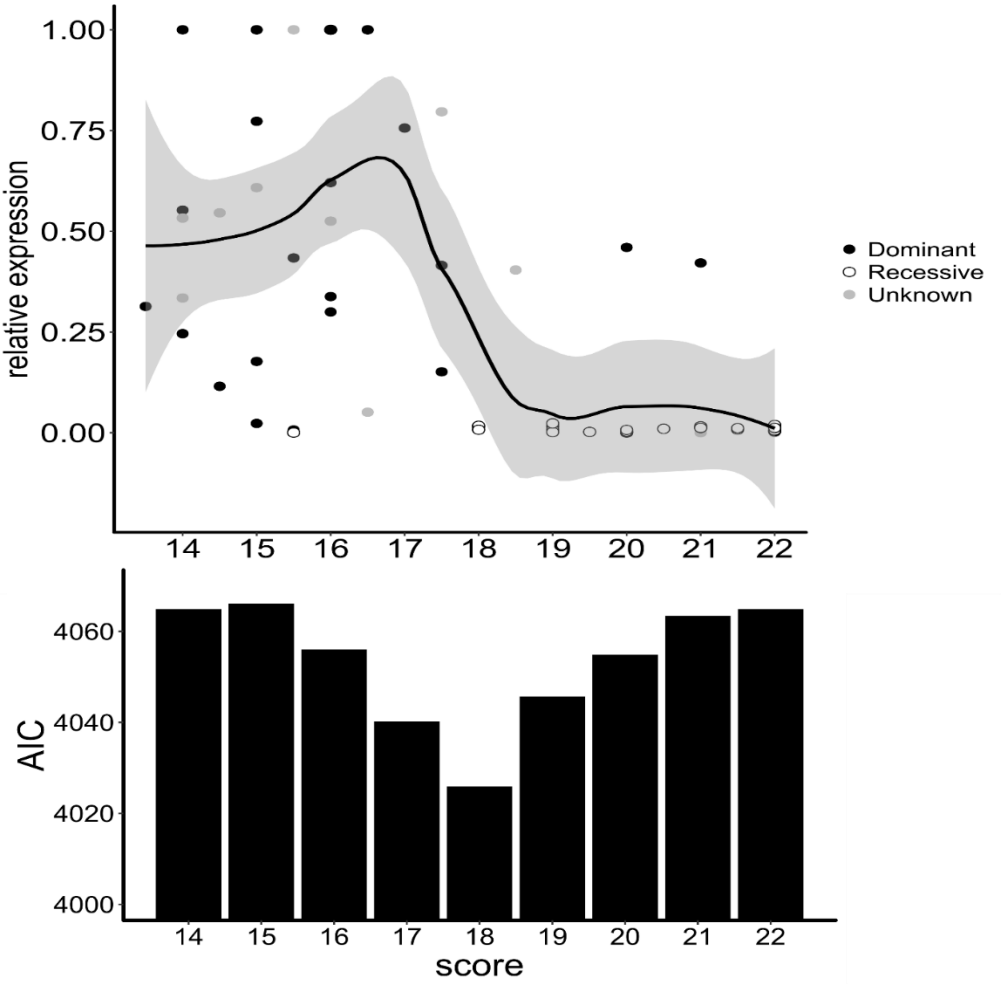


Figure 4: Base-pairing requirements for the transcriptional control of *SCR* alleles by sRNAs suggest a threshold model. **a.** Relative expression of *SCR* alleles as a function of the alignment score of the "best" interaction between the focal allele (including 2kb of sequence upstream and downstream of *SCR*) and the population of sRNAs produced by sRNA precursors of the other allele in the genotype. For each allele, expression was normalized relative to the genotype in which the $2^{-\Delta Ct}$ value was highest. Dots are coloured according to the dominance status of the focal *SCR* allele in each genotypic context (black: dominant; white: recessive; grey: undetermined). The black line corresponds to a local regression obtained by a smooth function (loess function, span=0.5) in the ggplot2 package (Wickham, 2009) and the grey area covers the 95% confidence interval. Vertical arrows point to observations that do not fit the threshold model of transcriptional control and are represented individually on Figure 5. **b.** Barplots of the Akaike Information Criteria (AIC) quantifying the fit of the generalized linear model for different target alignment scores used to define functional targets. Lower AIC values indicate a better fit.

a

```
Ah03 MirS3    3' UACAAGUUCCAUUAUAUAAUCGAA 5′
                 ||||||||||||||x||x|||o||
Ah29 target   5′ ATGTTCAAGGTAATTTACTAGTTT 3′
Score = 18.5


Ah04 Mir4239  3' CCUCGUACACCUUUAUUGCCUUUG 5′
                 |||x|||||||x||||||||||||
Ah20 target   5' GGAACATGTGGCAATAACGGAAAC 3′
Score = 20


Ah10 Mir4239  3' CCUCGUACACCUUUAUUGCCUUUGU 5′
                 |||x|||||||x|||||||||||||
Ah20 target   5' GGAACATGTGGCAATAACGGAAACA 3′
Score = 21
```

b

```
Ah04 MirS4    3' GUAUGAUUCUUGUUAGAUUCA 5‘
                 ~~~||||||||o||||||||~
Ah03 target   5' ACTACTAAGAATAATCTAAGA 3‘
Score = 15,5
```

Figure 5: Predicted sRNA/target interactions that do not fit with the documented dominance phenotype or the measured expression. For each alignment, the sequence on top is the sRNA and the bottom sequence is the best predicted target site on the *SCR* gene sequence (including 2kb of sequence upstream and downstream of *SCR*). **a.** sRNA targets with a score above 18, while the S-allele producing the sRNA is phenotypically recessive over the S-allele containing the *SCR* sequence. **b.** sRNA target with a score below 18, while the S-allele producing the sRNA (Ah04) is phenotypically dominant over the S-allele containing the *SCR* sequence and transcript levels of the *SCR03* allele is accordingly very low. This is the best target we could identify on SCR03 for sRNAs produced by Ah04.

Supplementary figures

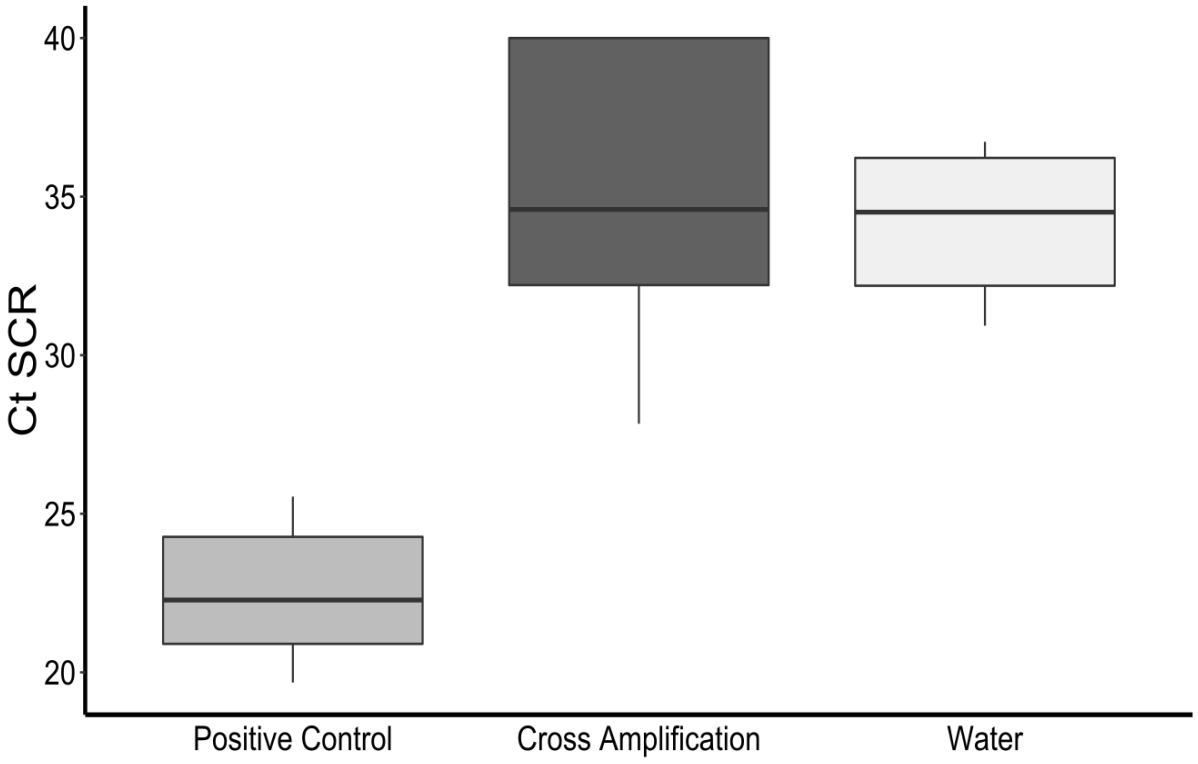


Figure S1. Validation of the *SCR* qPCR primers. "Positive control**"** corresponds to amplification with the Master Mix containing primers for *SCR* alleles that are present in the cDNA used. For the "Cross Amplification" assay, we used a Master Mix on cDNAs that do not contain alleles corresponding to the primer pair used. "Water**"**: master mix with water instead of cDNA. Thick horizontal bars represent the median of $2^{-\Delta Ct}$ values, 1st and 3rd quartile are indicated by the upper and lower limits of the. The upper whisker extends from the hinge to the largest value no further than 1.5 * Inter Quartile Range from the hinge (or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge and the black dots represents outlier values.

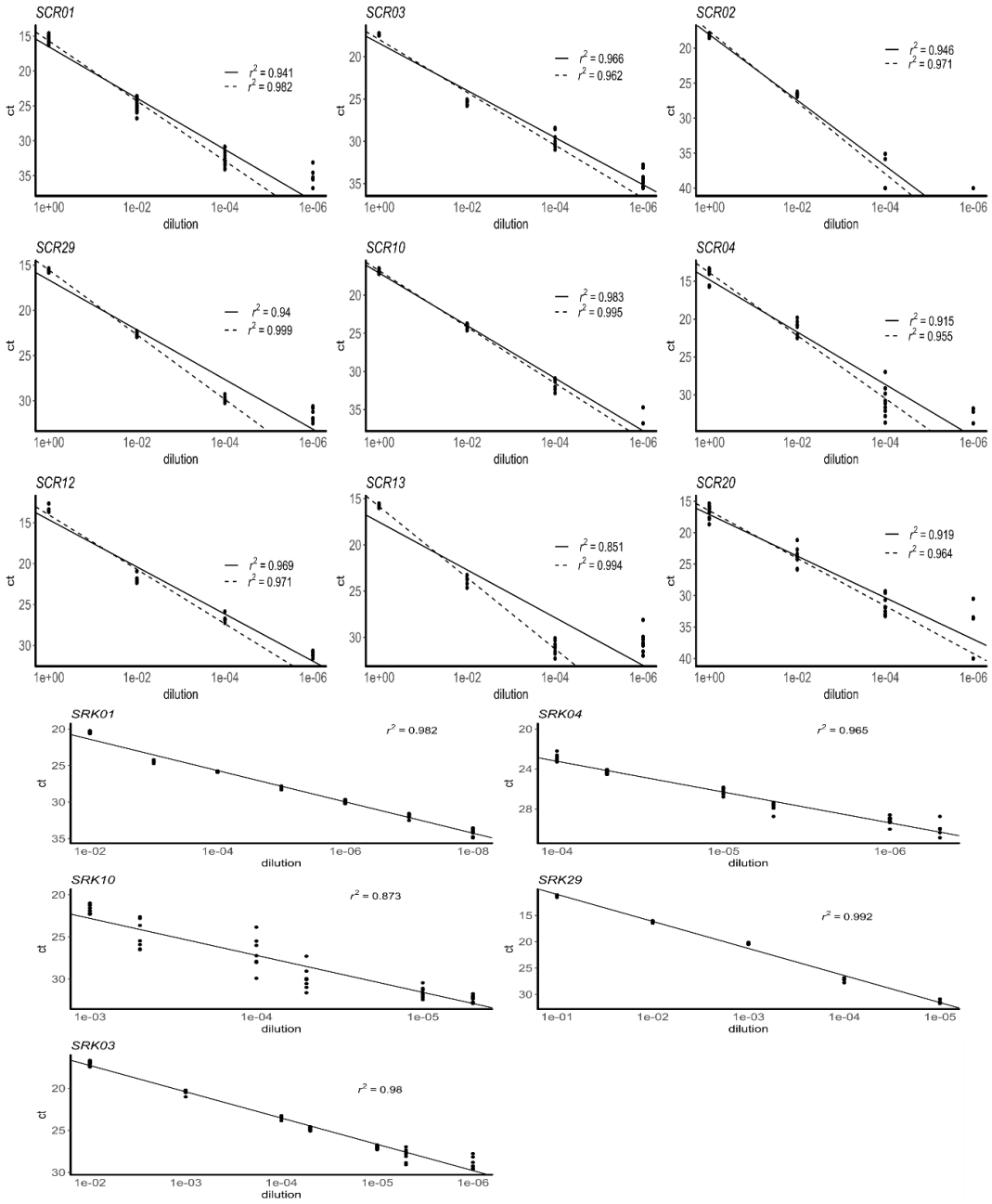


Figure S2: qPCR amplification (non-transformed Ct values) in serial dilutions for each *SCR* (**a**) and *SRK* (**b**) allele. Solid lines are the linear regressions over all Ct values. Dashed lines are linear regressions excluding the highest dilution level.

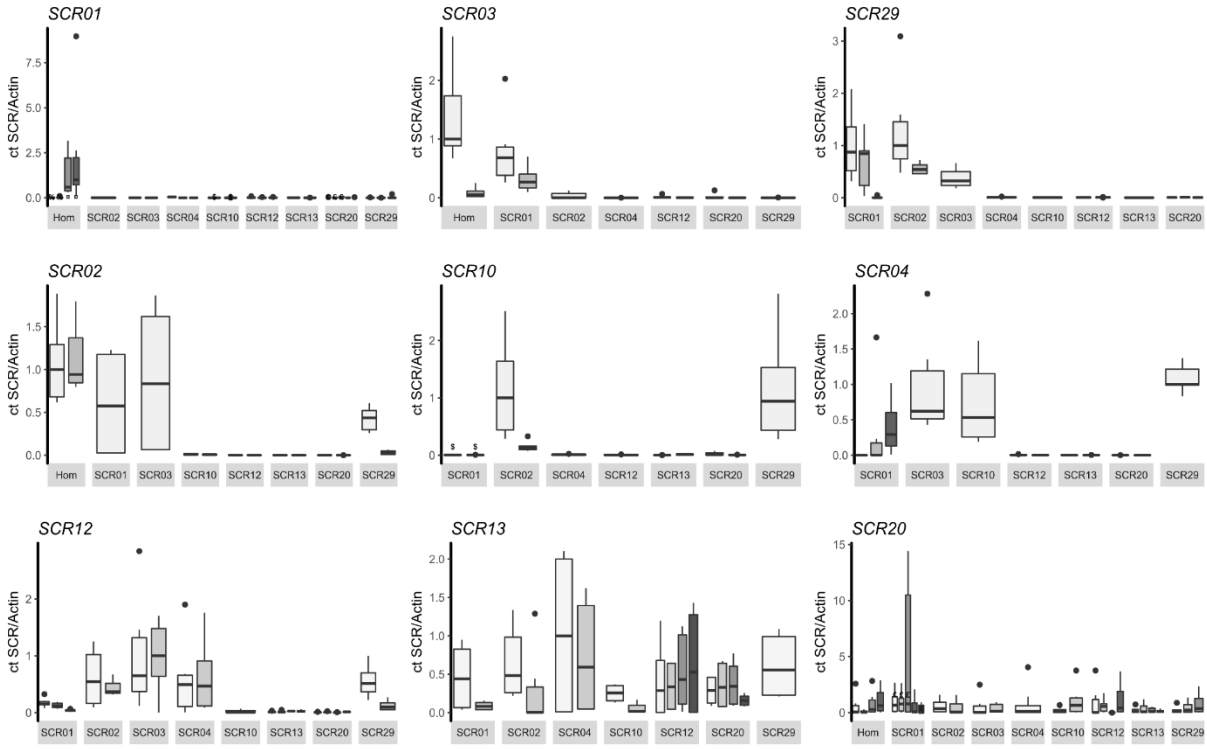


Figure S3. Expression of individual *SCR* alleles in different genotypic contexts, representing each biological and clone replicate separately. Symbols on top of the boxes indicate measures from identical clone replicates. See legend of Figure 2 for a full description.

```
lmer(log(s1$Ct_SCR.actine) ~
allele_measured:stade+stade*dom_phenotype+

(1|allele_measured/replicatBiol_genotype/replicat_Techclone) ,
                data =s1, na.action=na.omit)
```

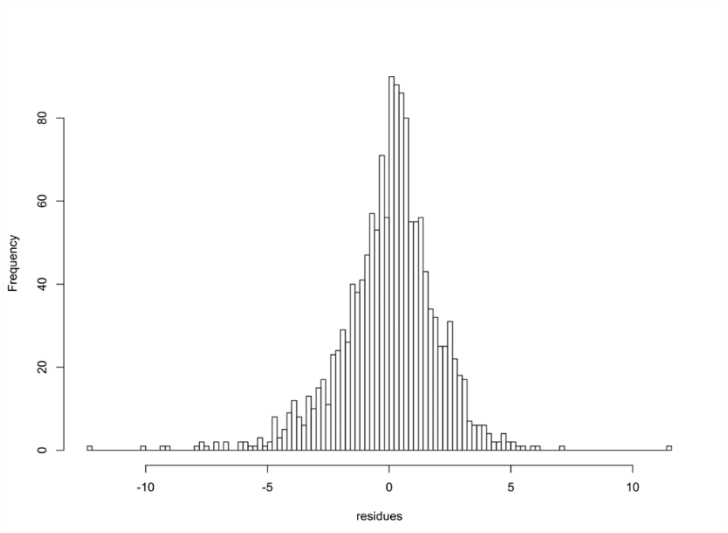


Figure S4. Generalized linear mixed model used to test the effect of developmental stage and dominance status on the expression of *SCR* alleles (Ct values). The distribution shows that the residues of the full model are approximately normally distributed when taking allele identity, developmental stage and dominance status into account and using a logarithmic transformation of the Ct/actin ratios.

	number of biological replicates	number of expression measure		
		allele 1	allele 2	actin
S1S1	2[$]	48	48	48
S1S2	1	12	12	12
S1S3	2	24	24	24
S1S4	3	36	36	36
S1S10	1[$]	24	24	24
S1S12	3	36	36	36
S1S13	2	24	24	24
S1S20	3[£]	60	60	60
S1S29	3	36	36	36
S2S2	2	24	24	24
S2S3*	2	18	18	18
S2S10	2	24	24	24
S2S12	2	24	24	24
S2S13	2	24	24	24
S2S20	2	24	24	24
S2S29	2	24	24	24
S3S3	2	24	24	24
S3S4	1	12	12	12
S3S12	2	24	24	24
S3S20*	3	30	30	30
S3S29	1	12	12	12
S4S10	1	12	12	12
S4S12	2	24	24	24
S4S13	2	24	24	24
S4S20	2	24	24	24
S4S29	1	12	12	12
S10S12	1	12	12	12
S10S13	2	24	24	24
S10S20	2	24	24	24
S10S29	1	12	12	12
S12S13	4	48	48	48
S12S20	4	48	48	48
S12S29	2	24	24	24
S13S20	4	48	48	48
S13S29	1	12	12	12
S20S20	4	48	48	48
S20S29	3	36	36	36

*: only the stages C and D were sampled for one of the biological replicates
$: two clone replicates per biological replicate
£: two of the three biological replicates are represented by two clone replicates

Table S1. *SCR* samples analysed for each S-locus genotype, showing the number of biological and clone replicates over the four developmental stages sampled. “Allele 1” refers to the first allele noted in the genotype (for example in the S1S2 genotype, “allele 1” is S1 and “allele 2” is S2).

	number of expression measure		
	allele 1	allele 2	actin
S1S1	12	-	11
S1S2	10	-	12
S1S3	11	-	12
S1S10	9	12	11
S1S12	11	12	12
S1S13	11	-	12
S1S20	8	-	9
S1S29	10	-	9
S3S3	12	-	12
S3S4	8	-	12
S3S12	12	11	12
S3S20	12	-	12
S3S29	12	12	12
S4S10	12	12	12
S4S12	12	9	9
S4S13	8	-	12
S4S29	12	11	12
S10S2	12	-	12
S10S12	11	-	12
S10S13	12	-	12
S10S20	12	-	12
S10S29	12	12	12
S12S20	12	-	12
S12S29	11	12	12
S29S2	12	-	12

Table S2: *SRK* samples analysed for each S-locus genotype, showing the number of biological and clone replicates over the four developmental stages sampled. The alleles are named accordingly to the Table S1.

pollen genotype	pistil phenotype	number of compatible crosses	dominance phenotype in pollen	reference	dominance phenotype in pistil	reference
S1S1	-	-	-	Durand et al. 2014	-	-
S1S2	[S1]	-	S2>S1	Llaurens et al. 2008	S2>S1	Llaurens et al. 2008
S1S2	[S2]	-		Llaurens et al. 2008		
S1S3	[S1]	-	S3>S1	Durand et al. 2014	-	-
S1S3	[S3]	-		Durand et al. 2014		-
S1S4	[S1]	-	S4>S1	Durand et al. 2014	S4>S1	Llaurens et al. 2008
S1S4	[S4]	-		Durand et al. 2014		
S1S10	[S1]	-	-	-	-	-
S1S10	[S10]			-		-
S1S12	[S1]	-	S12>S1	Durand et al. 2014	S12>S1	Llaurens et al. 2008
S1S12	[S12]	-		Durand et al. 2014		
S1S13	[S1]	-	S13>S1	Durand et al. 2014	-	-
S1S13	[S13]	-		Durand et al. 2014		-
S1S20	[S1]	-	S20>S1	Durand et al. 2014	S20>S1	Llaurens et al. 2008
S1S20	[S20]	-		Durand et al. 2014		
S1S29	[S1]	-	-	-	-	-
S1S29	[S29]			-		-
S2S2	-	-	-	Llaurens et al. 2008	-	-
S2S3	[S2]	-	-	-	-	-
S2S3	[S3]	-		-		-
S2S4	[S2]	-	S4>S2	Llaurens et al. 2008	S4>S2	Llaurens et al. 2008
S2S4	[S4]	-		Llaurens et al. 2008		
S2S10	[S2]	-	-	-	-	-
S2S10	[S10]	-		-		-
S2S12	[S2]	-	S12>S2	Llaurens et al. 2008	S12>S2	Llaurens et al. 2008
S2S12	[S12]	-		Llaurens et al. 2008		
S2S13	[S2]	-	-	-	-	-
S2S13	[S13]	-		-		-
S2S20	[S2]	-	-	-	-	-
S2S20	[S20]	-		-		-
S2S29	[S2]	0/5	S2>S29	This study	-	-
S2S29	[S29]	4/7		This study		-
S3S3	-	-	-	Durand et al. 2014	-	-
S3S4	[S3]	-	S4>S3	Durand et al. 2014	-	-
S3S4	[S4]	-		Durand et al. 2014		-
S3S10	[S3]	-	S10>S3	Durand et al. 2014	-	-
S3S10	[S10]	-		Durand et al. 2014		-
S3S12	[S3]	-	S12>S3	Durand et al. 2014	-	-
S3S12	[S12]	-		Durand et al. 2014		-
S3S13	[S3]	-	S13>S3	Durand et al. 2014	-	-
S3S13	[S13]	-		Durand et al. 2014		-
S3S20	[S3]	-	S20>S3	Durand et al. 2014	-	-
S3S20	[S20]	-		Durand et al. 2014		-
S3S29	[S3]	-	S29>S3	-	-	-
S3S29	[S29]	-		-		-
S4S4	-	-	-	Durand et al. 2014	-	-
S4S10	[S4]	-	S4>S10	Durand et al. 2014	-	-
S4S10	[S10]	-		Durand et al. 2014		-
S4S12	[S4]	-	S12>S4	Durand et al. 2014	S12>S4	Llaurens et al. 2008
S4S12	[S12]	-		Durand et al. 2014		
S4S13	[S4]	-	S13>S4	Durand et al. 2014	-	-
S4S13	[S13]	-		Durand et al. 2014		-

S4S20	[S4]	-	S20>S4	Durand et al. 2014	S20>S4	Llaurens et al. 2008
S4S20	[S20]	-		Durand et al. 2014		
S4S29	[S4]	-	S4>S29	Durand et al. 2014	-	-
S4S29	[S29]	-		Durand et al. 2014		-
S10S10	-	-	-	Durand et al. 2014	-	-
S10S12	[S10]	5/5	S12>S10	This study	-	-
S10S12	[S12]	0/5		This study		-
S10S13	[S10]	-	S13>S10	Durand et al. 2014	-	-
S10S13	[S13]	-		Durand et al. 2014		-
S10S20	[S10]	-	S20>S10	Durand et al. 2014	-	-
S10S20	[S20]	-		Durand et al. 2014		-
S10S29	[S10]	1/5	S10>S29	This study	-	-
S10S29	[S29]	3/3		This study		-
S12S12	-	-	-	Durand et al. 2014	-	-
S12S13	[S12]	-	S13>S12	Durand et al. 2014	-	-
S12S13	[S13]	-		Durand et al. 2014		-
S12S20	[S12]	-	S20>S12	Durand et al. 2014	S20>S12	Llaurens et al. 2008
S12S20	[S20]	-		Durand et al. 2014		
S12S29	[S12]	-	S12>S29	Durand et al. 2014	-	-
S12S29	[S29]	-		Durand et al. 2014		-
S13S13	-	-	-	Durand et al. 2014	-	-
S13S20	[S13]	-	S13=S20	Durand et al. 2014	-	-
S13S20	[S20]	-		Durand et al. 2014		-
S13S29	[S13]	-	S13>S29	Durand et al. 2014	-	-
S13S29	[S29]	-		Durand et al. 2014		-
S20S20	-	-	-	Durand et al. 2014	-	-
S20S29	[S20]	-	S20>S29	Durand et al. 2014	-	-
S20S29	[S29]	-		Durand et al. 2014		-
S29S29	-	-	-	Durand et al. 2014	-	-

Table S3: Dominance relationships between alleles from the different genotypes included in this study as determined by controlled crosses.

targeting allele	targeting sRNA	targeted allele	dominance phenotype	sRNA/target interaction				score	ct/actin	target location	relative from	relative to
Ah01	Ah01_Mir4239			sRNA	3'	AAAAUAAAAGCAAAAACGACGUA	5'	14				
						~~||||||--|||||x||||||						
		Ah01	H	target	5'	TATTATTTT--TTTTTACTGCAT	3'		1.1342E-02	promotor	-598	-581
	Ah01_Mir4239			sRNA	3'	AAAAU-AAAAGCAAAAACGAC-GUA	5'	13.5				
						~~||-|||||-||||||o||-||~						
		Ah02	D	target	5'	UAUUAUUUUUC-UUUUUGUUGUCAA	3'		1.3053E-01	intron	-196	-177
	Ah01_Mir4239			sRNA	3'	AAAAUA-AAAGCAAAAAC-GACGUA	5'	15				
						||||-|-|||||||||-||~~~~						
		Ah03	D	target	5'	UUUU-UCUUUCGUUUUUGUCUAUUU	3'		2.8473E-02	promotor	-505	-486
	Ah01 Mir4239			sRNA	3'	AAAUAAAAGCAAA-A-ACGACGUAC-UA	5'	14				
						||||x||||-|||-|-|-||-||||-||						
		Ah04	D	target	5'	UUUACUUUC-UUUAUAU-CU-CAUGUAU	3'		1.2298E-01	promotor	-517	-493
	Ah01 Mir4239			sRNA	3'	AGCAAAAACG-A-CGUAC-UAGGU--UCC	5'	15.5				
						||x|||||||-|-|o|||-|||||--|||						
		Ah10	D	target	5'	UCAUUUUUGCAUCGUAUGUAUCCAUCAGG	3'		2.8487E-04	exon	103	131
	Ah01 Mir4239			sRNA	3'	AAAAUAAA-AGCAAA-AA-CGACGUA	5'	14.5		2		
						||||||||-|x||||-||-|o|-|||						
		Ah12	D	target	5'	UUUUAUUUGUAGUUUAUUAGUU-CAU	3'		2.1209E-02	downstream	530	554
	Ah01_Mir4239			sRNA	3'	AG-CAAAAACGACGUACUAGGUUCC	5'	13.5				
						||-|||||||||x|o||---|||||						
		Ah13	D	target	5'	UCUGUUUUUGCUACGUG---CAAGG	3'		2.0399E-02	downstream	1940	1961
	Ah01_Mir4239			sRNA	3'	UCGUACGCCGUUUUUGCCUAUGUU	5'	16.5				
						~~||||o|||||x||||||x||||						
		Ah20	D	target	5'	CACAUGUGGCAAUAACGGAAACAA	3'		2.0770E-01	exon	55	76
	Ah01_Mir4239			sRNA	3'	AGCAAAAACGACGUACUA-GGUUCC	5'	15				
						~~|||||x||||||-||-|||||~						
		Ah29	D	target	5'	UUUUUUUUUCUGCAU-AUGCCAAGA	3'		4.9834E-02	downstream	910	929

Ah02	Ah02_Mir1887			sRNA	3'	GUGAAUCUAACAAGAAUCAUACUC	5'	19				
						~~||||||||||||||x|||||~						
		Ah01	R	target	5'	CUCUUAGAUUGUUCUUACUAUGAA	3'		5.4584E-05	promotor	-59	-39
	Ah02_MirS3			sRNA	3'	AACUAACCAAUGAGCAUUUU-CA	5'	15				
						||||||x|x|||-|||||||-||						
		Ah02	H	target	5'	UUGAUUUGCUAC-CGUAAAACGU	3'		2.3912E-01	upstream	-1545	-1524
	Ah02_Mir1887			sRNA	3'	GUGAAUCUAACAAGAAUCAUACUC	5'	16.5				
						~~||||||||x||||||||o|x||						
		Ah03	R	target	5'	CUCUUAGAUUAUUCUUAGUGUAAG	3'		1.8790E-03	promotor	-108	-87
	Ah02 MirS3			sRNA	3'	AACUAACCAAUGAGCAUUUUCA	5'	14				
						|||||||||||x|--||||~~						
		Ah04	D	target	5'	UUGAUUGGUUACAC--AAAAAA	3'		Na	intron	-60	-43
	Ah02 MirS3			sRNA	3'	AACUAACCAAUGAG-CAUUUUCA	5'	15.5				
						||||||||||||o-|-|||~~						
		Ah10	D	target	5'	UUGAUUGGUUACUUAG-AAAACU	3'		4.1045E-02	intron	-56	-37
	Ah02_MirS3			sRNA	3'	UUUAA-ACG-CUUAUU-GGUUAGUGU	5'	16				
						|||||-|x|-||||||-|||||||~~						
		Ah12	D	target	5'	AAAUUAUUCAGAAUAAGCCAAUCAGA	3'		9.6584E-02	intron	-601	-578
	Ah02_Mir1887			sRNA	3'	GUGAAUCUAA-CA-AGAAUCAUACUC	5'	14				
						~||||-||||-||-|||||||~~~~~						
		Ah13	D	target	5'	AACUU-GAUUAGUAUCUUAGUCACAA	3'		3.4676E-02	downstream	203	221
	Ah02_MirS3			sRNA	3'	UUUAAA-CGCUUAU-U--G-GUUAGUGU	5'	14				
						||||||-||x||||-|--|-||||||~~						
		Ah20	D	target	5'	AAAUUUAGCCAAUAUAUGCUCAAUCAAA	3'		6.9471E-02	downstream	1611	1636
	Ah02 Mir1887			sRNA	3'	GUGAAUCUAACAAGAAUCAUACUC	5'	16				
						|o|||||||x|||||||||o|~~~						
		Ah29	R	target	5'	CGCUUAGAUAGUUCUUAGUGUAAU	3'		8.1926E-02	promotor	-53	-33

Ah03	Ah03_MirS3			sRNA	3'	UACAAGUUCCAUUAUAUAAUCGAA	5'	22				
						||||||||||||||||x|||||						
		Ah01	R	target	5'	AUGUUCAAGGUAAUAUAUGAGCUU	3'		4.2927E-05	exon	59	82
	Ah03_MirS3			sRNA	3'	AGUUCCAUUAUAUAAUC-GAAGAAA	5'	17.5				
						|||||||||x||x|||-o|||||						
		Ah02	D	target	5'	UCAAGGUAAUUUACUAGUUUUCUUU	3'		1.9047E-01	intron-exon boundary	-19	5
	Ah03_MirS5			sRNA	3'	AUACGAAAGUACAGA--AAAAAAAG	5'	16.5				
						||||o||||||||-|--|||||||~						
		Ah03	H	target	5'	UAUGUUUUCAUGU-UGGUUUUUUUA	3'		3.6834E-02	exon	-21	1
	Ah03 MirS3			sRNA	3'	ACA-AGUU-CCAUUAUAUAAUCGAAG	5'	15				
						|||-||||-||-||||||x||-|||~						
		Ah04	D	target	5'	UGUGUCAAUGG-AAUAUACUA-CUUU	3'		5.0001E-01	intron	-1221	-1199
	Ah03 MirS5			sRNA	3'	AAAA-AAG-GUUCAGUACAAUUUC	5'	16				
						||||-|||-||||x|||x||||||						
		Ah10	D	target	5'	UUUUCUUCACAAGCCAUAUUAAAG	3'		Na	exon	84	107
	Ah03_MirS3			sRNA	3'	ACAAGUUCCAUUAUAUAAUCGAAG	5'	16				
						|||||||||||||-|-x||-||||						
		Ah12	D	target	5'	UGUUCAAGGUAAU-U-CUA-CUUC	3'		1.8389E-01	promotor	-569	-549
	Ah03_Mirs3			sRNA	3'	AGUUCCAUUAUAUAA-U-CGAAGAAA	5'	18.5				
						|||||||||||||||-|-x|||o|||						
		Ah13	D	target	5'	UCAAGGUAAUAUAUUAACACUUUUUU	3'		Na	intron-exon boundary	63	88
	Ah03_MirS5			sRNA	3'	AUACGAAAGUACAGAA-AAA--AAAG	5'	16				
						|||-||||x|||||||-|||--||||						
		Ah20	D	target	5'	UAU-CUUUAAUGUCUUGUUUAAUUUC	3'		6.2263E-02	downstream	1478	1502
	Ah03 MirS3			sRNA	3'	UACAAGUUCCAUUAUAUAAUCGAA	5'	18.5				
						|||||||||||||x||x|||o||						
		Ah29	R	target	5'	AUGUUCAAGGUAAUUUACUAGUUU	3'		3.3086E-02	intron-exon boundary	-8	15

Ah04	Ah04_MirS4			sRNA	3'	UAU-ACUUUUUUC-UUUUUUCU-UU	5'	17				
						|||-||-||||||-||||||||-||						
		Ah01	R	target	5'	AUAUUG-AAAAAGUAAAAAAGAGAA	3'		1.7497E-04	upstream	-1173	-1150
	Ah04_Mir867			sRNA	3'	ACAGAAAGGAUUAUUUGGUACAAGUU	5'	20.5				
						|||||||||x||||o|||||x||||						
		Ah02	R	target	5'	UGUCUUUCCUUAUAAGCCAUGGUCAA	3'		Na	exon	2	27
	Ah04_MirS4			sRNA	3'	GUAUGAUUCUUGUUAGAUUCA	5'	15.5				
						~~||||||||o||||||||~						
		Ah03	R	target	5'	ACUACUAAGAAUAAUCUAAGA	3'		1.8947E-05	intron-exon boundary	-20	3
	Ah04 MirS4			sRNA	3'	AA-AA-GAAAC-AGUAUUGUGAUAAAGA	5'	18.5				
						||-||-|||||-||||||o|||||||~~						
		Ah04	H	target	5'	UUCUUACUUUGAUCAUAAUACUAUUUAU	3'		Na	intron	-2057	-2032
	Ah04 MirS4			sRNA	3'	AUAGGUCUUUGUCGCUAACAAUGA	5'	19				
						||||||x|||o||o||||||||||						
		Ah10	R	target	5'	UAUCCAAAAAUAGUGAUUGUUACU	3'		6.6149E-04	intron	-872	-849
	Ah04_MirS4			sRNA	3'	UAUAC-UUU-UUUC-UUUUUUCUUU	5'	16				
						~||x|-|||-||||-||||||||||						
		Ah12	D	target	5'	UUAAGAAAAGAAAGAAAAAAAGAAA	3'		1.1418E-01	promotor	-273	-250
	Ah04_MirS4			sRNA	3'	UAUACUUUUUU--C-UUUUU--UCUUU	5'	16				
						~||||||||||--|-|||||--|||||						
		Ah13	D	target	5'	UUAUGAAAAAAGUGUAAAAAGUAGAAA	3'		6.5069E-02	downstream	1024	1049
	04_Mir4239_copy1			sRNA	3'	CCUCGUACACCUUUAUUGCCUUUG	5'	20				
						|||x||||||x||||||||||||						
		Ah20	D	target	5'	GGAACAUGUGGCAAUAACGGAAAC	3'		9.5550E-02	exon	51	74
	Ah04 Mir867			sRNA	3'	ACAGAAAGGAUUAUUUGGUACAAGUU	5'	20.5				
						|||||||x||x||||o|||||||||						
		Ah29	R	target	5'	UGUCUUUGCUUAUAAGCCAUGUUCAA	3'		8.0761E-04	exon	40	65

Ah10	Ah10_Mir1887			sRNA	3'	GGAGUAUGAUUCUUGUUAGAUUCA	5'	19.5				
						|o|||||x|||||||||||||~						
		Ah01	R	target	5'	CUUCAUAGUAAGAACAAUCUAAGA	3'		2.5148E-05	promotor	-61	-39
	Ah10_Mir867			sRNA	3'	AGAAAGGAAUAUUCGGUACAAGUU	5'	22				
						||||||||||||||||||x||||						
		Ah02	R	target	5'	UCUUUCCUUAUAAGCCAUGGUCAA	3'		2.0406E-03	exon	2	25
	Ah10_Mir1887			sRNA	3'	GGAGUAUGAUUCUUGUUAGAUUCA	5'	18				
						|||o|x|||||||o|||||||~						
		Ah03	R	target	5'	CCUUACACUAAGAAUAAUCUAAGA	3'		Na	promotor	-109	-87
	Ah10 Mir867			sRNA	3'	GCAACUUUUAACAAAUAUUCCU-UU	5'	17				
						~~||x|||||x||||||||||-||						
		Ah04	R	target	5'	UUUUCAAAAUGGUUUAUAAGGAGAA	3'		3.7813E-01	exon	90	112
	Ah10 Mir4239			sRNA	3'	UUGUUUC-UUACGUUUGUUCUG-GAA	5'	14.5				
						|||||||-||||--|||||-|o-|||						
		Ah10	H	target	5'	AACAAAGAAAUG--AACAA-AUGCUU	3'		Na	dowstream	136	158
	Ah10_Mir867			sRNA	3'	AGAAAGGAAUAUUCGGUACAAGUU	5'	15				
						||||||o||-|||xox||||||||						
		Ah12	D	target	5'	UCUUUCUUU-UAAAUGAUGUUCAA	3'		4.2743E-03	promotor	-555	-533
	Ah10_Mir4239			sRNA	3'	UUGUUUCUUAC-GUUUGUUCUGGAA	5'	15				
						|||-|||||||-|o||||x||o||~						
		Ah13	D	target	5'	AAC-AAGAAUGACGAACACGAUCUC	3'		1.1520E-02	downstream	714	736
	Ah10_Mir4239			sRNA	3'	CCUCGUACACCUUUAUUGCCUUUGU	5'	21				
						|||x|||||||x||||||||||||						
		Ah20	D	target	5'	GGAACAUGUGGCAAUAACGGAAACA	3'		8.7536E-02	exon	51	75
	Ah10 Mir867			sRNA	3'	AGAAAGGAAUAUUCGGUACAAGUU	5'	22				
						|||||x||||||||||||||||						
		Ah29	R	target	5'	UCUUUGCUUAUAAGCCAUGUUCAA	3'		4.2462E-04	exon	42	65

Ah12	Ah12_Mirs3			sRNA	3'	UUAGUUUUGGAUUUCCUCCAACUAU	5'	21				
						||||||||||x||||||x|||||||						
		Ah01	R	target	5'	AAUCAAAACCAAAAGGAUGUUGAUA	3'		1.8177E-04	intron	-238	-215
	Ah12_MirS3			sRNA	3'	UUUAGUUUUGGAUUUCCUCCAACUA	5'	21				
						|||||||||||x||||||x||||||						
		Ah02	R	target	5'	AAAUCAAAACCAAAAGGAUGUUGAU	3'		6.0314E-05	intron	-940	-916
	Ah12_MirS3			sRNA	3'	UUAGUUUUGGAUUUCCUCCAACUAU	5'	21.5				
						||||||||o|||||||x|||||||						
		Ah03	R	target	5'	AAUCAAAACUUAAAGGAUGUUGAUA	3'		4.1965E-04	intron	-211	-187
	Ah12 MirS3			sRNA	3'	UUAGUUUUGGA-UUUCCUCCAACUAU	5'	20				
						||||||||-||-||||||x|||||||						
		Ah04	R	target	5'	AAUCAAAA-CUAAAAGGACGUUGAUA	3'		1.6505E-03	intron	-79	-55
	Ah12 MirS3			sRNA	3'	UUUAGUUUUGGA-UUUCCUCCAACUA	5'	19				
						~||||||||-||-||||||x||||||						
		Ah10	R	target	5'	AAAUCAAAA-CUAAAAGGACGUUGAU	3'		3.2426E-04	intron	-74	-51
	Ah12_MirS3			sRNA	3'	GACGAAAG-AAAAUCGAAACUAAGC	5'	15.5				
						~~||||-||||o|||||||||~~						
		Ah12	H	target	5'	UUUCUUUCUUUUUGGCUUUGAUUGG	3'		Na	downstream	123	123
	Ah12_MirS3			sRNA	3'	UUUA-GUUUUGGAUUUCCUCCAACUA	5'	14				
						~~||-|||||-x|||||||--|||||						
		Ah13	D	target	5'	AUAUCCAAAA-AUAAAGGA--UUGAU	3'		3.5946E-02	promotor	-179	-159
	Ah12_MirS3			sRNA	3'	UUUAGUUUUGGA-UUUCC--UCCAACUA	5'	15.5				
						~||||||||-||-|||||--|x||||o|						
		Ah20	D	target	5'	AAAUCAAAA-CUAAAAGGAUAUGUUGGU	3'		9.0139E-02	intron	-193	-168
	Ah12 MirS3			sRNA	3'	UUAGUUUUGGA-UUUCCUCCAACUAU	5'	20				
						||||||||-||-||||||x|||||||						
		Ah29	R	target	5'	AAUCAAAA-CUGAAAGGACGUUGAUA	3'		5.8038E-04	intron	-225	-201

Ah13	Ah13_MirS3			sRNA	3'	CUCAUUGGUUAGUUUUGGAUUUCCU	5'	22				
						~|||||||||||||||x||||||						
		Ah01	R	target	5'	AAGUAACCAAUCAAAACCAAAAGGA	3'		6.6402E-05	intron	-231	-208
	Ah13_MirS3			sRNA	3'	UAGUUUUGGAUUUCCUUCAACUAU	5'	20				
						|||||||||x||||||x|||||||						
		Ah02	R	target	5'	AUCAAAACCAAAAGGAUGUUGAUA	3'		7.0608E-05	intron	-938	-915
	Ah13_MirS3			sRNA	3'	UAGUUUUGGAUUUCCUUCAACUAU	5'	20.5				
						||||||||o|||||||x|||||||						
		Ah03	R	target	5'	AUCAAAACUUAAAGGAUGUUGAUA	3'		Na	intron	-211	-188
	Ah13 MirS3			sRNA	3'	CUCAUUGGUUAGUUUUGGA-UUUCCU	5'	20				
						~~|||||||||||||||-||-||||||						
		Ah04	R	target	5'	GUGUAACCAAUCAAAA-CUAAAAGGA	3'		7.3639E-04	intron	-71	-49
	Ah13 MirS3			sRNA	3'	CU-AAAACGACUCAUUGGUUAGUUUUG	5'	21				
						||-||||-||x|||||||||||||||						
		Ah10	R	target	5'	GAGUUUU-CUAAGUAACCAAUCAAAAC	3'		4.9442E-04	intron	-59	-34
	Ah13_MirS2			sRNA	3'	AUAUAUACGCAUUACUGUAUUUGU	5'	22				
						|||||||||||||||||||||||~~						
		Ah12	R	target	5'	UAUAUAUGCGUAAUGACAUAAAAA	3'		3.5373E-03	promotor	-470	-449
	Ah13_MirS3			sRNA	3'	AAGAAAA-UCGAA-ACUAACCAAU	5'	15				
						||-||||-||-||-||||||||~						
		Ah13	H	target	5'	UU-UUUUGAG-UUCUGAUUGGUUG	3'		Na	intron	-95	-75
	Ah13_MirS3			sRNA	3'	UUAGUUUUGGA-UUUCCU-U-CAACUA	5'	17.5				
						||||||||-||-||||||-|-||||o|						
		Ah20	D	target	5'	AAUCAAAA-CUAAAAGGAUAUGUUGGU	3'		3.1419E-02	intron	-193	-168
	Ah13 MirS3		D	sRNA	3'	UUUGGA-UUUCCUUCAACUAUAUAU	5'	19				
						|||x||-||||||x|||||||||||						
		Ah29	R	target	5'	AAAACUGAAAGGACGUUGAUAUAUA	3'		1.4935E-04	intron	-229	-205

Ah20	Ah20_MirS3			sRNA	3'	AUAGUUGAAGGAAAACCAAAACUAG	5'	22				
						|||||||x|||||||||||||~						
		Ah01	R	target	5'	UAUCAACAUCCUUUUGGUUUUGAUU	3'		1.3491E-04	intron	-239	-216
	Ah20_MirS3			sRNA	3'	AUAGUUGAAGGAAAACCAAAACUAG	5'	22				
						|||||||x|||||||||||||~						
		Ah02	R	target	5'	UAUCAACAUCCUUUUGGUUUUGAUU	3'		1.9359E-05	intron	-938	-915
	Ah20_MirS3			sRNA	3'	AUAGUUGAAGGAAAACCAAAACUAG	5'	18				
						|||||||x||||||xx||||||||~						
		Ah03	R	target	5'	UAUCAACAUCCUUUAAGUUUUGAUU	3'		6.3463E-04	intron	-211	-188
	Ah20 MirS3			sRNA	3'	AUAGUUGAAGGAAAACCAAAACUAG	5'	20				
						|||||||x|||||||x||||||||~						
		Ah04	R	target	5'	UAUCAACGUCCUUUUAGUUUUGAUU	3'		3.0334E-04	intron	-79	-56
	Ah20 MirS3			sRNA	3'	UAGUUGAAGGAAAACCAAAACUAG	5'	19				
						||||||x|||||||x||||||||~						
		Ah10	R	target	5'	AUCAACGUCCUUUUAGUUUUGAUU	3'		9.5621E-04	intron	-74	-52
	Ah20_MirS2			sRNA	3'	CGUCGUAUUGUGCAUUUGUGUAU	5'	21.5				
						||||||||||||||||o||||						
		Ah12	R	target	5'	GCAGCAUAACACGUAAACGCAUA	3'		1.4779E-03	promotor	-501	-479
	Ah20_MirS2			sRNA	3'	CACAAAUAC-ACAAAUUCACAAUAC	5'	16				
						||-||||||-|-||||||x|||||~						
		Ah13	D	target	5'	GU-UUUAUGAU-UUUAAGCGUUAUU	3'		2.1991E-02	downstream	1322	1343
	Ah20_MirS3			sRNA	3'	GUUG-A-AGGAAAACCAAAACUAGCC	5'	17.5				
						||||-|-|||||||x||||||||o|~						
		Ah20	H	target	5'	CAACAUAUCCUUUUAGUUUUGAUUGU	3'		8.6244E-02	intron	-191	-167
	Ah20 MirS2			sRNA	3'	CACAAAUACAGAAACACAUAUACA	5'	18				
						|||||||||x|||-|||||||~~						
		Ah29	R	target	5'	GUGUUUAUGUGUUU-UGUAUAUUC	3'		6.2265E-04	promotor	-205	-185

Brassica S44	S44_Smi2			sRNA	3'	UUCUGUAUGUGCUUAUUUCACACA	5'	17		-
						|o||x|x|||||||||o||||||				
		S29-pro	R	target	5'	AGGAAACACACGAAUAAGGUGUGU	3'		-	-
	S44_Smi2			sRNA	3'	UUC-UGUAUGUGCUUAUUUCACACA	5'	16.5		-
						~||-|||x||||x|||||o||||||				
		S40-pro	R	target	5'	AAGUACACACACAAAUAAGGUGUGU	3'		-	-
	S44_Smi2			sRNA	3'	UUCUGUAUGUGCUUAUUUCACACA	5'	16.5		-
						~~|||x|||||||||o||||\~				
		S60-pro	R	target	5'	ACAACACACACGAAUAAGGUGUGA	3'		-	-
	S44_Smi2			sRNA	3'	UUCUGUAUGUGCUUAUUUCACACA	5'	13.5		-
						|o||o|x||o||||-||||-||||				
		S44-pro	H	target	5'	AGGAUACACGCGAA-AAAG-GUGU	3'		-	-
Brassica S60	S60_Smi2			sRNA	3'	UUCUAUAUGUGUUUAUUCCACACA	5'	17		-
						|o||x|x|||o||||||||||||				
		S29-pro	R	target	5'	AGGAAACACACGAAUAAGGUGUGU	3'		-	-
	S60_Smi2			sRNA	3'	UUC-UAUAUGUGUUUAUUCCACACA	5'	18		-
						~||-|x|x|||||||||||||||				
		S40-pro	R	target	5'	AAGUACACACACAAAUAAGGUGUGU	3'		-	-
	S60_Smi2			sRNA	3'	UUCUAUAUGUGU--UUAUUCCACACA	5'	15		-
						~~||x|||||--||||||||||\~				
		S60-pro	H	target	5'	UACAUACACACACGAAUAAGGUGUGA	3'		-	-
	S60_Smi2			sRNA	3'	UUCUAUAUGUG-UUUAUUCCACACA	5'	14		-
						|o||||x||o|-|||-|||||||~~				
		S44-pro	D	target	5'	AGGAUACACGCGAAA-AAGGUGUAU	3'		-	-

Brassica S40	S40_Smi2			sRNA	3'	UUCUAUAUGUGCUUAUUUCACACA	5'	17		-
						|o||x|x|||||||||o||||||				
		S29-pro	R	target	5'	AGGAAACACACGAAUAAGGUGUGU	3'		-	-
	S40_Smi2			sRNA	3'	UUC-UAUAUGUGCUUAUUUCACACA	5'	14.5		-
						~||-|x|x||||x|||||o||||||				
		S40-pro	H	target	5'	AAGUACACACACAAAUAAGGUGUGU	3'		-	-
	S40_Smi2			sRNA	3'	UUCUAUAUGUGCUUAUUUCACACA	5'	14.5		-
						~~||||||||||o||||\~				
		S60-pro	D	target	5'	ACACGAAACACGAAUAAGGUGUGA	3'		-	-
	S40_Smi2			sRNA	3'	UUCUAUAUGUGCUUAUUUCACACA	5'	15		-
						|o||||x||o||||-||||-||||				
		S44-pro	D	target	5'	AGGAUACACGCGAA-AAAG-GUGU	3'		-	-
Brassica S29	S29_Smi2			sRNA	3'	UUCUAUAUGUGCUUAUUGCACACA	5'	16.5		-
						|o||x|x|||||||||x||||||				
		S29-pro	H	target	5'	AGGAAACACACGAAUAAGGUGUGU	3'		-	-
	S29_Smi2			sRNA	3'	UUC-UAUAUGUGCUUAUUGCACACA	5'	14		-
						~||-|x|x|||x|||||x||||||				
		S40-pro	D	target	5'	AAGUACACACACAAAUAAGGUGUGU	3'		-	-
	S29_Smi2			sRNA	3'	UUCUAUAUGUGCUUAUUGCACACA	5'	14		-
						~~||||||||||x||||\~				
		S60-pro	D	target	5'	ACACGAAACACGAAUAAGGUGUGA	3'		-	-
	S29_Smi2			sRNA	3'	UUCUAUAUGUGCUUAUUGCACACA	5'	13		-
						|o|||x||o|||x||x|||\~\~				
		S44-pro	D	target	5'	AGGAUACACGCGAAAAAGGUGUAU	3'		-	-

Brassica S9	S9-Smi			sRNA	3'	ACAUUGAUAAAAUGUGCAUUUGUA	5'	17		-
						~~||||||x|||||||||||				
		S29-SP11	R	target	5'	CUAUUCUAUUUCACACGUAAACAU	3'		-	-
	S9-Smi			sRNA	3'	ACAUUGAUAAAAUGUGCAUUUGUA	5'	17		-
						~~||||||x|||||||||||				
		S40-SP11	R	target	5'	CUAUUCUAUUUCACACGUAAACAU	3'		-	-
	S9-Smi			sRNA	3'	ACAUUGAUAAAAUGUGCAUUUGUA	5'	18		-
						~~||||||||||||||||||~				
		S60-SP11	R	target	5'	CUAUUCUAUUUUACACGUAAACAA	3'		-	-
	S9-Smi			sRNA	3'	ACAUUGAUAAAAUGUGCAUUUGUA	5'	17		-
						~~||||||x|||||||||||				
		S44-SP11	R	target	5'	CUAUUCUAUUUCACACGUAAACAU	3'		-	-
Brassica S60	S60-smi			sRNA		ACAUUGAUAAAAUGAGCAUUUGUA	5'			-
						~~||||||x||x||||||||				
		S29-SP11	R	target		CUAUUCUAUUUCACACGUAAACAU	3'	15	-	-
	S60-smi			sRNA		ACAUUGAUAAAAUGAGCAUUUGUA	5'			-
						~~||||||x||x||||||||				
		S40-SP11	R	target		CUAUUCUAUUUCACACGUAAACAU	3'	16	-	-
	S60-smi			sRNA		ACAUUGAUAAAAUGAGCAUUUGUA	5'			-
						~~||||||||x|||||||~				
		S60-SP11	H	target		CUAUUCUAUUUUACACGUAAACAA	3'	15	-	-
	S60-smi			sRNA		ACAUUGAUAAAAUGAGCAUUUGUA	5'			-
						~~||||||x||x||||||||				
		S44-SP11	D	target		CUAUUCUAUUUCACACGUAAACAU	3'	15	-	-

Table S4: sRNA and target identified as the best match for every pair of alleles for *SCR*. Ct/actin are given for the target allele in the interaction, calculated from the mean of Ct/actin across the two earliest developmental stages (buds below 1mm, see Figure 1). The position of the targets are given relative to the beginning of the closest exon of SCR for targets upstream from the gene or in the intron), and relative to the stop codon for downstream targets. R: Recessive; D: dominant; H: homozygote.

a. analyse of variance for biological, clone, technical replicates and allele's expression dynamic

lmer(log(Ct_SCR.actine) ~ stage*dom_sRNA_prediction+(1|allele_measured:stage)+(1|replicatBiol_genotype)+(1|replicat_Techclone), na.act

Random effects:	Groups	Name	Variance	Std.Dev.
	replicat_Techclone	(Intercept)	0.4091	0.6396
	replicatBiol_genotype	(Intercept)	1.0795	1.0390
	allele_measured:stage	(Intercept)	4.5674	2.1372
	Residual (Technical replicates)		6.0815	2.4461

---

b. statistical model to test the hypothesis of a variation of expression dynamic across allele

model 1: lmer(log(Ct_SCR.actine) ~ stage*dom_phenotype+(1|replicatBiol_genotype/replicat_Techclone))

model 2 : lmer(log(Ct_SCR.actine) ~ stage*dom_phenotype+(1|allele_measured:stage)+(1|replicatBiol_genotype/replicat_Techclone), na.act

	Df	AIC	BIC	logLik	deviance	Chisq Chi	Df	Pr(>Chisq)	
model 1:	11	6946.0	7004	-3462.0	6924.0				
model 2:	12	6730.6	6794	-3353.3	6706.6	217.32	1	< 2.2e-16	***

---

c. statistical model to test the hypothesis of a variation among stages and effect of dominance

lmer(log(Ct_SCR.actine)~(1|allele_measured:stage)+stade*dom_phenotype+(1|replicatBiol_genotype/replicat_Techclone), na.action=na.omit

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)		
stade	213.51	71.17	3	27.53	13.805	1.107e-05	***	
dom_phenotype	814.70	814.70	1	302.35	158.025	< 2.2e-16	***	
stade:dom_phenotype	596.73	198.91	3	1358.37	38.582	< 2.2e-16	***	

---

d.statistical model to test the hypothesis of a threshold model based on the alignment score that might explain the silencing phenotype in SCR

lmer(log(Ct_SCR.actine) ~ (1|allele_measured:stage)+stage+Dom_score_based_xx+(1|replicatBiol_genotype/replicat_Techclone) , na.action=

	Df	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)	
Dom_score_based_14	23	4064.9	4175.5	-2009.5	4018.9				
Dom_score_based_15	24	4066.1	4181.5	-2009.0	4018.1	0.8294	1	0.3624	
Dom_score_based_16	24	4056.0	4171.5	-2004.0	4008.0	10.0691	0	<2e-16	***
Dom_score_based_17	24	4040.2	4155.6	-1996.1	3992.2	15.8111	0	<2e-16	***
Dom_score_based_18	24	4025.9	4141.3	-1988.9	3977.9	14.3695	0	<2e-16	***
Dom_score_based_19	24	4045.7	4161.1	-1998.9	3997.7	0.0000	0	1.0000	
Dom_score_based_20	24	4054.9	4170.3	-2003.5	4006.9	0.0000	0	1.0000	
Dom_score_based_21	24	4063.4	4178.8	-2007.7	4015.4	0.0000	0	1.0000	
Dom_score_based_22	24	4064.9	4180.3	-2008.5	4016.9	0.0000	0	1.0000	

---

e. statistical model to test the hypothesis of a variation in the intensity of selencing regarding the position of the target site for SCR

lmer(log(Ct_SCR.actine)~allele_measured:stage+stage+position_target+(1|allele_measured/replicatBiol_genotype/replicat_Techclone), na.a

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)			
position_target	27.019	6.755	4	33.275	1.4432	0.241609			
---									

f. analyse of variance for biological, clone, technical replicates and allele's expression dynamic for SRK

lmer(log(Ct_SRK.actine) ~ stade*dom_phenotype+(1|allele_measured:stade)+(1|replicat_Techclone), na.action=na.omit)

Random effects:	Groups	Name	Variance	Std.Dev.					
	allele_measured:stade	(Intercept)	3.1451	1.7734					
	replicat_Techclone	(Intercept)	0.6974	0.8351					
	Residual		0.8089	0.8994					
---									

g. statistical model to test the hypothesis of a variation of expression dynamic of *SRK across alleles*

model 1 :lmer(log(Ct_SRK.actine) ~ stade*dom_phenotype+(1|replicatBiol_genotype/replicat_Techclone) , na.action=na.omit)

model 2 :lmer(log(Ct_SRK.actine) ~ (1|allele_measured:stade)+stade*dom_phenotype+(1|replicatBiol_genotype/replicat_Techclone), na.acti

	Df	AIC	BIC	logLik	deviance	Chisq Chi	Df	Pr(>Chisq)	
model 1:	11	280.74	307.07	-129.37	258.74				
model 2:	12	275.82	304.56	-125.91	251.82	6.9103	1	0.00857	**
---									

h. statistical model to test the hypothesis of a variation among stages and effect of dominance on *SRK*

lmer(log(Ct_SRK.actine) ~ (1|allele_measured:stade)+stade*dom_phenotype+(1|replicatBiol_genotype/replicat_Techclone), na.action=na.omit)

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)	
stade	2.0520	0.6840	3	3.346	0.8456	0.546472	
dom_phenotype	5.5723	5.5723	1	3.442	6.8884	0.068244	.
stade:dom_phenotype	10.7038	3.5679	3	64.867	4.4107	0.006943	**
---							

i. statistical model to test the hypothesis of an effect of the other allele on the expression measured for one SCR allele

lmer(log(Ct_SCR.actine) ~ (1|allele_measured:stade)+stage+Other_allele_inGenot+(1|replicatBiol_genotype/replicat_Techclone) , na.actio

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(>F)	
Other_allele_inGenot	52.061	10.412	5	32.843	2.2217	0.07558	.
---							

j. statistical model to test the hypothesis of an effect of age on aligement score

lm(score ~ age)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	0.066	0.06586	0.0362	0.8504
Residuals	29	52.708	1.81753		

Table S5: Detailed results from the generalized mixed model. **a**. Decomposition of the sources of variance across the hierarchical levels biological, clones and technical replicates. **b**. Test of the dominance and stage effects on *SCR* transcript levels, showing a significant interaction. **c**. Comparison of the fit of the model under different base-pairing score thresholds. **d**. Test of the effect of the position of the target on the strength of silencing. **e**. Test of the effect of stage and dominance on *SRK* transcript levels. **f**. Test of the effect of the identity of the companion allele on *SCR* transcript levels. **g**: Test of the effect of age on alignment score above the threshold of 18.

Chapitre 2

Polymorphisme intra-allélique et diversité naturelle des modificateurs de dominance au locus d'auto-incompatibilité : développement d'une approche par capture de séquences

Auteurs

Nicolas Burghgraeve, Mathieu Genete, Christelle Blassiau, Vincent Castric.

Introduction

La question des bases moléculaires et de l'évolution de la dominance a été l'objet d'un vif débat qui a largement divisé la communauté scientifique depuis les années 1930 (revue dans Billiard & Castric, 2011). Le phénomène de dominance a initialement été considéré comme résultant de l'action d'éléments génétiques dédiés, appelés « modificateurs de dominance » par R.A. Fisher (Fisher, 1928). Très rapidement cependant cette vision a été contestée et la dominance a été considérée dans le cadre de la théorie enzymatique comme une propriété émergente de la non-linéarité de la courbe dose-réponse lorsque les allèles formant le génotype hétérozygote sont associés à des activités enzymatiques différentes (Wright, 1934). Le point d'échappement principal de ce débat a concerné la faiblesse de l'intensité de la sélection naturelle qui agirait sur les hypothétiques modificateurs de dominance proposés par Fisher (1928) : quand bien même ils existeraient, la sélection serait tellement faible que leur importance évolutive en serait presque nulle (Wright, 1929). Au cours du XXème siècle, l'existence de modificateurs de dominance a largement été rejetée au profit la théorie enzymatique (Wright, 1929, 1934; Haldane, 1930; Kacser & Burns, 1981). Toutefois, des modèles récents (Otto & Bourguet, 1999; Nuismer & Otto, 2005; Llaurens *et al.*, 2009) ont montré que dans certaines conditions la sélection pouvait être substantielle, en particulier lorsqu'un mécanisme comme la sélection balancée maintient de façon active un ensemble de lignées alléliques sur des échelles de temps importantes, formant une proportion importante de génotypes hétérozygotes au sein des populations.

En accord avec ces modèles, la première mise en évidence expérimentale de l'existence de tels modificateurs de dominance a été publiée récemment (Tarutani *et al.*, 2010; Durand *et al.*, 2014), permettant de rouvrir en de nouveaux termes le débat sur cette question. Ces modificateurs contrôlent les relations de dominance/récessivité entre allèles au locus d'auto-incompatibilité chez les Brassicaceae et prennent la forme de petits ARNs non codants produits par les allèles dominants, et de leurs cibles respectives sur les allèles récessifs. Chez *Arabidopsis halleri*, les relations de dominance entre allèles sont contrôlées par un réseau de régulation impliquant au moins 8 familles de petits ARNs non codants (Durand *et al.*, 2014) et offrent une occasion unique d'examiner les contraintes fonctionnelles qui pèsent sur l'évolution de ces modificateurs de dominance. Cette question a un écho tout particulier dans

le cadre du débat d'origine sur l'intensité de la sélection sur les modificateurs de dominance proposés par Fisher (1928).

Au-delà de ces seuls éléments génétiques spécifiques du locus S, les patrons de polymorphisme des gènes producteurs de microARNs (miR) ont été étudiés à l'échelle génomique chez plusieurs espèces. Ces analyses de polymorphisme le long de ces précurseurs ont permis de mettre en évidence différents aspects des contraintes fonctionnelles qu'ils subissent. Tout d'abord, les locus producteurs de miR semblent accumuler sur l'ensemble de leur structure tige-boucle environ trois fois moins de mutations en comparaison avec les positions synonymes des régions codantes, comme montré par Jovelin & Cutter, (2014) chez *Caenorhabditis remanei*. Collectivement, le polymorphisme des locus producteurs de miR chez cette espèce est proche de celui des positions non-synonymes des gènes codant pour des protéines, indiquant une contrainte fonctionnelle importante (Jovelin & Cutter, 2014). De plus, les régions du miR mature et du miR star (miR complémentaire produit par le même précurseur et généralement présent sur la tige opposée du miR mature) accumulent généralement moins de polymorphisme que le reste de la structure tige boucle, ce qui semble indiquer que les contraintes sur ces régions sont plus fortes. Les analyses de polymorphisme des miRs et en particulier leur distribution le long des motifs tige-boucle ont donc permis de montrer une sélection purifiante intense contre les nouvelles mutations, probablement pour maintenir leurs fonctions régulatrices de l'expression des gènes. Globalement, évaluer les patrons de polymorphisme des miRs nous renseigne donc sur l'intensité de la contrainte fonctionnelle qu'ils subissent et ses éventuelles variations le long de leur séquence.

Sur la base de ces observations, on peut faire plusieurs prédictions quant au niveau de polymorphisme des miRs au locus d'auto-incompatibilité. D'une part, un des critères principaux déterminant la capacité d'un petit ARN à réprimer l'expression de l'allèle cible, est la complémentarité de séquence entre ces deux éléments. Nous avons vu que ce critère de complémentarité peut jouer un rôle critique, notamment quand le score d'alignement est proche du seuil de 18 (Chapitre 1). Ainsi, un simple SNP sur le miR mature peut avoir des conséquences fonctionnelles, avec l'apparition ou l'abolition du phénomène de silencing, et ces polymorphismes peuvent avoir un rôle fonctionnel essentiel comme révélé chez Brassica (Yasuda *et al.*, 2016). A ce jour, on ne sait pourtant rien des niveaux de polymorphisme des éléments de cette machinerie en populations naturelles. D'autre part, on s'attend à ce qu'ils

soient sous sélection purifiante et montrent donc un très faible niveau de polymorphisme, à la fois au sein de la structure tige boucle comparé au reste des régions codantes du locus S, mais aussi sur les séquences du miR mature et du miR star par rapport au reste du précurseur. Par ailleurs, l'annotation des miRs a révélé la présence sur plusieurs haplotypes S de motifs dont la séquence nucléotidique rappelle fortement celle des précurseurs décrits ci-dessus mais pour lesquels Durand *et al.* (2014) n'a pas pu détecter d'expression de miR mature et que nous avons donc appelés « silencieux ». A ce jour, nous ne savons pas si ces précurseurs sont effectivement des motifs silencieux qui ne s'expriment jamais, étant par exemple des reliquats d'anciens miRs ayant perdu leur fonction, ou alors des motifs ne s'exprimant que sous certaines conditions, ou ayant une expression très faible que nous n'avons pas été en mesure de détecter. On s'attend donc à ce que leur éventuel plus faible rôle fonctionnel se traduise par une accumulation plus importante de polymorphisme.

Dans l'objectif d'évaluer l'intensité de la contrainte fonctionnelle qui pèse sur les miRs du locus S, j'ai mis en œuvre dans le cadre du travail présenté dans ce chapitre une approche permettant d'accéder aux patrons de polymorphisme de ces éléments. Un aspect important de l'approche est qu'elle repose sur l'analyse du polymorphisme encore copies d'allèles S au sein des lignées alléliques (appelé par la suite « polymorphisme intra-allélique »). Etant donné le caractère sporophytique du système d'auto-incompatibilité chez les Brassicaceae, les relations de dominance entre allèles provoquent une asymétrie de la sélection. Il en résulte que les allèles récessifs devraient atteindre dans les populations une fréquence à l'équilibre plus élevée et présenter un temps de coalescence plus long que les allèles dominants (Schierup *et al.*, 1997; Billiard *et al.*, 2007; Castric *et al.*, 2010). On s'attend donc à ce que les niveaux de polymorphisme intra-allélique soit différents, avec des niveaux de polymorphisme intra-allélique plus fort pour les allèles récessifs que pour les dominants. Ces prédictions ont été étudiés dans (Castric *et al.*, 2010), mais sur un jeu de données limité de quatre allèles seulement. Si, en accord avec les prédictions, le niveau moyen de polymorphisme était globalement faible, le lien avec la dominance n'a pas pu être validé. Toutefois, ces résultats ont été obtenus sur des fragments partiels du seul gène *SRK*, sur seulement quatre lignées alléliques S et sur relativement peu d'échantillons. A ce jour, l'ampleur du polymorphisme intra-allélique et sa variation entre allèles le long de la hiérarchie de dominance reste donc largement inexplorés. En particulier, le polymorphisme des régions non-codantes (dont les

miRs contrôlant la dominance font partie) est totalement inconnu. Une étude récente chez *Arabidopsis thaliana* (Tsuchimatsu *et al.*, 2018) a révélé un polymorphisme intra-allélique important pour au moins deux des trois allèles S résiduels ségrégeant chez cette espèce, comportant en particulier de nombreuses variations structurales, mais il n'est pas clair si ces résultats sont transposables à une espèce qui n'a pas perdu son système d'auto-incompatibilité.

Dans ce chapitre, nous avons mesuré le polymorphisme des modificateurs de dominance en populations naturelles afin de déterminer si des variations de la machinerie de contrôle de la dominance étaient présentes, et de tenter d'identifier les contraintes fonctionnelles qui pèsent sur ces éléments. Pour cela, nous avons reséquencé le locus S de différentes lignées alléliques de différentes classes de dominance, via deux approches. Nous avons dans un premier temps exploré la possibilité d'une approche par simple séquençage SANGER, puis devant l'ampleur et la difficulté de la tâche nous avons ensuite opté pour une approche NGS par capture de séquences. Nous avons alors mesuré le polymorphisme intra-allélique pour différentes classes de dominance, incluant le polymorphisme des gènes codant pour le phénotype d'auto-incompatibilité (*SCR* et *SRK*) ainsi que les sRNAs et leurs cibles respectives impliqués dans le réseau de régulation de la dominance et le reste des régions intergéniques. Nos résultats confirment l'absence de corrélation entre dominance et polymorphisme intra-allélique, révèlent que l'intensité de la sélection exercée sur les modificateurs de dominance est du même ordre de grandeur que celle qui s'exerce sur les régions codante et montrent que l'approche par capture de séquence représente un outil efficace pour l'étude de régions sous sélection balancée. Nous discutons les limites de l'approche et développons les perspectives possibles à ce travail.

Matériel & Méthodes

Approche exploratoire par séquençage SANGER

Dans un premier temps, nous avons envisagé d'amplifier par PCR puis de reséquencer en SANGER les différents précurseurs miRs des allèles à partir de couples d'amorces définis sur les séquences des différents clones BACs que nous avions à notre disposition. Pour cette analyse exploratoire, nous avons choisi l'unique précurseur miR de Ah01 (miR4239) et défini des amorces de PCR de part et d'autre de la structure tige-boucle, pour une région à amplifier de 119pb. Cette approche représentait un défi en soi, car les amorces étaient situées en régions intergéniques de part et d'autre de la structure tige-boucle qui produit le miR, ce qui pouvait poser problème si ces régions étaient trop polymorphes. Etant donné le peu de diversité intra-allélique attendu (Castric *et al.*, 2010) nous pensions pouvoir trouver des régions conservées en sein des lignées alléliques à utiliser comme point d'ancrage pour des amorces spécifiques. Nous avons sélectionné 10 individus possédant l'allèle Ah01 (7 plantes de trois populations Allemandes et 3 plantes provenant de France) à partir d'un ensemble de plantes en serre ayant fait l'objet d'un génotypage basé sur des amorces spécifiques du gène *SRK.* Nous avons testé l'amplification du précurseur d'Ah01_miR4239 via des amorces spécifiques et nous avons séquencé les produits d'amplification obtenus (Tableau 1).

Oligo name	Sequence	Longueur (nt)
Ah01mir4239_806F	TGCCCATTCCCATTAAGTCG	20
Ah01mir4239_846F :	CACAACGTGGAGCAAAATTAACT	23
Ah01mir4239_1273R	GACTGCGAGTACTTACCAACAC	22
Ah01mir4239_1313R	TGGGACATTAACATCAGCATTGT	23

Tableau 1 : Séquence des amorces utilisées pour amplifier le miR4239 chez Ah01

Approche par capture de séquence

Cette méthode consiste à utiliser des sondes d'oligonucléotides liées à des billes magnétiques en solution. Ces sondes, par complémentarité de séquence, s'hybrident alors aux fragments d'ADN individuels ou de librairies génomiques poolés, que l'on peut alors séquencer via NGS.

Echantillonnage

A cette étape, nous avons étendu l'échantillonnage à un ensemble de plantes issues de deux campagne d'échantillonnage (Allemagne en 2014, puis Autriche et Slovénie en 2015) ou disponibles au sein d'une collection d'ADN issue de plusieurs populations réparties en Europe, représentant les trois groupes génétiques identifiés (Meyer *et al.*, 2010; Pauwels *et al.*, 2012; Šrámková-Fuxová *et al.*, 2017). Une partie de ces individus avait été génotypée par PCR allèle-spécifique avec des amorces ciblées sur certains des allèles du gène *SRK* (Llaurens *et al.*, 2008), dans l'intention d'assurer l'obtention d'un nombre minimal de copies des allèles les plus dominants (qui sont généralement les plus rares). Cependant, le génotypage avec ces amorces étant incomplet (seuls les allèles recherchés peuvent être identifiés), de nombreux autres allèles peuvent être présents. Par ailleurs, ce génotypage repose sur des patrons de présence/absence de bandes d'amplification PCR sur gel d'agarose (Llaurens *et al.*, 2008) et est dont susceptible à la présence de faux positifs dont la fréquence est inconnue.

Définition des sondes de capture

L'approche par capture de séquence nous a permis d'étendre les séquences analysées en capturant l'intégralité du locus S, soit les gènes *SCR* et *SRK* mais également les régions intergéniques qui comprennent en particulier tous les locus producteurs de miRs et leurs cibles pour l'ensemble des allèles pour lesquels nous disposions d'un clone BAC (36 allèles S). Nous avons ajouté à ce dispositif toutes les séquences de *SRK* complètes (*n*=21) ou partielles du domaine S de *SRK* (*n*=102, taille moyenne = 684pb, soit 56% du domaine S) de différentes espèces de Brassicaceae (*A. halleri, A. lyrata, A. thaliana, A. kamchatica, Capsella grandiflora, C. rubella, Brassica rapa, B. oleraceae*). Nous avons également ajouté d'autres régions qui seront utilisées pour le chapitre 3, à savoir les régions flanquantes au locus S et des régions de contrôle tirées aléatoirement dans le génome (Tableau 2, voir chapitre 3) et qui ne seront pas utilisées ici.

Nom	Description	Longueur (bp)
S-locus	Séquence de chaque locus S disponible sous forme de clone BAC (*A. halleri* et *A. lyrata)*	966 930
Régions flanquantes	Séquence de deux clones BAC couvrant chacun une région flanquante du locus S	2*100 000
Base de données *SRK*	Séquences complètes ou partielles des domaines S de *SRK* connues chez l'ensemble des Brassicaceae (129 allèles provenant de 8 espèces)	173 000
Régions de contrôles	Ensemble de régions de 25kb tirées aléatoirement dans le génome d'*A. halleri*, ayant la même densité en gènes que les région flanquantes (+-10%)	100*25 000

Tableau 2 : Résumé des différentes régions incluses dans la définition des sondes.

De façon à s'assurer de la spécificité des sondes définies, nous avons exclu les éléments transposables à partir d'une librairie d'éléments transposables d'*A. halleri* et *A. lyrata* (Sylvain Legrand et Thibault Caron, in prep), en excluant 100pb de part et d'autre de chaque séquence identifiée comme répétée.

Caractéristiques des sondes

Les sondes ont été synthétisées par le prestataire externe MyBaits (Ann Arbor, Michigan (USA), http://www.arborbiosci.com/products/targeted-sequencing-kits/). Nous avons défini des K-mers de 120pb avec un taux de recouvrement de 1.15, pour un total de 48 151 sondes. Leur définition et leur synthèse a été réalisée par Mybaits, et nous avons vérifié que l'ensemble des sites visés dans les régions d'intérêt telles que les éléments de la machinerie de la dominance ou les gènes du locus S, étaient bien couverts par au moins une sonde. Collectivement, nous souhaitions être en mesure de capturer un total de 4.1Mb. Après exclusion des sites de faible complexité et des sites pour lesquels des sondes ne pouvaient pas être définies, notre dispositif de capture de séquences couvre 3.8Mb, soit 1,48% des 256Mb du génome *A. halleri*, permettant une capacité de multiplexage importante.

Préparation des banques

Lors de cette étude, nous avons procédé à la capture de deux librairies poolées distinctes (Tableau 3). Dans un premier temps, nous nous sommes focalisés sur deux allèles, Ah01 (le plus récessif) et Ah20 (très dominant), sur un premier groupe de 16 individus avec comme objectif de tester le protocole, tout en limitant le niveau de multiplexage. Une fois les

paramètres de multiplexage validés, nous avons capturé les séquences de 56 individus supplémentaires lors d'une seconde expérience de capture. La première capture a été fragmentée grâce à l'utilisation d'un « kit nextera DNA library prep » (FC-121-1031), la deuxième via un « kit kapa » (KK8503* 07962355001), reposant toutes deux sur une fragmentation chimique par des enzymes de restriction. Les échantillons ont ensuite subi un traitement à la RNAse puis une purification sur colonne (kit NucleoSpin® Plant II, 740770) puis ont été dosés au Qubit. C'est à partir de ces séquences fragmentées que nous avons créé les banques NGS. Pour cela, les extrémités de ces séquences ont d'abord été réparées et « A-tailed » pour le processus de ligation avec des adaptateurs possédant les index de multiplexage en plus des amorces spécifiques, nécessaires pour l'étape d'amplification. Après amplification, les banques ont été nettoyées afin de ne garder que les fragments dont la taille est comprise entre 150 et 800pb grâce à des billes AMPures. Les banques ont ensuite été passées au BioAnalyseur via une puce DNA HS (1µl) et dosée au Qubit. Elles sont ensuite « poolées » de manière équimolaire pour la capture.

librarie	individus (n)	allèles (n)	kit de construction de banque	methode de séquençage	Nb de read obtenus
1	16	Ah01(n=12); Ah20(n=5); Ah13(n=3); Ah12(n=2); Ah10(n=1); Ah03 (n=1)	kit nextera DNA library prep	MiSeq	13 728 804
2	56	Ah01(n=19); Ah03(n=15); Ah12(n=8); Ah10(n=7); Ah13(n=5); Ah20(n=2)	kit kapa	HiSeq	397 456 418

Tableau 3 : tableau résumé des deux expériences de capture, les effectifs par allèles après génotypage NGS

Capture et séquençage

La capture elle-même a suivi le protocole fourni par MyBaits. Brièvement, elle a consisté à récupérer les séquences d'intérêt grâce aux sondes de capture. Lorsque ces sondes se sont hybridées sur les fragments d'ADN constituant les banques NGS, il est possible de les récupérer grâce à des billes marquées à la streptavidine. Une fois les banques liées aux billes, on enlève la solution restante, tandis que les billes sont gardées dans le tube grâce à un portoir

aimanté. Les billes subissent alors plusieurs lavages, puis les séquences capturées sont amplifiées et dosées au BioAnalyseur. Les molécules d'ADN ainsi produites sont ensuite prêtes pour le séquençage, qui a été réalisé lors de deux expériences de séquençage Illumina, la première sur MiSeq et la seconde sur HiSeq, et diffèrent entre autres par la taille des lectures obtenues (300 pb et 150 pb respectivement).

Spécificité et sensibilité

L'efficacité de la capture a été mesurée grâce à deux critères. Tout d'abord la spécificité, qui est le nombre de lectures qui s'alignent sur les séquences ciblées par les sondes par rapport au nombre de lectures total obtenu par individu. Ensuite la sensibilité, qui quantifie la proportion des régions ciblées qui sont couverte par au moins une couverture seuil, qui peut varier selon le type d'analyse que l'on souhaite réaliser (Figure12).

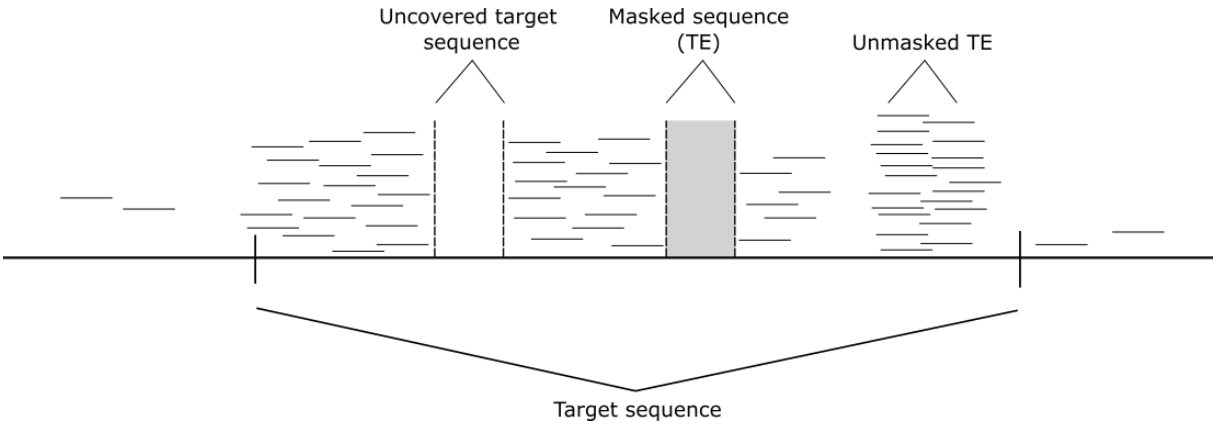


Figure 12 : Schéma présentant l'hétérogénéité possible dans la répartition des lectures pour une région cible. Cette figure schématique correspond à une vision théorique des différents cas de figure possible.

Nous avons également estimés le taux d'enrichissement des régions ciblées comme le ratio entre le nombre de lectures alignées sur le nombre total de lectures, divisé par le ratio de la taille des régions ciblées sur la taille du génome haploïde (cf formule ci-dessous).

$$EF = \frac{\left(\frac{number\ of\ reads\ that\ map\ on\ target}{total\ number\ of\ reads}\right)}{\left(\frac{target\ region\ size}{haploid\ genome\ size}\right)}$$

Génotypage des allèles *SRK* avec des données NGS

Nous avons dans un premier temps utilisé ces données de séquençage pour génotyper les individus via une nouvelle approche qui consiste à aligner avec BOWTIE2 (Langmead & Salzberg, 2012, version : 2.2.6) les reads sur la séquence du domaine S de chaque allèle du gène *SRK* que nous avions alors à disposition dans la base de données *SRK*¸ en ajoutant le gène paralogue *ARK3* (ou *Aly8*). Nous avons raisonné que si les allèles sont suffisamment différents entre eux pour que les reads s'alignent de manière spécifique, alors on devrait être capable de déterminer si un individu possède tel allèle ou non en comptant le nombre de read alignés et le taux d'erreur (mismatchs entre le read et la référence sur laquelle il est aligné), et en les comparant avec le nombre de reads alignés sur le gène paralogue (supposé être présent chez tous les individus). Ainsi, un individu de génotype S1S2 devrait au total montrer autant de reads alignés sur les références *SRK1* et *SRK2* que sur le paralogue *ARK3*, et des niveaux de couverture très faibles lorsque d'autres séquences *SRK* sont utilisées comme références. Cette approche a été développée au laboratoire par Mathieu Genete dans le contexte de l'analyse de données de séquençage génomique, et nous l'avons ici utilisée sans modification dans le cadre de données de capture de séquences.

Détection des variant nucléotidiques sur l'intégralité du locus S

Les séquences obtenues ont ensuite été alignées sur la totalité de la longueur des séquences de références (clones BACs) par BOWTIE2 (Langmead & Salzberg, 2012, version : 2.2.6, « end to end » par défaut), puis nettoyées grâce à l'utilisation de Mark Duplicate (Picard-tools, version 1.119) afin d'éliminer les duplicats de PCR et optiques. Les alignements ont ensuite été recoupés avec les séquences d'intérêt (bedtools, Quinlan & Hall, 2010, version2.27.0). La divergence des séquences des allèles S étant très importante, l'alignement des lectures de la séquence de référence d'un allèle à celle d'un autre est improbable, de telle sorte que les individus peuvent être considérés comme haploides pour chacun des deux allèles qu'ils portent pris successivement comme références. En conséquence, nous avons procédé à l'identification des variant nucléotidiques (SNPs) via Genome Analysis Toolkit (GATK, McKenna *et al.*, 2010, version 3.8.0) avec une approche individu par individu (GVCF), en supposant un génotype haploïde. Nous avons ensuite éliminé les variant détectés dont le seuil de qualité était inférieur à 60, retiré les indel, ainsi que les positions dont la couverture était en moyenne

inférieure à 10 lectures pour l'ensemble des individus confondus (VCFtools, Danecek *et al.*, 2011, version 0.1.15). Les données ont ensuite été visualisées grâce à IGV (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2013, version2.4.4), et les valeurs de polymorphisme nucléotidique ont été calculées suivant la formule : $H = 1 - \sum f^2$ (« f » étant la fréquence du mutant dans notre échantillon). La profondeur par position a été obtenue via SAMTools (Li *et al.*, 2009, version 1.3.1). Nous avons ensuite obtenu le cadre de lecture pour les gènes *SCR* et *SRK* grâce aux annotations des clones BACs de références (réalisées grâce à Fgenesh ; Solovyev *et al.*, 2006 ; selon Goubet *et al.*, 2012), afin de déterminer si les variant observés sont synonymes ou non-synonymes. La distinction entre les miRs matures et les miRs star a été obtenue en comparant le nombre de reads obtenus par RNAseq, le miR mature ayant été identifié comme celui présentant le plus grand nombre de reads. Le miR le plus abondant sur la tige opposée est appelé ici miR star. Les autres régions des tiges pour lesquelles la quantité de reads est inférieure au miR mature et star sont considérées comme des isomiRs (Figure 13).

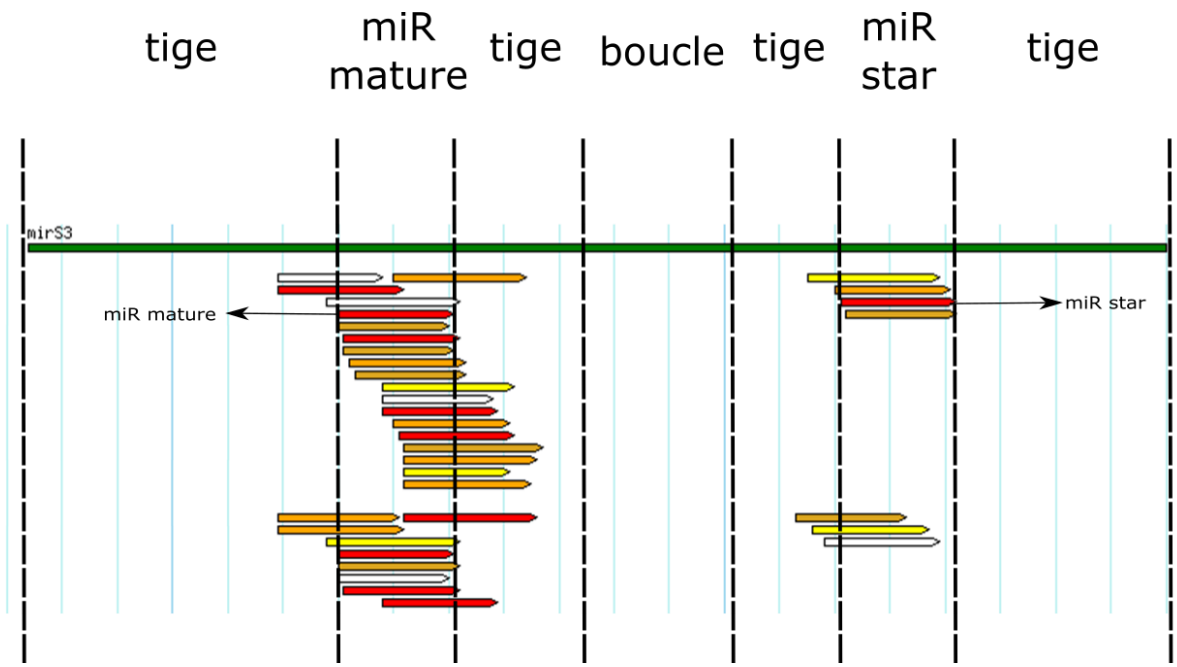


Figure 13 : Illustration de la méthode d'annotation d'un précurseur, ici MirS3. L'ensemble du précurseur est représenté en vert, sur laquelle s'alignent les différents reads détectés par RNAseq colorés en fonction de la quantité (blanc pour les plus rares, jusqu'à rouge pour les plus représentés). Le miR retrouvé en plus grande quantité est appelé par défaut miR mature, le miR présent en plus grande quantité sur la tige opposée est appelé miR star (tous deux indiqués par des flèches). L'annotation de la structure du précurseur dépend de la position du miR mature et du miR star. Les autres miRs sont appelés IsomiRs.

Au sein de la machinerie de contrôle de la dominance, nous avons distingué plusieurs sous-parties. D'une part les différents éléments de la structure tige/boucle du précurseur, à savoir les tiges, la boucle, le mir mature et le mir star, mais aussi les mirs silencieux, les cibles, ainsi que les cibles dites inactives. L'annotation des cibles a été faite à partir des données du chapitre 1 en prenant, pour chaque paire d'allèles, la meilleure cible du premier allèle sur le deuxième et inversement, sans tenir compte des relations de dominance. Certaines cibles (que appelons par la suite cibles « inactives ») n'en sont donc pas de véritables, mais sont des cibles potentielles, qui sous l'effet d'une substitution pourraient passer au-dessus du seuil de score d'alignement (de 18) identifié dans le chapitre 1.

Test par bootstrap

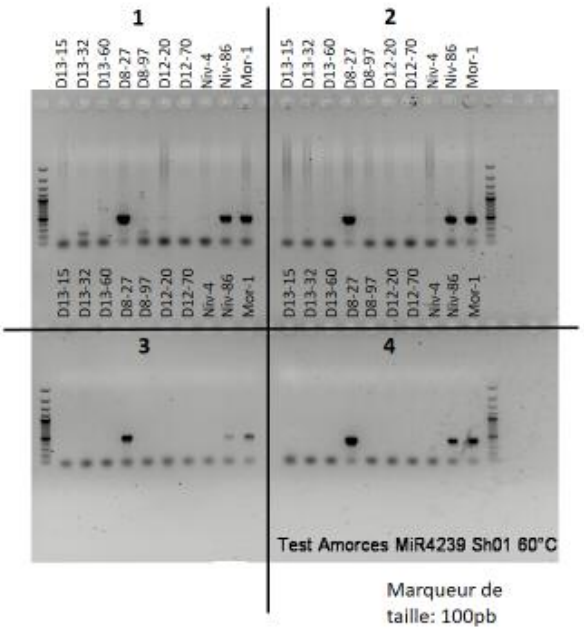
Afin de tester la significativité des différences de polymorphisme observées entre catégories de sites au sein du locus S (sites totaux, synonymes, non synonymes, intergéniques, précurseur de miRNA, miRNA mature, cibles de sRNAs), une procédure de bootstrap a été mise en œuvre en re-échantillonnant les sites de manière aléatoire dans des séquences de taille identique à la séquence étudiée, afin d'en calculer une valeur de polymorphisme, et ce 1 000 fois (10 000 pour les régions intergéniques). Ces valeurs de polymorphisme nous permettent ensuite de définir un intervalle de confiance à 95% de la distribution des valeurs simulées afin de déterminer si les éléments comparés sont significativement différents. Nous avons alors comparé le polymorphisme moyen observé au locus S sur les gènes *vs.* les régions intergéniques *vs.* les précurseurs. Nous avons ensuite comparé les niveaux de polymorphisme des gènes (π_N et π_S), des précurseurs ainsi que des régions intergéniques allèle par allèle, selon leur position le long de la hiérarchie de dominance. Enfin, pour les précurseurs, nous avons comparé les différentes parties entre elles (tiges, boucles, miRs, miRs star, miRs silencieux, cibles et cibles inactives) sur l'ensemble des haplotypes, sans considérer les allèles séparément.

Résultats

Approche en SANGER

Amplification du Ah01_mir4239

L'objectif de cette première approche était de savoir si, compte tenu du faible niveau de polymorphisme intra-allélique attendu, un reséquençage via des amorces dans les régions intergéniques autour des précurseurs de miR était envisageable afin d'en mesurer le polymorphisme. Pour cela nous avons testé simultanément plusieurs couples d'amorces, sur dix individus génotypés comme portant l'allèle S01, en suivant un protocole classique de PCR (Figure 14).



Sh01

MiR4239:

F1→Ah01mir4239_806F
R1→Ah01mir4239_1273R
F2→Ah01mir4239_846F
R2→Ah01mir4239_1313R

N°	Couple	Taille attendue
1	F1/R1	467
2	F2/R2	467
3	F1/R2	507
4	F2/R1	427

T°	Temps	
95	5 min	
95	40 sec	x35
60	45 sec	
72	1 min	
72	10 min	
16	∞	

Figure 14 : Résultat d'amplification sur quatre combinaisons de couples d'amorces Ah01_Mir4239, les différents couples sont numérotés de 1 à 4 puis représentés sur leurs parties respectives d'un gel d'agarose à 1%.

Aucun de ces quatre couples d'amorces n'était capable d'amplifier l'allèle S01 chez l'ensemble des individus, seuls les trois mêmes individus (D8-27 ; Niv-86 et Mor-1) étant à chaque fois correctement amplifiés. Ces résultats suggèrent la présence de plusieurs haplotypes, posant

un problème quand à notre capacité à correctement procéder à la mise au point des conditions d'amplification de nos amorces spécifiques. Cette approche semble peu compatible avec l'analyse du polymorphisme que nous souhaitons réaliser sur de multiples individus, de multiples précurseurs et de multiples allèles, et nous nous sommes donc tournés vers une approche NGS.

Création des banques NGS et capture

Profil taille des fragments des banques

La construction des banques a été réalisée avec succès pour l'ensemble des 76 échantillons. Dans l'ensemble, les profils de banques sont bons, avec toutefois trois individus qui présentaient deux pics (à 350 et 1000pb) au lieu d'un seul à 300-400pb (Figure 15). Ces trois échantillons ont été inclus dans l'expérience de séquençage (voir ci-dessous), et ne montrent pas de résultats aberrants lors des contrôles de qualités, seules les quantités de reads après filtrage par individus sont plus faibles que la moyenne.

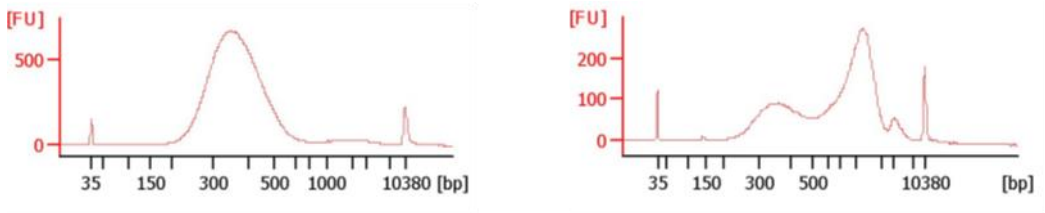


Figure 15 : Exemples de profils de banque NGS. A gauche, échantillon présentant un profil typique observé chez la majorité des échantillons, avec un pic à 300-400pb. A droite, un profil anormal, avec deux pics, l'un à 300-400pb, l'autre vers les 1000pb.

Résultat de la capture

Nous avons obtenu un total global sur les deux expériences de séquençage de 411 185 222 reads. Nous avons d'abord vérifié si le fait de multiplexer les individus ne provoquait pas de déséquilibre dans les quantités de lectures obtenues par individus, se traduisant par une sur- ou une sous-représentation de certains individus. Même si les quantités de lectures varient fortement entre la première et la seconde série de captures, il n'y a pas de déséquilibre important entre individus. Pour la 1ère série, le nombre de reads varie entre individus d'un facteur de 32%, soit un écart type de 281 649 reads pour une moyenne de 858 050 reads. Pour

la 2ème série, le nombre de reads varie d'un facteur de 35%, soir un écart type de 2 530 569 reads pour une moyenne de 7 097 436 reads (Tableau S1).

La spécificité est le nombre de lectures qui s'alignent correctement (« properly paired ») sur les régions que l'on voulait capturer, par rapport au nombre de reads total. Elle est en moyenne de 50%, avec un minimum à 14% (individu A26-D05) et un maximum à 59% (individu b8h3, Figure S1). La spécificité varie peu entre la 1ère et la 2e série de captures (spécificité moyenne de 41% et 53% pour la 1ère et la 2e série respectivement, Tableau S1). Ces valeurs sont satisfaisante et restes comparables avec celles que l'on peut trouver sur des différents modèles (environ 19% pour des amphibiens à génome complexe (McCartney-Melstad *et al.*, 2016) et entre 80.1% et 33.8% en moyenne pour une expérience de capture d'exons comparative entre espèces modèles et divergentes (Portik *et al.*, 2016).

La sensibilité de notre expérience est élevée, 99% des régions contrôle et des régions flanquantes du locus S étant couvertes par au moins 1 read, et 90% d'entre elles étant couvertes à X=20, permettant une étape de découverte des SNPs relativement puissante. Concernant le locus S, le calcul de la sensibilité est à appréhender de manière différente, dans la mesure où il faut la calculer allèle par allèle. En effet, nous ne pouvions pas obtenir un alignement total sur les différents allèles du locus S (chaque individu n'en présentant que deux au maximum), et avons donc traité chaque allèle séparément. Pour le locus S, en moyenne, 93% des positions sont couverte pas au moins 1 read (77% et 67% des positions sont couverte par au moins 10 et 20 reads respectivement). Mais les résultats de couverture varient entre les deux séries de capture avec des couvertures à 10 et 20X beaucoup plus faible pour la 1ère série (Tableau 4). Le taux d'enrichissement calculé comme étant le ratio entre la spécificité et la taille des régions ciblées, est en moyenne de 25,7. C'est-à-dire que l'on a séquencé en moyenne 25.7 fois plus les régions ciblées avec la méthode de capture par rapport à un séquençage de ces mêmes régions sans les cibler spécifiquement. Ces résultats sont satisfaisants et nous placent dans des conditions adéquates pour l'analyse du polymorphisme des différents compartiments du génome que nous voulions capturer. Les résultats obtenus sur les trois individus qui présentaient un profil de banque particuliers indiquent une quantité totale de séquences plutôt faibles (en moyenne 2 728 929 contre 5 840 557 pour le reste des individus, Figure S1), mais leur spécificité est comparable aux autres (en moyenne 0.65 vs 0.68, Tableau S1). La sensibilité de ces individus à X=1 est en moyenne de 0.93, contre 0.96 pour les

individus de la même série de capture, ce qui n'indique pas de problème particulier et nous les avons inclus dans la suite des analyses.

Serie	individus	allele	Coverage		
			X=1	X=10	X=20
1	PL22-6	Ah03	0.892	0.466	0.268
	D12-3	Ah01	0.810	0.365	0.133
	A03-F01	Ah01	0.886	0.503	0.273
	PL22-6	Ah01	0.940	0.587	0.332
	D8-18	Ah01	0.811	0.432	0.231
	NIV99	Ah01	0.862	0.384	0.138
	A06-C12	Ah01	0.941	0.722	0.524
	Niv6	Ah01	0.958	0.766	0.456
	A03-H09	Ah01	0.784	0.457	0.268
	A02-F10	Ah01	0.878	0.649	0.407
	A02-D09	Ah01	0.911	0.553	0.306
	aha2-18-2-8	Ah01	0.934	0.711	0.576
	PL22-11	Ah01	0.832	0.250	0.070
	A26-D05	Ah10	0.682	0.167	0.079
	D12-3	Ah12	0.819	0.244	0.035
	Niv6	Ah12	0.890	0.524	0.250
	A06-C12	Ah13	0.884	0.654	0.478
	A03-H09	Ah13	0.900	0.535	0.293
	PL22-11	Ah13	0.850	0.539	0.347
	A26-D05	Ah20	0.696	0.152	0.072
	A22-C12	Ah20	0.905	0.493	0.198
	Niv18	Ah20	0.952	0.745	0.471
	A02-F10	Ah20	0.853	0.623	0.360
	A02-D09	Ah20	0.880	0.601	0.293
		mean	0.865	0.505	0.286
		std.dev	0.07162761	0.175488118	0.150893066

Tableau 4 : Sensibilité pour chaque individu à X=1, 10 et 20.

Serie	individus	allele	Coverage		
			X=1	X=10	X=20
	a24-c01	Ah03	0.995	0.951	0.923
	a28-a01	Ah03	0.978	0.932	0.887
	a28-a06	Ah03	0.990	0.942	0.916
	a22-g06	Ah03	0.993	0.938	0.923
	a06-h11	Ah03	0.981	0.921	0.901
	a01-d06	Ah03	0.982	0.932	0.891
	a20-b06	Ah03	0.990	0.950	0.921
	b8h3	Ah03	0.969	0.920	0.900
	b4b11	Ah03	0.992	0.931	0.885
	a06-h05	Ah03	0.972	0.907	0.877
	a22-b03	Ah03	0.981	0.930	0.887
	a03-a09	Ah03	0.983	0.943	0.922
	a24-g09	Ah03	0.985	0.942	0.924
	a22-a11	Ah03	0.974	0.937	0.913
	a06-g05	Ah03	0.949	0.884	0.822
	a20-b04	Ah01	0.955	0.842	0.742
	a02-h03	Ah01	0.996	0.955	0.908
	b6a9	Ah01	0.987	0.942	0.902
	b4b10	Ah01	0.979	0.928	0.871
	a03-f03	Ah01	0.895	0.830	0.801
	a02-g02	Ah01	0.979	0.846	0.789
	a02-d02	Ah01	0.993	0.974	0.926
	a03-d05	Ah01	0.988	0.970	0.939
	a06-h11	Ah01	0.993	0.952	0.892
	b4g10	Ah01	0.989	0.857	0.708
	a02-a11	Ah01	0.987	0.927	0.872
	a03-f10	Ah01	0.988	0.941	0.898
2	b4b11	Ah01	0.984	0.907	0.852
	a02-b04	Ah01	0.946	0.873	0.820
	b8c2	Ah01	0.980	0.927	0.896
	a06-d12	Ah01	0.999	0.990	0.981
	a06-h09	Ah01	0.883	0.828	0.801
	b4-d10	Ah01	0.979	0.895	0.834
	a06-h05	Ah01	0.984	0.938	0.907
	b6a9	Ah10	0.956	0.876	0.825
	a06-e02	Ah10	0.961	0.890	0.819
	a21-a03	Ah10	0.875	0.721	0.624
	b1b9	Ah10	0.967	0.881	0.843
	b6b9	Ah10	0.950	0.842	0.747
	b2g06	Ah10	0.961	0.871	0.805
	b2e10	Ah10	0.954	0.874	0.816
	a28-a01	Ah12	0.954	0.875	0.827
	b4b10	Ah12	0.975	0.885	0.844
	b2e10	Ah12	0.963	0.882	0.829
	b8e6	Ah12	0.752	0.531	0.458
	b4-d10	Ah12	0.956	0.875	0.805
	b1c10	Ah12	0.972	0.890	0.834
	a22-a11	Ah12	0.959	0.890	0.837
	a06-g05	Ah12	0.901	0.745	0.637
	a06-g12	Ah13	0.978	0.873	0.823
	a03-f10	Ah13	0.959	0.807	0.749
	a26-d09	Ah13	0.967	0.857	0.786
	a02-c08	Ah13	0.929	0.786	0.735
	a01-g06	Ah13	0.960	0.821	0.753
	a20-g05	Ah20	0.949	0.881	0.830
	b2g06	Ah20	0.966	0.897	0.871
		mean	0.964	0.888	0.838
		std.dev	0.0394884	0.072827032	0.088949833

Tableau 4(suite)

Effectif par allèle après génotypage

Nous avons obtenus un total de 31, 16, 10, 8, 8 et 7 copies pour les allèles Ah01, Ah03, Ah12, Ah10, Ah13 et Ah20 respectivement, plus au moins 24 copies d'allèles présents en un plus faible nombre de copies (Figure 16, Tableau S2). Nous avons décidé de ne garder pour la suite des analyses que les allèles dont on en possède 6 copies au minimum, c'est à dire les six allèles listés ci-dessus, représentant un total de 80 copies d'allèles S différentes. Sur les 76 individus, nous avons retrouvé les deux allèles chez 56 d'entre eux, tandis que seul un des deux allèles a été retrouvé pour 12 échantillons. Parmi ces 12 échantillons, 4 possèdent Ah01 (l'allèle le plus récessif de la série allélique) comme unique allèle, suggérant qu'ils puissent être des homozygotes. Quatre échantillons présentent une incertitude sur le génotypage (avec 3 allèles distincts pour lesquels les couvertures sont comparables) et ont été exclus des analyses suivantes (Figure 16, Tableau S2). Il est à noter que le choix des individus s'étant basé sur une connaissance *a priori* des génotypes S des individus, ces données ne peuvent pas être utilisées pour estimer leurs fréquences relatives en populations naturelles. Par ailleurs, nous notons que la proportion des individus pour lesquels le génotype déterminé par génotypage PCR diffère de celui obtenu par génotypage NGS est plus élevée que celle à laquelle nous nous attendions (33%), suggérant une forte incertitude sur la première dont la lecture des bandes d'amplification est parfois difficile. Dans le cadre de cette thèse je n'ai pas pu déterminer précisément l'origine de cette différence importante, mais le séquençage NGS réalisé révélant de nombreux reads à forte identité avec les allèles *SRK* utilisés comme référence, nous avons choisi de considérer le génotypage NGS pour la suite de cette étude.

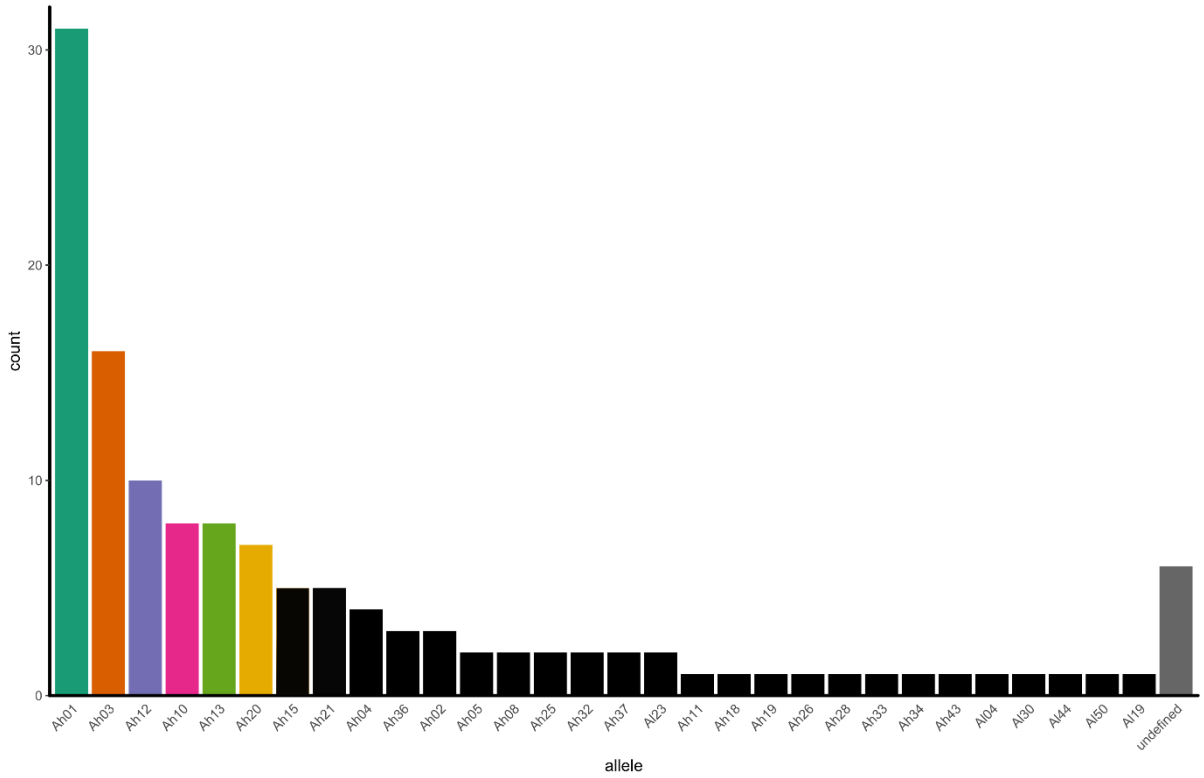


Figure 16 : Effectifs de chacun des allèles S dans l'échantillon de 72 individus d'*A. halleri*. Les allèles gardés pour analyse sont colorés (Ah01 en Jade ; Ah03 en Orange ; Ah12 en Mauve ; Ah10 en Fuchsia ; Ah13 en Vert Clair et Ah20 en Jaune), les allèles en noirs n'ont pas été retenus pour la suite car présents en trop faible nombre de copies dans l'échantillon. La barre grise représente les allèles non identifiés (pouvant correspondre à des individus homozygotes pour l'un des allèles identifiés). Noter que l'échantillonnage ayant été biaisé vers l'obtention d'individus portant des allèles spécifiques, cette distribution ne peut pas être utilisée pour évaluer la fréquence des allèles S dans les populations naturelles échantillonnées.

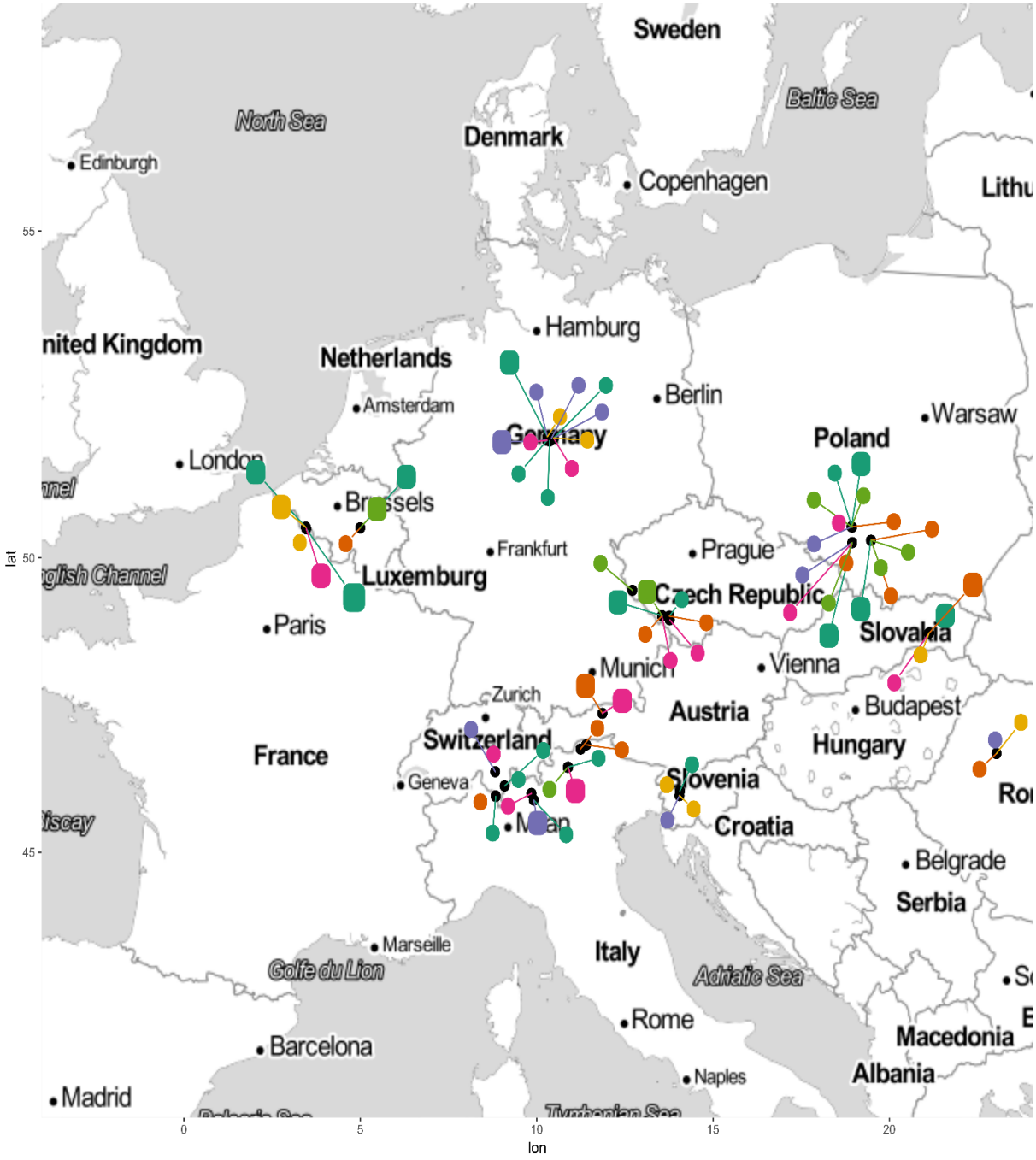


Figure 17 : Répartition des copies d'allèles par population, montrant leur large répartition géographique. Les couleurs représentent les différents allèles en accord avec la Figure 6. La taille des points représente le nombre de copies de chaque allèle. Pour éviter la superposition des points, ceux-ci sont déportés et reliés à leur point d'origine (en noir) par un court segment.

Polymorphisme du locus S

Nombre et localisation des SNPs

Au total, pour l'ensemble des 80 copies des allèles d'intérêt au locus S que nous avons pu étudier, nous avons pu analyser 111 455 pb, dont 9 345 pour les séquences codantes des gènes *SCR* et *SRK* ; 3 298pb pour les différents précurseurs des miRNAs et 98 815 pb de régions intergéniques. Sur l'ensemble, nous avons pu identifier 1 765 SNPs, dont 55 dans les séquences codantes des gènes *SCR* et *SRK* ; 33 sur l'ensemble de la machinerie de régulation de la dominance (précurseurs et cibles), et 1 677 pour les régions intergéniques de l'ensemble des allèles. Un exemple de couverture est représenté en Figure 18.

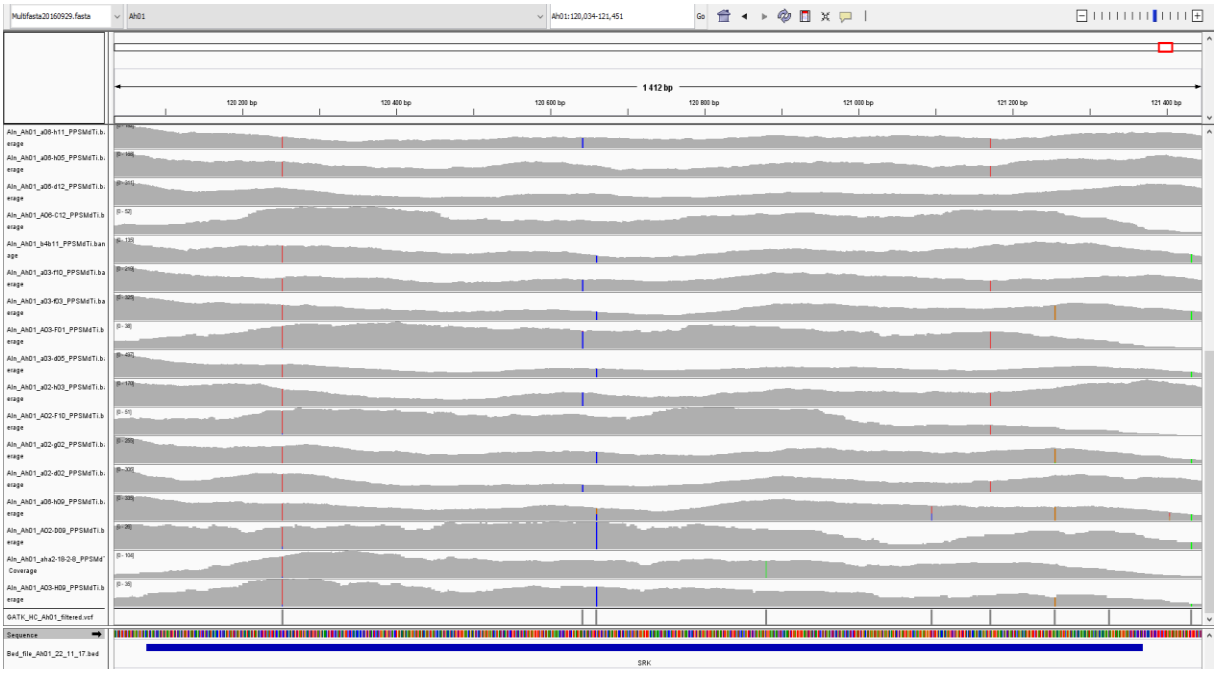


Figure 18 : Exemple de patrons de couverture et de position des SNPs présents sur le domaine S du gène *SRK* pour l'allèle Ah01. Les traits verticaux de couleur indiquent la présence de SNPs dans la séquence des individus représentés en lignes. Les barres grises correspondent à la couverture de chaque position nucléotidique successive de la séquence de référence (la position du domaine S de *SRK* est représentée par une barre horizontale bleue) Les traits verticaux gris en bas indiquent la position des SNPs retenus après le filtre de qualité par position (ici, 8 SNPs dans la séquence de SRK).

Absence de polymorphismes structuraux détectés par analyse de la couverture

A l'inverse de Tsuchimatsu et al. (2018), nous n'observons pas de grandes délétions sur des portions du locus S, qui se seraient traduites par des régions sans couverture. Toutefois, nous observons quelques petites délétions dans les régions intergéniques (et notamment une délétion plus grande d'environ 2kb autour du miR4239 d'Ah01 chez certains individus dont nous parlerons plus tard), voire dans certains cas sur les régions codantes. Un individu provenant d'une population en Slovaquie possédant l'allèle Ah20 présente une délétion de 24 pb au début du domaine S de *SRK* (Figure S2), ce qui a de forte chance de créer une protéine tronquée dont la fonction de reconnaissance du pistil pour l'allèle Ah20 n'est plus fonctionnelle. L'allèle Ah20 étant parmi les plus dominants (Durand *et al.*, 2014), il est probable que les individus porteurs de cet allèle soient auto-compatibles. Sur une autre plante, de génotype S10S20 (individu A26-D05), nous n'observons de couverture ni sur une portion du 2^{e} exon de *SCR* ni sur une partie du domaine S de *SRK* pour chacun des deux allèles Ah10 et Ah20 là où le reste des séquences codantes ou des régions intergéniques présente des mêmes niveaux de couverture comparables à celles des autres individus (Figure S3-S6). Toutefois cet individu est aussi celui ayant le plus faible niveau de spécificité et une sensibilité faible comparée aux autres échantillons. Le séquençage de ces séquences par une approche alternative (*e.g.* Sanger) sera maintenant nécessaire pour confirmer la présence de ces haplotypes non-fonctionnels. Il s'agirait de la première mise en évidence de la ségrégation d'haplotypes S non fonctionnels dans des populations naturelles d'*A. halleri*.

Analyse comparée du polymorphisme

Parmi les régions codantes de *SCR* et *SRK*, nous observons un π synonyme (π_S) moyen de 0,003, et un π non-synonyme (π_N) moyen de 0,001 (Figure 19). Ces valeurs varient d'un allèle à l'autre, de 1 seul site ségrégant pour l'allèle Ah13 à 28 sites ségrégant pour l'allèle Ah12, formant de 2 à 7 haplotypes distincts respectivement pour Ah13 et Ah01. Tous les sites variant détectés chez Castric *et al.*, (2010) pour AhSRK01 ont été retrouvés lors de notre analyse excepté une position non-synonyme, et nous avons identifié un site non-synonyme supplémentaire. Les valeurs de π_S moyennes sont très similaires à celles de cette étude.

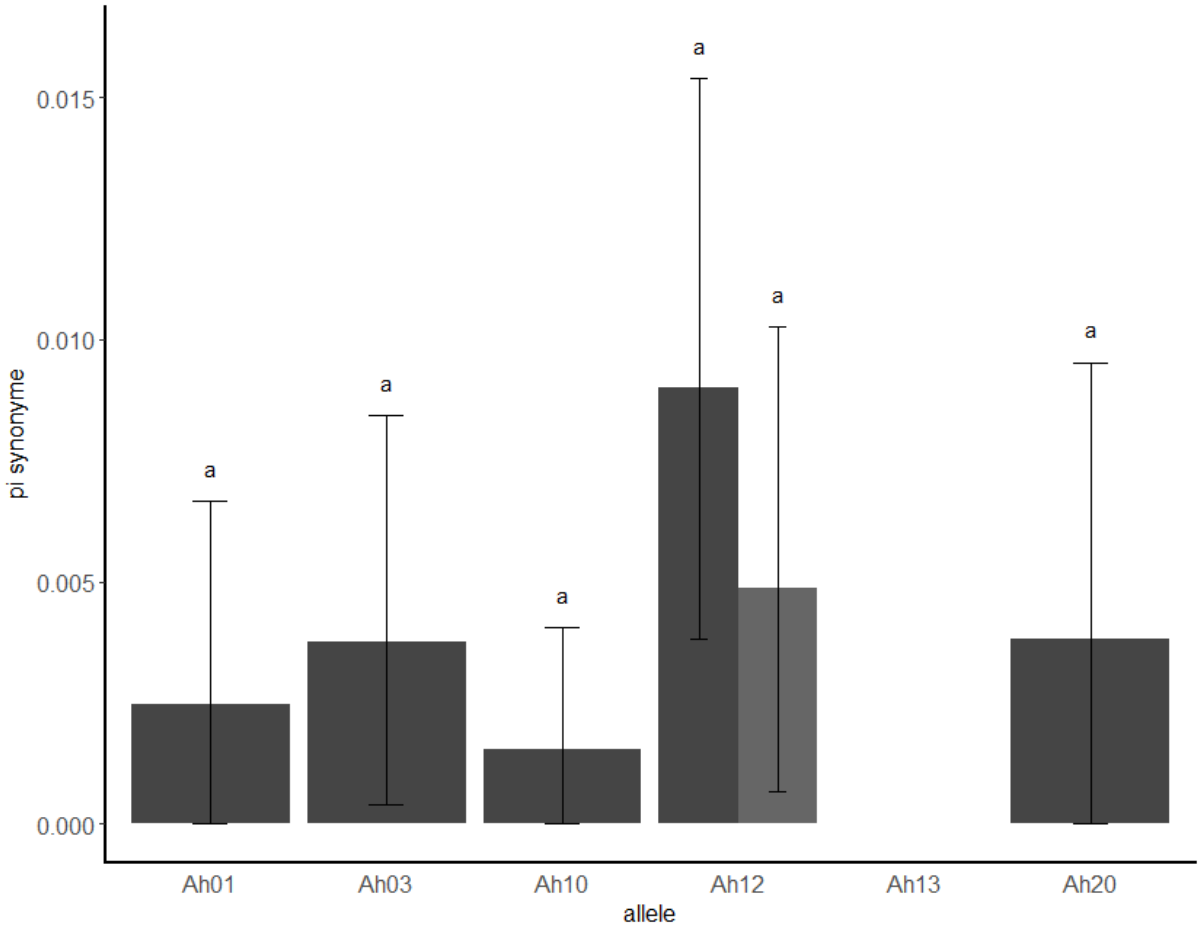


Figure 19 : Valeurs des π synonyme (Gris) au sein des différentes lignées alléliques (sans distinction entre *SCR et SRK*). La valeur en gris claire pour Ah12 représente la valeur de π_S si l'on retire l'haplotype divergent (vraisemblablement issu d'introgression). Les allèles sont triés de gauche à droite, du plus récessif au plus dominant.

L'allèle Ah12 présente un haplotype (individu b8e6, issu d'une population Suisse) très différent des autres, tant au niveau du polymorphisme de *SRK* que des régions intergéniques (Figure S7), résultant en un niveau de polymorphisme non-synonyme relativement élevé par rapport aux autres allèles. L'alignement des séquences du domaine S de *SRK* des individus possédant l'allèle Ah12 avec la séquence partielle (534pb) du domaine S de *SRK* dont on dispose pour l'orthologue *A. lyrata* Al42 nous permet de constater que sur les douze SNPs présents sur ces séquences, sept sont spécifiquement partagés entre l'orthologue *A. lyrata* et l'individu b8e6, là où les huit autres individus sont similaire à la séquence de référence AhSRK12 (Tableau 5).

	Position	0715	0746	0824	0834	0909	0913	1046	1065	1075	1161	1227	1233
	Reference	A	C	C	C	A	G	G	A	C	C	T	G
A.lyrata	AlSRK42	**C**	.	.	**G**	G	.	**A**	**G**	.	**T**	**G**	**T**
A.halleri	b8e6	**C**	A	T	**G**	G	A	**A**	**G**	A	**T**	**G**	**T**
	a06-g05	.	.	.	.	G	.	.	.	.	.	.	.
	a22-a11	.	.	.	.	G	.	.	.	.	.	.	.
	b1-c10	.	.	.	.	G	.	.	.	.	.	.	.
	b2-e10	.	.	.	.	G	.	.	.	.	.	.	.
	d12-3	.	.	.	.	G	.	.	.	.	.	.	.
	b4-b10	.	.	.	.	.	.	.	.	.	.	.	.
	b4-d10	.	.	.	.	.	.	.	.	.	.	.	.
	niv6	.	.	.	.	.	.	.	.	.	.	.	.

Tableau 5 : Alignement des séquences du domaine-S de *SRK* avec une séquence partielle de l'orthologue *A. lyrata* (AlSRK42). Les nucléotides en gras sont spécifiquement partagés entre l'orthologue *A. lyrata* et l'individu b8e6. Les positions sont données en référence au codon start du domaine S de la séquence de référence.

Ces résultats semblent indiquer que cet haplotype très différent est issu d'un phénomène d'introgression de l'allèle AlSRK42. Si on le retire dans le calcul du polymorphisme, le π_S pour cet allèle passe alors à 0.004, du même ordre de grandeur que celui des autres allèles (Figure 19).

Pour la première fois, nous avons été en mesure de mesurer le polymorphisme des régions non-codantes pour ces différents allèles, ce qui représente l'avantage de constituer un nombre de sites bien supérieur aux seules régions codantes. En moyenne, le polymorphisme calculé tous allèles confondus est de 0.0043. Si les allèles Ah01 et Ah03 possèdent des valeurs élevées (π=0.005 et 0.006 respectivement), on constate dans un premier temps qu'Ah01 n'est pas le plus polymorphe. En effet, c'est l'allèle Ah13, très dominant, avec une valeur de π=0.007 qui est le plus polymorphe dans notre échantillonnage. Dans un second temps, ces trois allèles (Ah01, Ah03 et Ah13) forment ensemble un groupe d'allèles très polymorphes, tandis que les allèles Ah10, Ah12 (introgression incluse) et Ah20 semblent accumuler moins de substitution avec des valeurs de π de 0.002, 0.001 et 0.003 respectivement. De fait, il est intéressant de constater que les allèles Ah13 et Ah20, les plus dominants, sont très différents en terme de polymorphisme des régions non-codantes (Figure 20).

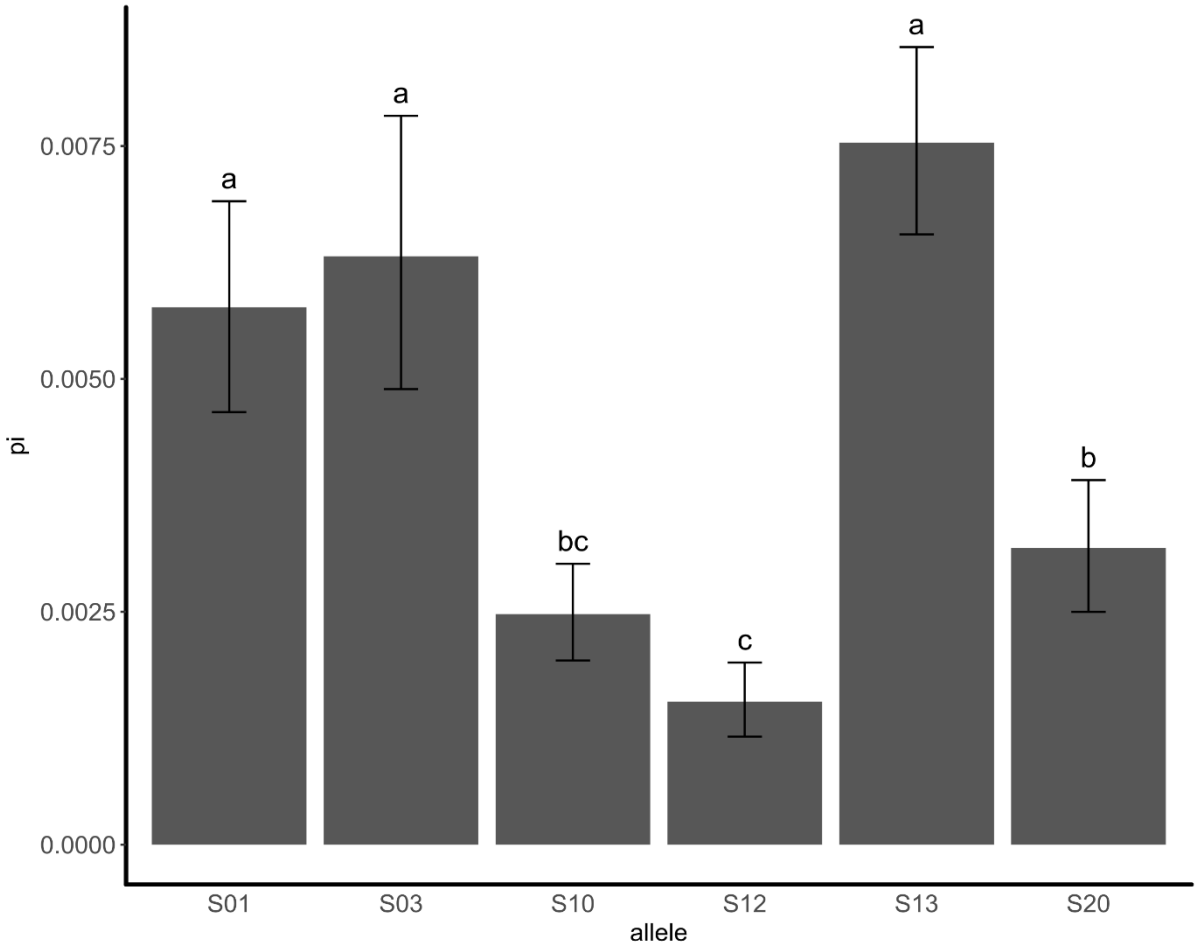


Figure 20 : Distribution des valeurs de polymorphisme mesurées sur les régions intergéniques des différents allèles S. Les barres d'erreur représentent les intervalles de confiance à 95% obtenus par bootstrap sur les sites. Les allèles sont triés de gauche à droite du plus récessif au plus dominant.

Globalement, et contrairement aux attendus théoriques (Castric *et al.*, 2010), nous ne trouvons donc pas de corrélation entre la position dans la hiérarchie de dominance et le niveau de polymorphisme intra-allélique, que ce soit pour les régions codantes ou pour les régions intergéniques.

Les valeurs ratios π_N/π_S vont de 0.7 pour l'allèle Ah01 à 0.11 pour l'allèle Ah03 (Figure 21).Figure 21 : Valeur des ratios π_N / π_S des allèles, *SCR* et *SRK* confondu. Les ratios sont représentés en noir. La valeur en gris représente le ratio de l'allèle Ah12 sans l'haplotype divergent. Les allèles sont classés de gauche à droite par position dans la hiérarchie de dominance. La valeur de π_N / π_S n'est pas définie pouf Ah13 car π_S est nul. De façon intéressante, la valeur de π_N/π_S est maximale pour l'allèle le plus récessif Ah01, qui est pourtant l'allèle dont les fréquences en populations naturelles sont les plus élevées et au sein duquel la sélection purifiante devrait donc être la plus efficace.

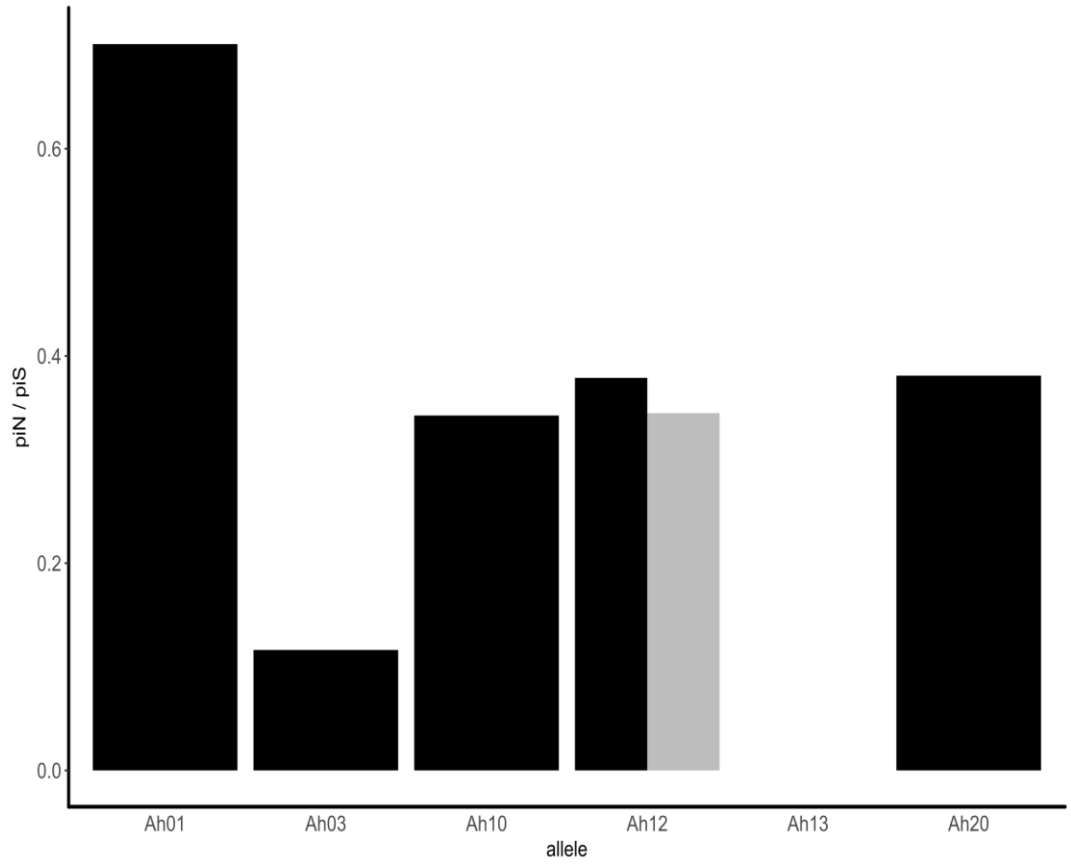


Figure 21 : Valeur des ratios π_N / π_S des allèles, *SCR* et *SRK* confondu. Les ratios sont représentés en noir. La valeur en gris représente le ratio de l'allèle Ah12 sans l'haplotype divergent. Les allèles sont classés de gauche à droite par position dans la hiérarchie de dominance. La valeur de π_N / π_S n'est pas définie pouf Ah13 car π_S est nul.

Contraintes fonctionnelles sur la machinerie de régulation de la dominance

Nous trouvons un total de 33 sites polymorphes sur l'ensemble des 3 241 nt des éléments de la machinerie de régulation. Parmi ces 33 sites, 16 sont situés au sein des différentes partie des précurseurs (2 sur les boucles, 7 sur les tiges, 5 sur les mirs et 2 sur les mirs*). Le reste des sites étant retrouvés sur les mirs silencieux, les cibles ainsi que les cibles dites inactives (3, 1 et 13 respectivement). Sur l'ensemble de ces éléments, nous obtenons un π moyen de 0.002, ce qui est significativement moins élevé que pour les régions non-codantes du locus S et du même ordre que pour les régions codantes du locus S (Figure 22).

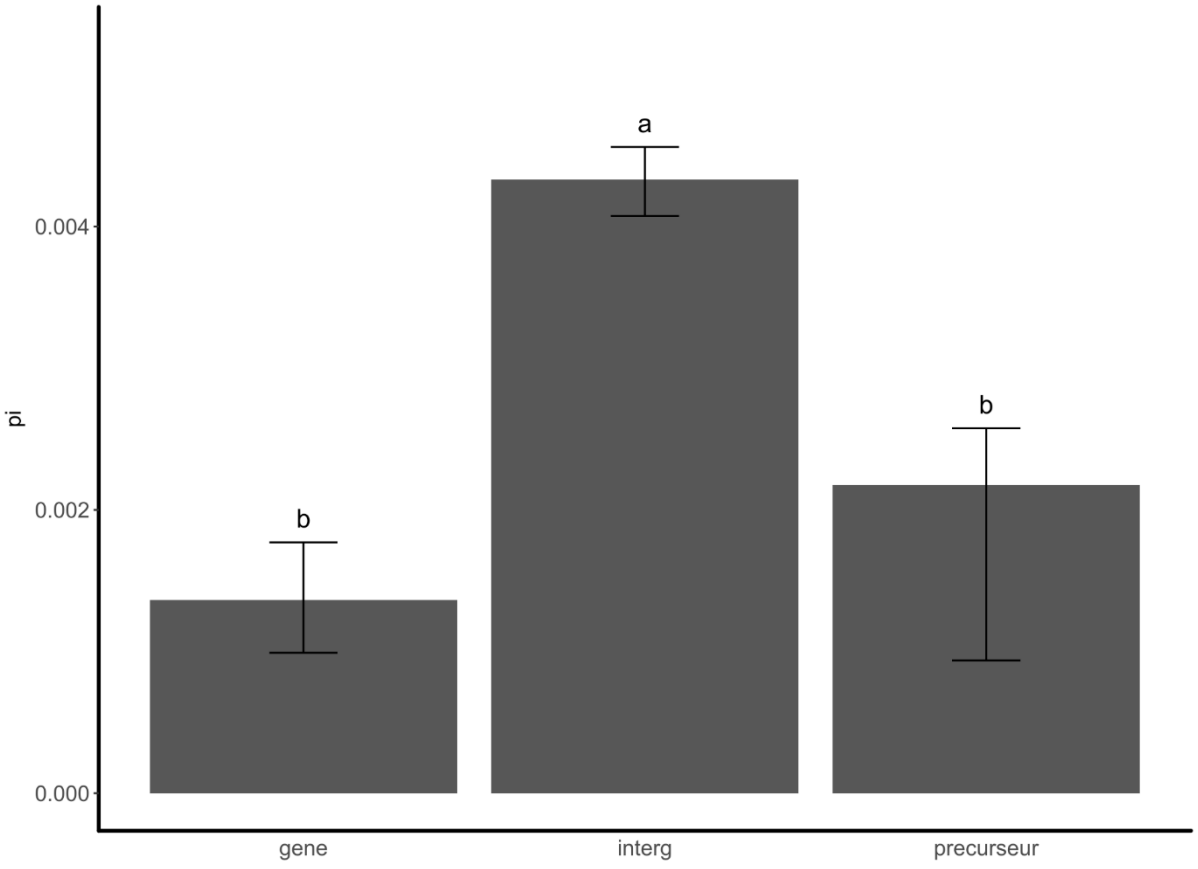


Figure 22 : Valeur de polymorphisme mesuré en moyenne sur tous les allèles pour les gènes (*SRK* et *SCR*), les régions intergéniques, les précurseurs. Les barres d'erreur représentent les intervalles de confiance à 95% obtenus par bootstrap.

A une seule exception, nous n'observons pas de différences de polymorphisme significatives entre les différentes parties de la structure du précurseur (Figure 24). Sur les quinze miRs

annotés sur les six allèles S pour lesquels nous avons pu analyser le polymorphisme, seuls cinq possèdent un site variant, quatre sur des miRs d'allèles très récessifs, et un seul sur le miR d'un allèle très dominant. Ainsi, le miR4239 de l'allèle Ah01, les miRS1, miRS3, miRS5 d'Ah03 et le miR1887 de l'allèle Ah13 possèdent chacun 1 SNP. Toutefois, ces mutations observées ne semblent dans aucun des cas étudiés pouvoir perturber les relations de dominance entre allèles. En effet, elles ne sont présentes que sur des mirs qui n'interviennent pas dans les relations de dominance entre allèles et dont nous ne connaissons pour la plupart pas la fonction (cf Chapitre 1, Table S4). Seul le mirS3 de l'allèle Ah03 est responsable du silencing du seul allèle plus récessif Ah01. Toutefois, après analyses de la position du SNP, il s'avère que celui-ci n'est pas situé sur l'isomiR responsable de cette interaction (Tableau S3). On note également la présence d'une délétion de 20 nucléotides de la tige en amont du miR4239 d'Ah01 pour 8 individus Polonais (Figure 23). De plus, trois individus Suisses portent une délétion d'environ 2kb sur la même région, et donc incluant ce même Mir4239 chez Ah01 (Figure S8). Ces résultats sont intéressants au regard du fait que ce miR chez Ah01 ne semble pas avoir de fonction connue.

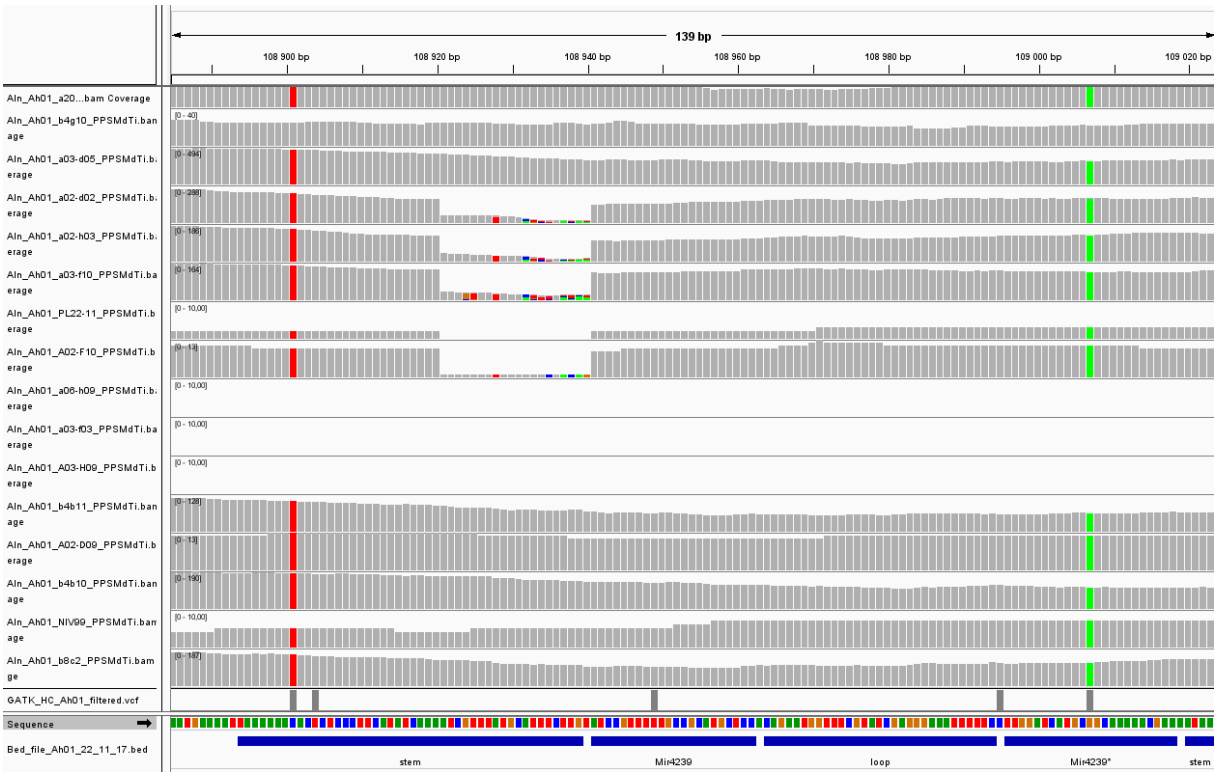


Figure 23 : Délétion d'une partie de la tige en amont du miR4239 chez Ah01. On remarque aussi que certains individus ne sont pas couverts à cette région car il semble qu'ils aient subis une délétion d'environ 2kb (voir Figure S8)

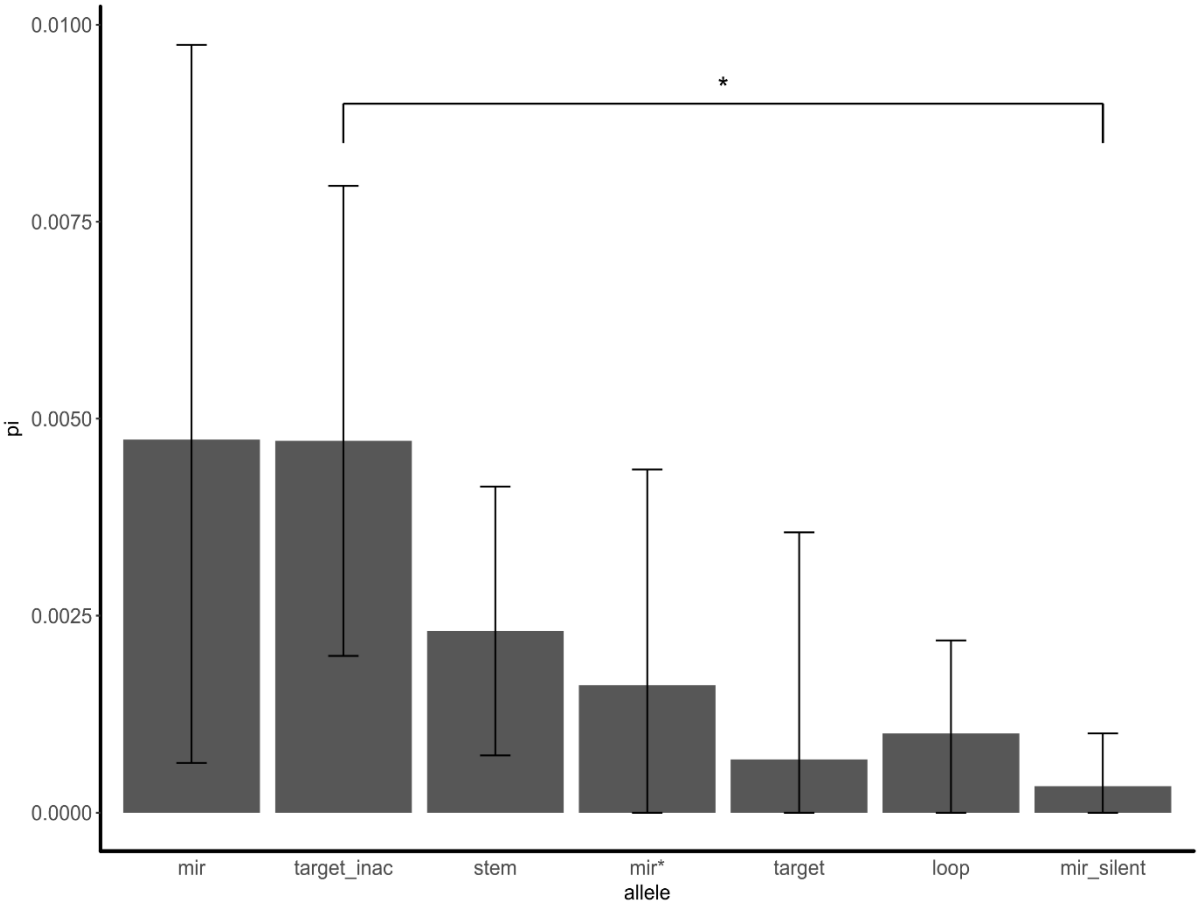


Figure 24 : polymorphisme mesuré au sein des éléments de la machinerie de régulation de la dominance. Les barres verticales représentent les intervalles de confiance à 95% obtenus par bootstrap sur les sites nucléotidiques. Seules les catégories « target_inac » et « mir_silent » sont significativement différentes l'une l'autre.

Contrairement à notre attendu, les mirs silencieux ne présentent pas significativement une valeur de polymorphisme plus élevée que pour le reste de la machinerie, laissant suggérer qu'ils soient toujours sous un régime de sélection purifiante.

Absence de mutations compensatrices

Parmi les 14 SNPs présents dans des sites cibles, nous n'avons pas pu mettre en évidence de variant au niveau des séquences des miRs correspondant, qui auraient pu correspondre à des mutations compensatrices. Réciproquement, aucun des 7 sites variable de miR ne semble être associé à des variant correspondant au niveau de leurs séquences cibles. Chacun de ces sites contribue donc de façon indépendante à modifier le patron d'interaction, sans que cela n'ait de conséquences fonctionnelles prévisibles car les scores d'alignements des différents couples

miR/cible concernés sont tous éloignés du seuil de 18 susceptible d'altérer les relations de dominances établies.

Discussion

Le locus S est une région qui diffère sur bien des aspects par rapport au reste du génome, caractérisée par une très forte divergence entre haplotypes et une accumulation d 'éléments répétés (Goubet *et al.*, 2012). Ces caractéristiques limitent notre capacité à séquencer les éléments de la machinerie de régulation de la dominance, et nos essais basés sur l'approche par séquençage SANGER ont confirmé ces difficultés. La capture de séquence nous a permis de reséquencer ces régions hautement variables de façon quasi exhaustive. Grâce à la capture de séquence, c'est en particulier la première fois que nous pouvons obtenir une estimation du polymorphisme sur l'intégralité des régions intergéniques d'un ensemble d'allèles en population naturelle. Jusqu'à maintenant, nous savions que les régions intergéniques divergeaient considérablement entre allèles différents (au point qu'elles ne peuvent pas être alignées), mais nous n'avions pas d'idée quant au degré de conservation au sein d'une même lignée allélique. Contrairement à ce qui a été observé chez *A. thaliana* (Tsuchimatsu *et al.*, 2018), nous n'avons pas observé de grandes délétions, qui se traduiraient par des portions d'allèles S dont la couverture serait très faible. Ces résultats confirment que la perte de fonction de l'auto-incompatibilité chez *A. thaliana,* synonyme de relâchement des pressions de sélection, a pu permettre l'accumulation de ces grands réarrangements, qui ne peuvent se produire chez *A. halleri* qui conserve un système d'auto-incompatibilité fonctionnel.

Dans l'ensemble, le polymorphisme au locus S est faible, en accord avec les résultats obtenus chez Miege *et al.*, (2001); Charlesworth *et al.*, (2003a,b) et Castric et al. (2010)). Nous confirmons l'absence de corrélation entre niveau de polymorphisme et position dans la hiérarchie de dominance, et ce quelles que soient les régions du locus S étudiées. Ces résultats confirment ceux obtenus dans Castric *et al*, (2010) et sont en contradiction claire avec les attendus théoriques présentés dans cette même étude. Cette absence de corrélation pourrait s'expliquer par un taux de turnover allélique plus important pour les allèles récessifs que pour les allèles dominants, alors que les modèles font l'hypothèse d'un taux de turnover égal entre allèles. Cette interprétation est cohérente avec le ratio π_N/π_S plus élevé pour l'allèle Ah01, qui n'est justement pas attendu pour les allèles les plus récessifs car ils peuvent former des

combinaisons homozygotes au sein desquelles la recombinaison peut se produire, diminuant ainsi les effets d'interférence sélective. Par ailleurs, leur plus forte fréquence en populations naturelles devrait être associée à une efficacité de la sélection naturelle plus élevée. Il est possible qu'un taux de turnover allélique plus élevé puisse contrebalancer cet effet et mener à un taux d'accumulation de mutations délétères plus important. Tout comme Castric *et al.*, (2010), nous avons retrouvé un haut niveau de polymorphisme pour Ah03. Castric *et al.*, (2010) a montré qu'un évènement de recombinaison intra allélique entre AhSRK28 et AhSRK03 était responsable de ce niveau élevé de polymorphisme chez Ah03. Par contre, nous avons mis en évidence des traces d'introgression entre deux lignées alléliques proches : l'allèle Ah12 et l'orthologue Al42. De tels évènements ont aussi été décrits chez *A. halleri* (Castric *et al.*, 2008, 2010) et représentent une des explications possibles concernant la divergence entre les deux groupes de SRK2 phylogénétiquement distincts chez *Brassica oleracea* (Miege *et al.*, 2001). Nos résultats montrent clairement que le polymorphisme d'Ah12 a été augmenté par une introgression récente, produisant deux clades divergents et illustrant le caractère récent de l'introgression entre les deux espèces.

Si des mutants auto-compatibles ont été identifiés chez *A. lyrata* (Mable *et al.*, 2017), il s'agit de la première observation de tels mutants chez *A. halleri* en populations naturelles. Les modèles théoriques montrent que des mutants auto-compatibles peuvent servir d'intermédiaires à l'émergence de nouveaux allèles, en particulier via des mutants pollen (Uyenoyama, 2000, Gervais et al. 2011). Nos résultats suggèrent la présence de deux mutants auto-compatibles pistil et pollen. Il est toutefois peu probable que les mutants auto-compatibles identifiés donnent lieu à une diversification car ce sont des mutants KO pour lesquels des mutations compensatrices sont difficiles à envisager.

Au total, étant donnés leur petite taille et le faible niveau de polymorphisme à l'échelle intra-allélique, on détecte très peu de sites polymorphes sur les précurseurs de petits ARNs. Toutefois, nos résultats montrent que la pression de la sélection sur ces motifs est du même ordre de grandeur que pour les régions codantes du locus S, et donc que contrairement à ce qu'avait suggéré Wright en 1929 (mais dans le contexte spécifique des mutations délétères), la sélection naturelle sur les modificateurs de dominance n'est pas faible. Il est intéressant de constater que la région du miR4239 de l'allèle Ah01 est fortement perturbée, avec la présence de délétion au sein du précurseur voire même de la région entière chez certaines populations.

Etant donnée la position basale de l'allèle Ah01 le long de la hiérarchie de dominance, la présence d'un motif de modificateur de dominance était intrigante dans le sens où celui-ci ne pouvait pas avoir de fonction sans rendre l'allèle Ah01 dominant par rapport à un autre. L'hypothèse que ce motif subit de faibles contraintes fonctionnelles semble ici confirmée. Il pourrait s'agir d'une relique d'une histoire évolutive passée et nous sommes probablement en présence d'un motif ancien, non fonctionnel ou en cours de dégradation et qui évolue de manière neutre.

Bien que le faible nombre de sites polymorphes identifié limite très fortement notre puissance à détecter des différences de polymorphisme entre précurseurs, le fait que nous trouvons peu voire pas de polymorphisme sur les miRs silencieux est intriguant, car ils semblent faire l'objet d'une sélection purifiante de même intensité que celle qui pèse sur les motifs exprimés (alors que ce relâchement des contraintes est détectable sur le miR4239 d'Ah01). Ceci nous amène à penser que ces motifs ne sont peut-être pas silencieux, mais que nous n'avons pas été en mesure de détecter leur expression, ou que celle-ci soit soumise à certaines conditions. Concernant les cibles plus particulièrement, nous n'avons pas pu mettre en évidence de différences significative en termes de pression de sélection entre les cibles fonctionnelles et les cibles inactives. L'absence de mutations compensatrices semble suggérer que tant que la spécificité entre le miR et sa cible n'est pas compromise, ces motifs puissent tolérer quelques substitutions, sans que celles-ci soient éliminées par la sélection purifiante (Liu *et al.*, 2014).

Perspectives

L'approche par capture de séquence s'est relevée être efficace pour obtenir la séquence de l'intégralité du locus S chez *A. halleri*. Les résultats présentés ici, bien que préliminaires pour certains aspects, se sont révélés assez puissants pour mettre en évidence des phénomènes d'introgression, mais aussi pour conduire une analyse comparative du polymorphisme des différentes régions du locus-S. Une limite évidente à notre étude est le faible nombre de lignées alléliques que nous avons pu étudier, et le faible nombre de copies d'allèles au sein de chacune de ces lignées. Maintenant que l'expérience de capture de séquence a été mise au point et s'est révélée puissante, il sera relativement aisé de poursuivre ce travail pour remédier à ces deux limitations.

Bibliographie

Billiard S, Castric V. **2011**. Evidence for Fisher's dominance theory: How many 'special cases'? *Trends in Genetics* **27**: 441–445.

Billiard S, Castric V, Vekemans X. **2007**. A general model to explore complex dominance patterns in plant sporophytic self-incompatibility systems. *Genetics* **175**: 1351–1369.

Castric V, Bechsgaard JS, Grenier S, Noureddine R, Schierup MH, Vekemans X. **2010**. Molecular evolution within and between self-incompatibility specificities. *Molecular Biology and Evolution* **27**: 11–20.

Castric V, Bechsgaard J, Schierup MH, Vekemans X. **2008**. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genetics* **4**.

Charlesworth D, Bartolome C, Schierup MH, Mable BK. **2003a**. Haplotype structure of the stigmatic self-incompatibility gene in natural populations of Arabidopsis lyrata. *Molecular Biology and Evolution* **20**: 1741–1753.

Charlesworth D, Mable BK, Schierup MH, Bartolomé C, Awadalla P. **2003b**. Diversity and linkage of genes in the self-incompatibility gene family in Arabidopsis lyrata. *Genetics* **164**: 1519–1535.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Durand E, Méheust R, Soucaze M, Goubet PM, Gallina S, Poux C, Fobis-loisy I, Guillon E, Gaude T, Sarazin A, *et al.* 2014. Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. *Science* **346**: 1200–1205.

Fisher RA. **1928**. The Possible Modification of the Response of the Wild Type to Recurrent Mutations. *The American Naturalist* **62**: 115–126.

Goubet PM, Bergès H, Bellec A, Prat E, Helmstetter N, Mangenot S, Gallina S, Holl A-C, Fobis-Loisy I, Vekemans X, *et al.* 2012. Contrasted pattern of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. *PLoS genetics* **8**.

Haldane JBS. **1930**. A Note on Fisher's Theory of the Origin of Dominance, and on a Correlation between Dominance and Linkage. *The American Naturalist* **64**: 87–90.

Jovelin R, Cutter AD. **2014**. Microevolution of nematode miRNAs reveals diverse modes of selection. *Genome biology and evolution* **6**: 3049–63.

Kacser H, Burns JA. **1981**. The molecular basis of dominance. *Genetics* **97**: 639–666.

Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. **2009**. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liu Q, Wang F, Axtell MJ. **2014**. Analysis of complementarity requirements for plant MicroRNA targeting using a Nicotiana benthamiana quantitative transient assay. *The Plant Cell* **26**: 741–753.

Llaurens V, Billiard S, Castric V, Vekemans X. **2009**. evolution of dominance in sporophytic self-incompatibility systems: i. genetic load and coevolution of levels of dominance in pollen and pistile. *Evolution* **63**: 2427–2437.

Llaurens V, Billiard S, Leducq JB, Castric V, Klein EK, Vekemans X. **2008**. Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in *Arabidopsis halleri*? *Evolution* **62**: 2545–2557.

Mable BK, Hagmann J, Kim S-T, Adam A, Kilbride E, Weigel D, Stift M. **2017**. What causes mating system shifts in plants? Arabidopsis lyrata as a case study. *Heredity* **118**: 52–63.

McCartney-Melstad E, Mount GG, Shaffer HB. **2016**. Exon capture optimization in amphibians with large genomes. *Molecular ecology resources* **16**: 1084–1094.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**: 1297–1303.

Meyer C-L, Kostecka A a, Saumitou-Laprade P, Créach A, Castric V, Pauwels M, Frérot H. **2010**. Variability of zinc tolerance among and within populations of the pseudometallophyte species Arabidopsis halleri and possible role of directional selection. *The New phytologist* **185**: 130–42.

Miege C, Ruffio-Châble V, Schierup MH, Cabrillac D, Dumas C, Gaude T, Mark Cock J. **2001**. Intrahaplotype polymorphism at the brassica S locus. *Genetics* **159**: 811–822.

MyBaits. http://www.arborbiosci.com/products/targeted-sequencing-kits/.

Nuismer SL, Otto SP. **2005**. Host-parasite interactions and the evolution of gene expression. *PLoS Biology* **3**: 1283–1288.

Otto SP, Bourguet D. **1999**. Balanced Polymorphisms and the Evolution of Dominance. *The American Naturalist* **153**: 561–574.

Pauwels M, Vekemans X, Godé C, Frérot H, Castric V, Saumitou-Laprade P. **2012**. Nuclear and chloroplast DNA phylogeography reveals vicariance among European populations of the model species for the study of metal tolerance, Arabidopsis halleri (Brassicaceae). *The New phytologist* **193**: 916–28.

Picard-tools. http://broadinstitute.github.io/picard/.

Portik DM, Smith LL, Bi K. **2016**. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular ecology resources* **16**: 1069–1083.

Quinlan AR, Hall IM. **2010**. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. **2011**. Integrative genomics viewer. *Nature Biotechnology* **29**: 24.

Schierup MH, Vekemans X, Christiansen FB. **1997**. Evolutionary dynamics of sporophytic self-incompatibility alleles in plants. *Genetics* **147**: 835–846.

Solovyev V, Kosarev P, Seledsov I, Vorobyev D. **2006**. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**: 10.1-10.12.

Šrámková-Fuxová G, Záveská E, Kolár F, Lucanová M, Španiel S, Marhold K. **2017**. Range-wide genetic structure of Arabidopsis halleri (Brassicaceae): Glacial persistence in multiple refugia and origin of the Northern Hemisphere disjunction. *Botanical Journal of the Linnean Society* **185**: 321–342.

Tarutani Y, Shiba H, Iwano M, Kakizaki T, Suzuki G, Watanabe M, Isogai A, Takayama S. **2010**. Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. *Nature* **466**: 983–986.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. **2013**. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**: 178–192.

Tsuchimatsu T, Kakui H, Yamazaki M, Marona C, Tsutsui H, Hedhly A, Meng D, Sato Y, Stadler T, Grossniklaus U, *et al.* 2018. Adaptive reduction of male gamete number in a selfing species. *bioRxiv*: 272757.

Uyenoyama MK. **2000**. Evolutionary dynamics of self-incompatibility alleles in Brassica. *Genetics* **156**: 351–359.

Wright S. **1929**. Fisher's Theory of Dominance. *The American Naturalist* **63**: 274–279.

Wright S. **1934**. Physiological and Evolutionary Theories of Dominance. *American Naturalist* **68**: 25.

Yasuda S, Wada Y, Kakizaki T, Tarutani Y, Miura-uno E, Murase K, Fujii S, Hioki T, Shimoda T, Takada Y, *et al.* 2016. A complex dominance hierarchy is controlled by polymorphism of small RNAs and their targets. *Nature Plants* **16206**: 1–6.

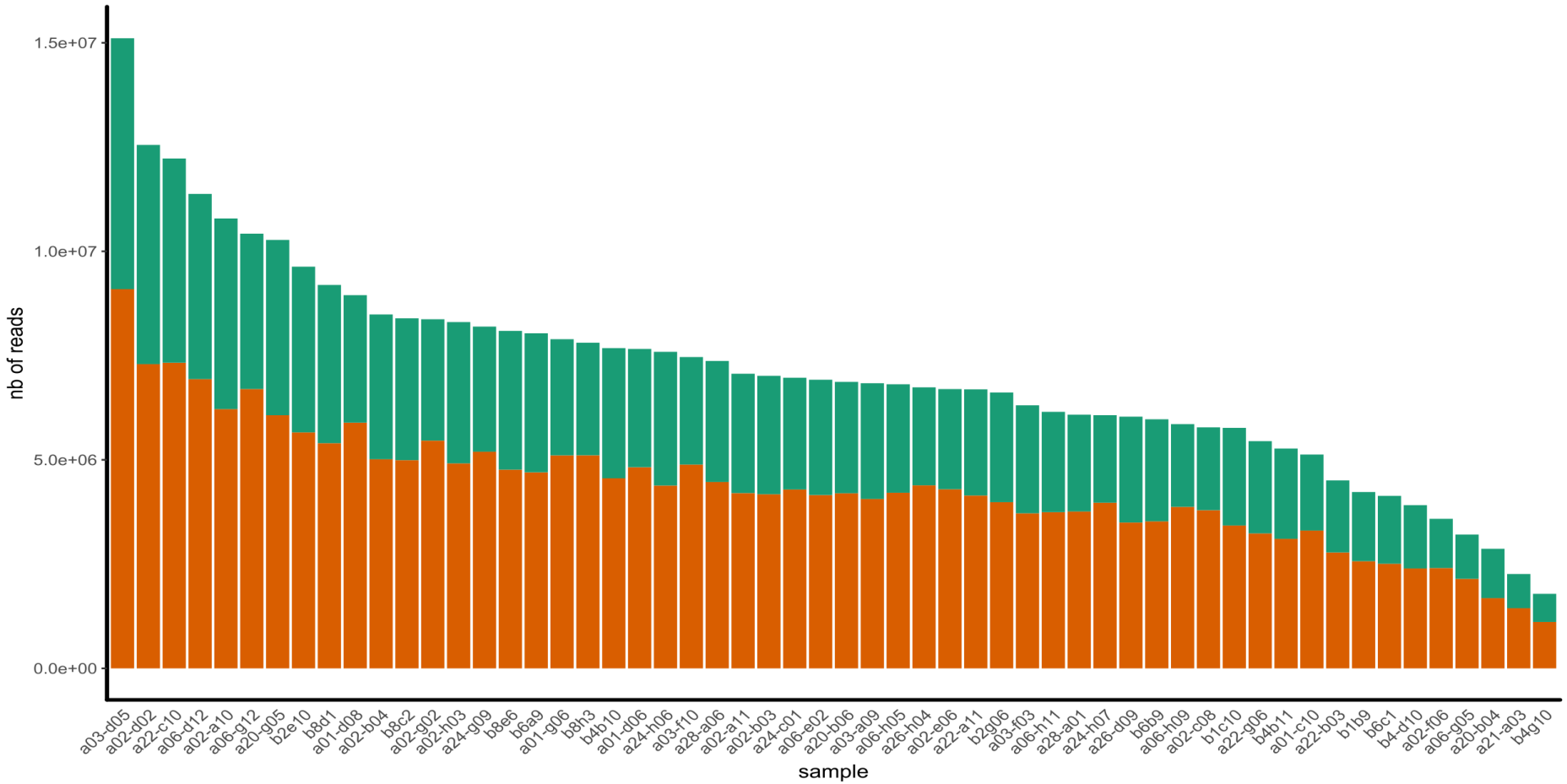


Figure S1 : Représentation du nombre de lecture total obtenu pour chaque individu. Les deux séries de capture sont incluse. En orange : le nombre de read alignés sur une cible. En Jade : le nombre de read restant. La somme des deux représente le total de read obtenus par individu. Les échantillons sont triés de la gauche vers la droite, en fonction du nombre de read total.

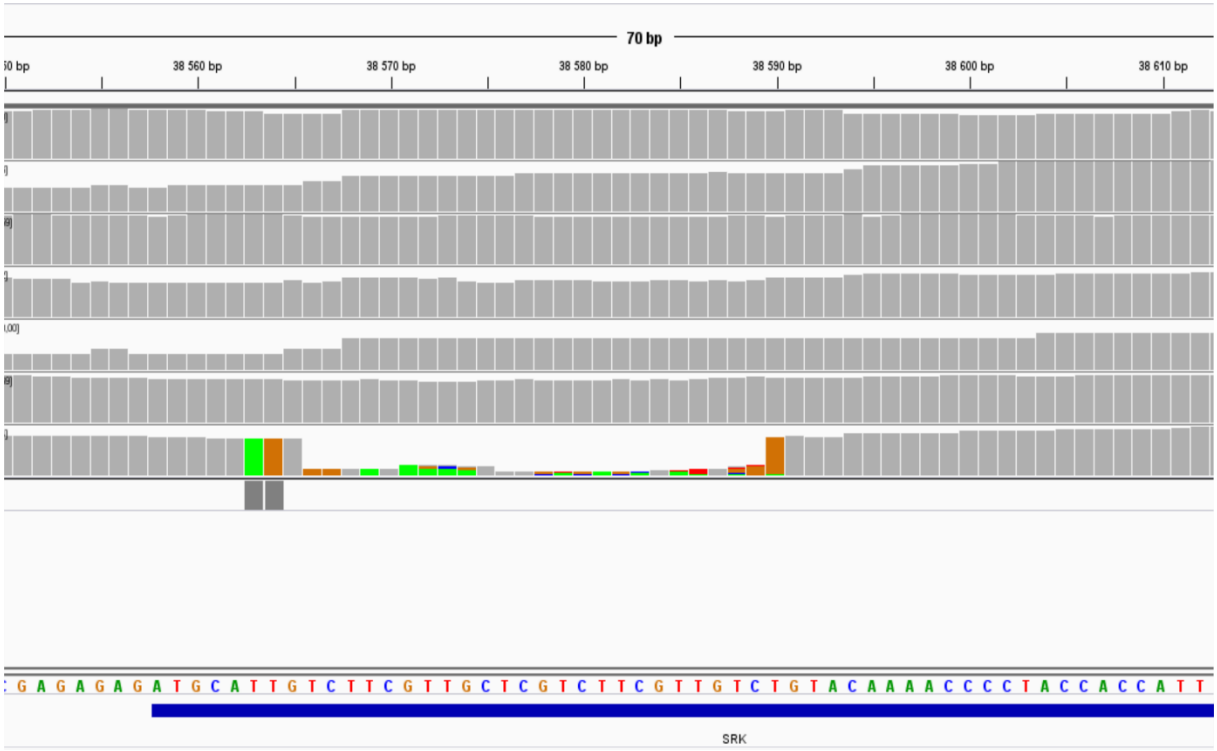


Figure S2 : Illustration de la délétion de 24 nt sur la séquence du domaine S de SRK20 pour l'individu Slovaque A02-F10.

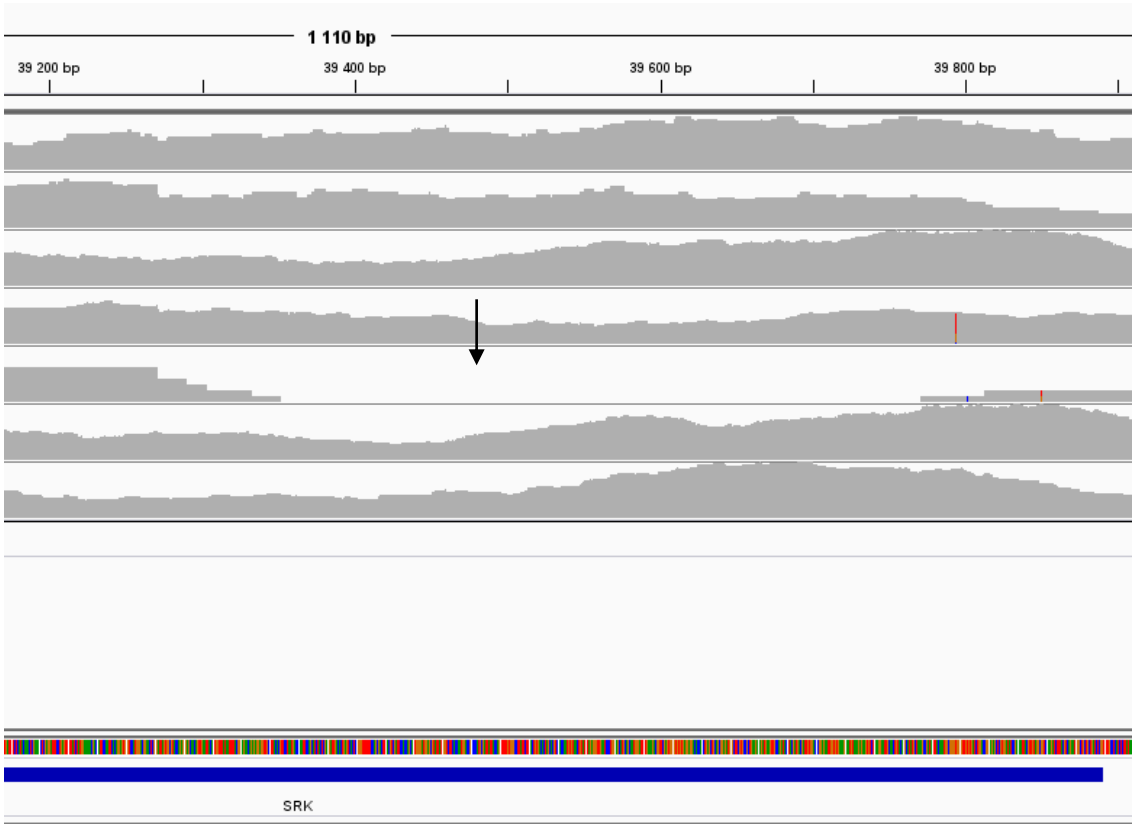


Figure S3 : Illustration de la partie non couverte du domaine S de SRK20 pour l'individu A26-D05.

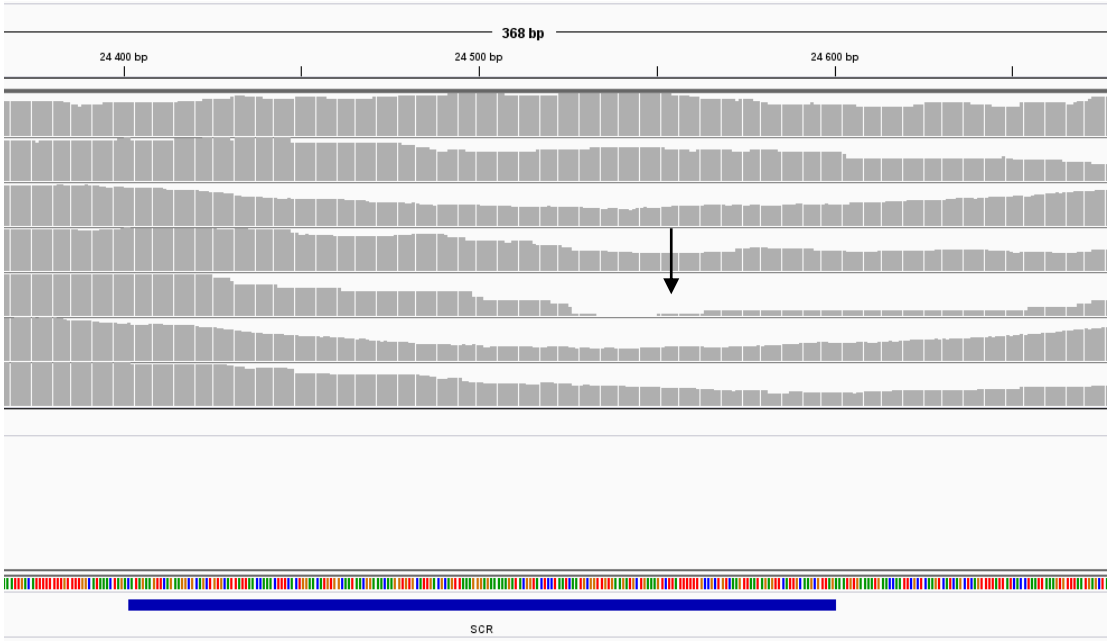


Figure S4 : Illustration de la partie non couverte du 2e exon de SCR20 pour l'individu A26-D05.

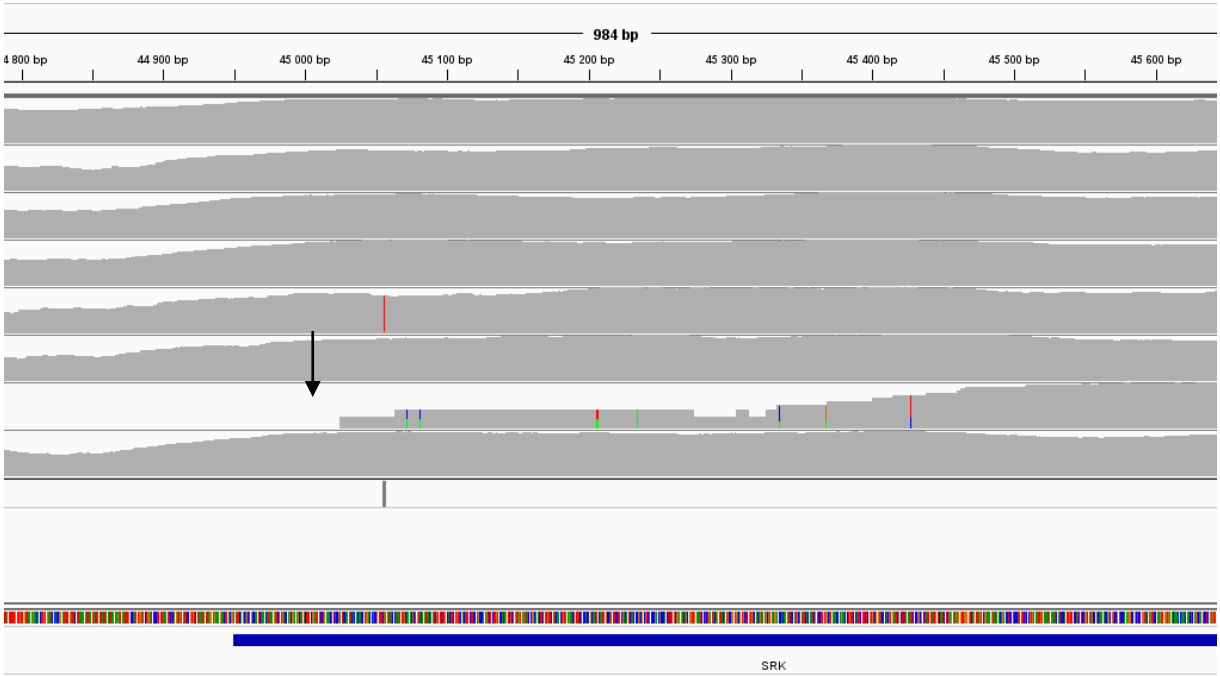


Figure S5 : Illustration de la partie non-couverte de la fin de la partie 3' du domaine S de SRK10 chez l'individu A26-D05.

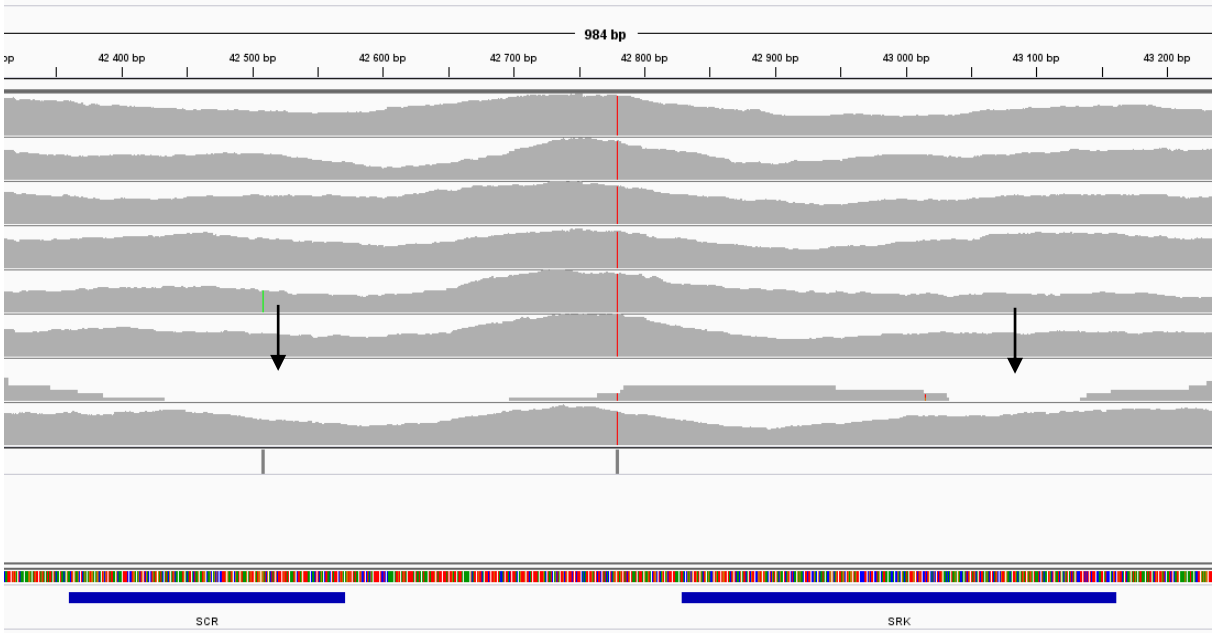


Figure S6 : Illustration des parties non-couvertes du 2^{e} exon de SCR10 et d'une partie du dernier exon du domaine kinase de SRK10 chez l'individu A26-D05.

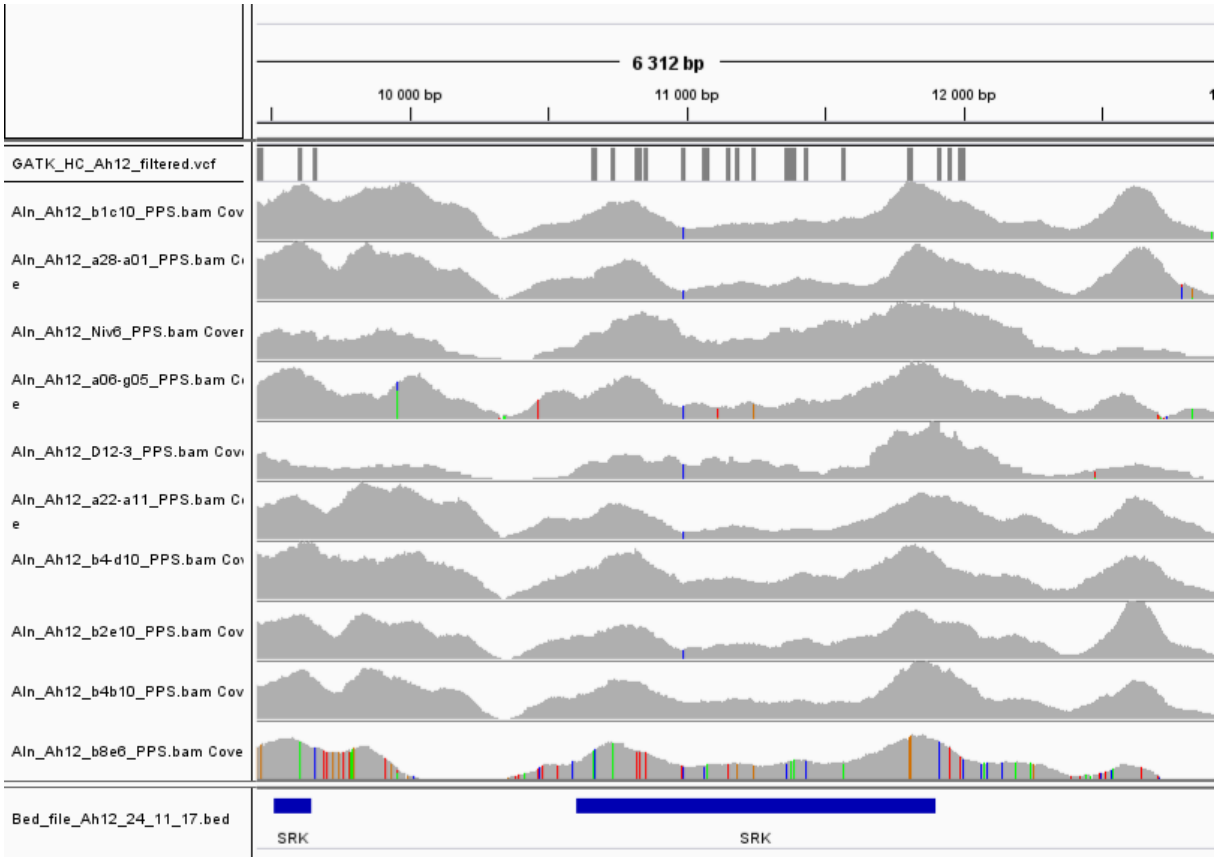


Figure S7 : illustration de l'empreinte de l'introgression d'Al42 dans un contexte Ah12 chez l'individu b8e6. Outre les différences dans la région codante de *SRK*, on en constate également les traces dans les régions intergéniques.

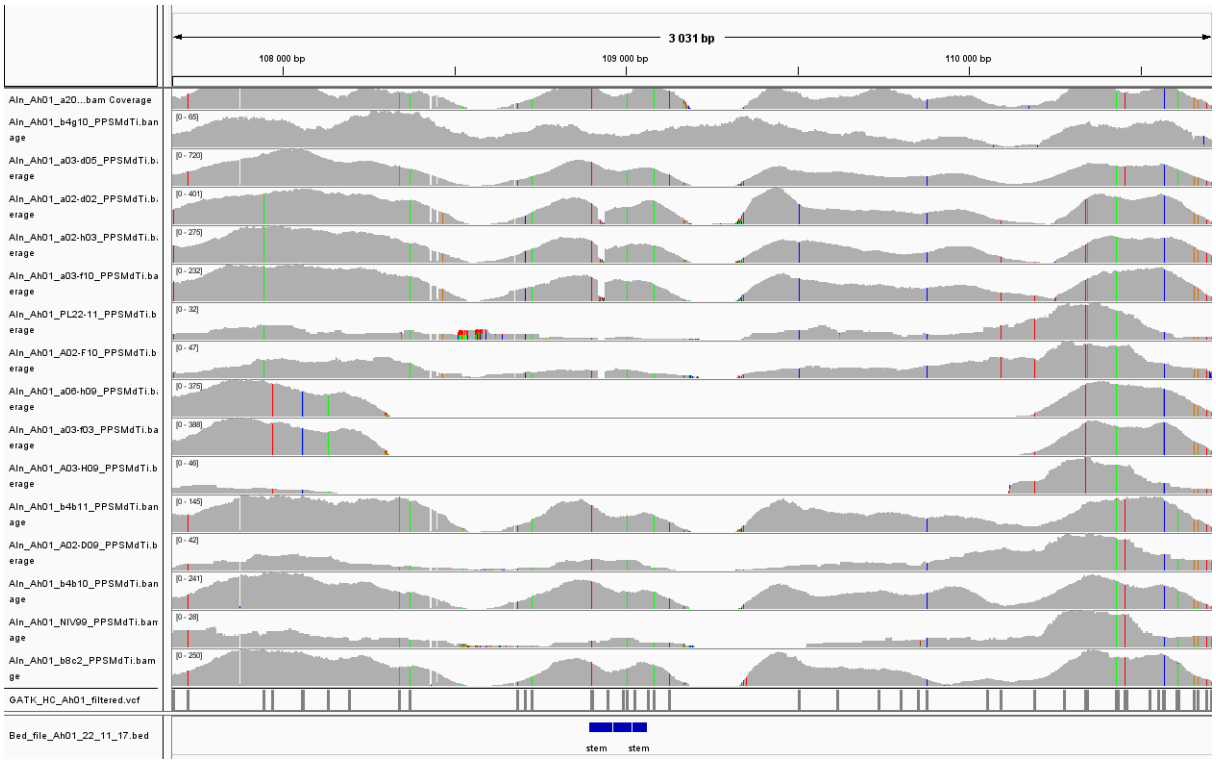


Figure S8 : Délétion d'environ 2kb de la région intergénique comprennant le miR4239 chez Ah01.

Sample	Total reads nb	Reads PP	PP on target	specificity	Enrichment factor EF	Serie
A02-D09	938 054	494 504	365 591	0,390	19,796	1
A02-F10	1 067 606	628 798	465 015	0,436	22,125	1
A03-F01	814 252	518 800	377 565	0,464	23,553	1
A03-H09	580 034	415 816	303 325	0,523	26,563	1
A06-C12	1 135 382	657 756	481 110	0,424	21,524	1
A22-C12	623 696	376 888	277 352	0,445	22,588	1
A26-D05	696 304	137 888	98 334	0,141	7,173	1
aha2-18-2-8	1 296 924	702 872	515 512	0,397	20,190	1
D12-3	510 204	279 810	206 587	0,405	20,567	1
D8-18	661 532	422 260	311 583	0,471	23,924	1
MOR6	948 364	638 754	471 169	0,497	25,236	1
Niv18	962 418	645 040	473 311	0,492	24,980	1
Niv6	953 060	527 982	390 428	0,410	20,808	1
NIV99	261 276	165 490	120 816	0,462	23,488	1
PL22-11	1 080 236	661 884	489 914	0,454	23,037	1
PL22-6	1 199 462	454 236	337 301	0,281	14,284	1
mean	858 050	483 049	355 307	0,418	21,240	
S.D	232 382	141 119	103 981	0,061	3,078	

Tableau S1 : Valeurs résumées de la capture. Pour chaque échantillons sont donné: « total reads nb » : le nombre de read total ; « Reads PPé : le nombre de reads « properly paired » ou « PP » ; « PP on target » : reads PP qui s'alignent sur les régions cibles ; specificity : calcul de

la spécificité comme étant le nombre de reads PP on target sur le nombre de reads total ; le facteur d'enrichissement (calculé selon la formule cf. matériel et méthode) en enfin la série de capture pour cet'échantillon.

Sample	Total reads nb	Reads PP	PP on target	specificity	Enrichment factor EF	Serie
a01-c10	5 127 532	4 290 316	2 939 935	0,573	29,124	2
a01-d06	7 659 512	6 382 948	4 345 768	0,567	28,819	2
a01-d08	8 948 638	7 442 308	5 280 337	0,590	29,972	2
a01-g06	7 892 600	6 620 738	4 594 439	0,582	29,569	2
a02-a10	10 787 394	8 609 676	5 434 158	0,504	25,588	2
a02-a11	7 064 776	5 453 258	3 498 065	0,495	25,151	2
a02-b03	7 014 408	5 504 294	3 557 849	0,507	25,764	2
a02-b04	8 486 350	6 679 340	4 278 613	0,504	25,609	2
a02-c08	5 779 410	4 719 542	3 313 830	0,573	29,125	2
a02-d02	12 550 802	10 219 082	6 527 615	0,520	26,418	2
a02-e06	6 696 902	5 422 550	3 747 562	0,560	28,424	2
a02-f06	3 586 156	2 920 396	2 069 703	0,577	29,315	2
a02-g02	8 370 980	6 931 520	4 867 817	0,582	29,538	2
a02-h03	8 303 592	6 794 064	4 388 561	0,529	26,846	2
a03-a09	6 836 784	5 471 690	3 503 063	0,512	26,026	2
a03-d05	15 108 568	12 412 526	8 131 882	0,538	27,339	2
a03-f03	6 307 756	5 028 570	3 211 328	0,509	25,860	2
a03-f10	7 464 940	6 222 206	4 357 388	0,584	29,650	2
a06-d12	11 376 382	9 323 742	6 152 610	0,541	27,471	2
a06-e02	6 920 960	5 498 108	3 584 048	0,518	26,304	2
a06-g05	3 208 678	2 657 530	1 889 813	0,589	29,916	2
a06-g12	10 422 938	8 627 302	5 984 590	0,574	29,165	2
a06-h05	6 811 918	5 603 632	3 741 630	0,549	27,900	2
a06-h09	5 856 532	4 848 084	3 432 506	0,586	29,771	2
a06-h11	6 151 220	5 070 960	3 289 187	0,535	27,161	2
a20-b04	2 867 452	2 196 880	1 413 388	0,493	25,037	2
a20-b06	6 869 202	5 427 550	3 588 737	0,522	26,537	2
a20-g05	10 273 084	8 180 360	5 361 720	0,522	26,511	2
a21-a03	2 262 792	1 779 550	1 202 980	0,532	27,004	2
a22-a11	6 690 490	5 325 714	3 561 730	0,532	27,041	2
a22-b03	4 507 370	3 632 722	2 402 487	0,533	27,074	2
a22-c10	12 225 596	9 526 272	6 261 108	0,512	26,013	2
a22-g06	5 448 110	4 125 978	2 671 332	0,490	24,906	2
a24-c01	6 968 682	5 639 766	3 746 961	0,538	27,312	2
a24-g09	8 194 988	6 396 988	4 409 059	0,538	27,328	2
a24-h06	7 590 098	5 829 020	3 684 483	0,485	24,657	2
a24-h07	6 071 036	4 829 560	3 434 066	0,566	28,732	2
a26-d09	6 035 716	4 726 590	3 013 881	0,499	25,364	2
a26-h04	6 738 566	5 524 836	3 857 395	0,572	29,077	2
a28-a01	6 082 790	4 807 862	3 199 335	0,526	26,716	2
a28-a06	7 371 014	5 819 284	3 814 354	0,517	26,285	2
b1b9	4 228 340	3 248 774	2 115 800	0,500	25,417	2
b1c10	5 766 770	4 599 128	2 963 699	0,514	26,105	2
b2e10	9 631 840	7 175 902	4 694 898	0,487	24,759	2
b2g06	6 614 346	5 114 856	3 328 490	0,503	25,561	2
b4-d10	3 913 846	3 030 964	1 990 224	0,509	25,829	2
b4b10	7 678 898	5 971 404	3 861 013	0,503	25,540	2
b4b11	5 271 340	4 189 986	2 700 555	0,512	26,023	2
b4g10	1 787 656	1 415 486	945 156	0,529	26,856	2
b6a9	8 034 472	6 308 468	4 019 549	0,500	25,412	2
b6b9	5 972 584	4 578 828	2 961 064	0,496	25,183	2
b6c1	4 136 340	3 303 346	2 158 206	0,522	26,503	2
b8c2	8 393 438	6 741 192	4 411 058	0,526	26,694	2
b8d1	9 193 518	7 236 516	4 661 557	0,507	25,755	2
b8e6	8 092 060	6 384 100	4 117 060	0,509	25,843	2
b8h3	7 808 256	6 540 194	4 622 108	0,592	30,068	2
mean	7 097 436	5 685 044	3 773 138	0,53	27,02	
S.D	1 840 053	1 497 646	995 691	0,03	1,34	

TableauS1 (suite)

allele list	Nb	Bac clone availability
AhSRK01	31	1
AhSRK03	16	1
AhSRK12	10	1
AhSRK10	8	1
AhSRK13	8	1
AhSRK20	7	1
AhSRK15	5	1
AhSRK21	5	0
AhSRK04	4	1
AhSRK36	3	1
AhSRK02	3	1
AhSRK05	2	0
AhSRK08	2	0
AhSRK25	2	0
AhSRK32	2	1
AhSRK36/AlSRK33	2	0
AhSRK37	2	0
AlSRK23	2	0
AhSRK03/AhSRK08	1	0
AhSRK05/AhSRK08	1	0
AhSRK11	1	0
AhSRK18	1	0
AhSRK19	1	0
AhSRK26	1	0
AhSRK28	1	0
AhSRK33	1	0
AhSRK34	1	0
AhSRK43	1	1
AhSRK43/AhSRK08	1	0
AlSRK04	1	0
AlSRK30	1	0
AlSRK44	1	0
AlSRK45	1	0
AlSRK50	1	1
AlSRK19	1	0

Tableau S2 : Effectif des allèles après génotypage. La disponibilité d'un BAC clone est également indiquée. Les individus ont été exclus de l'expérience dans les cas où plusieurs allèles sont suggérés (e.g. AhSRKx/AhSRKx).

h03	MirS3_precursor	TATTGTATTCTATTTTGCACGATCAGTAACAACCAATGGTTTCAGATTTTGCGAGTAACCAATAAAACCCAAAAGAAGCTAATATATTACCTTGAACATAGTCA…
	MirS3 chap 1	AAGCTAATATATTACCTTGAACAT
	MirS3 chap 2	AACAACCAATGGTTTCAGATT
	MirS3_alt chap 2	AACAAACAATGGTTTCAGATT

Tableau S3 : Partie 3’ du précurseur du miRS3 d’Ah03. Le miRS3 utilisés lors du chapitre 1 est celui responsable de la dominance d’Ah03 sur Ah01 mais n’est pas le plus abondant. Le miRS3 du chapitre 2 est le plus abondant de cette partie du précurseur. Tous deux ne sont pas situés au même endroit au sein de la structure tige/boucle et forment par conséquent des isomiRs. La mutation découverte au chapitre 2 est indiquée en rouge et n’est donc pas présente au sein du miRS3 du chapitre 1, et ne perturbe donc pas la relation de dominance entre les deux allèles

Chapitre 3

Sélection balancée et structure de la diversité des régions liées au locus d'auto-incompatibilité chez *Arabidopsis halleri*

Auteurs

Nicolas Burghgraeve, Mathieu Genete, Christelle Blassiau, Anne-Catherine Holl, Elisa Prat, William Marande, Vincent Castric.

Introduction

La sélection naturelle est le processus par lequel une mutation qui apparait par hasard dans une population, est favorisée ou au contraire défavorisée au regard de l'avantage ou du désavantage que cette mutation apporte aux individus qui la portent. Si ce processus va à son terme, cette mutation peut soit se fixer soit disparaître, aboutissant à la perte du polymorphisme généré par la mutation. Dans les cas de sélection dite « balancée » cependant, le polymorphisme est protégé et les mutations tendent à se perdre plus difficilement soit parce qu'elles sont avantagées lorsqu'elles deviennent rares (sélection fréquence-dépendante négative), soit parce que les hétérozygotes ont une valeur sélective supérieure à celle des homozygotes (superdominance), soit enfin parce que les pressions de sélection sont hétérogènes dans l'espace ou dans le temps. Quelle que soit sa forme, la sélection modifie en profondeur les patrons de polymorphisme observés en populations naturelles pour les sites concernés.

Dans de nombreux cas cependant, il est clair que la sélection sur un locus peut influencer la diversité aux locus neutres associés (Cutter & Payseur, 2013), et la question de l'importance de ce phénomène est d'une grande actualité maintenant qu'il est possible de décrire de façon exhaustive les patrons de polymorphisme à l'échelle de génomes complets (Kern & Hahn, 2018). Ces phénomènes d'entrainement de la variabilité neutre peuvent concerner toutes les formes de sélection, qu'elle soit positive, négative ou balancée. Les premiers sont appelés « auto-stop » (Hitchhiking en anglais) et ont été proposés par Maynard Smith & Haigh (1974). Ils ont proposé le fait que la propagation d'une mutation favorable devrait réduire la diversité à un site neutre lié, à la condition que cet allèle favorable soit apparu au sein d'un seul haplotype. Si aucune recombinaison n'a lieu, la propagation de cette mutation favorable va aussi propager tous les variant neutres présents sur cet haplotype. L'association entre le locus sous sélection et les sites neutres liés diminuant rapidement en fonction du taux de recombinaison. De la même manière, la sélection d'arrière-plan (ou Background Selection, Charlesworth *et al.*, 1993) provoque l'élimination de variant neutres liés à un locus présentant une mutation délétère. Ce qui amène aussi à une faible diversité aux locus neutres, plus particulièrement dans des régions où la recombinaison est faible. Dans les cas de sélection

balancée, les sites liés vont accumuler des mutations qui différencient les haplotypes entre eux, ce qui, à l'échelle de la population, va être représenté par un pic de polymorphisme neutre autour du locus sous sélection (DeGiorgio *et al.*, 2014).

Au-delà des effets d'entrainement sur le polymorphisme neutre, il se peut aussi que deux sites proches soient soumis à des régimes de sélection distincts, auquel cas les processus de fixation des deux sites s'influencent mutuellement. Ces phénomènes d'interférences sélectives, (ou Hill-Roberston effect, Hill & Robertson, 1966) peuvent amener à plusieurs situations. Lorsqu'une mutation favorable apparait sur un haplotype, celle-ci va augmenter en fréquence via un balayage sélectif. Si deux mutations favorables sont liées, il peut se produire plusieurs cas de figures. Si elles sont sur le même haplotype, alors le coefficient de sélection de l'haplotype en question en sera d'autant plus fort, ce qui accélèrera la vitesse à laquelle il va se fixer. Par contre la seconde mutation favorable peut apparaitre sur un autre haplotype, et dans ce cas l'une ou l'autre des mutations peut se fixer (même si cela nécessitera plus de temps) en fonction de la différence dans l'intensité de l'avantage fourni, ou alors les deux mutations peuvent se maintenir dans un état d'équilibre plus ou moins long en fonction du taux de recombinaison, il en résulte que l'efficacité de la sélection positive sur chacun des sites est réduite (Comeron *et al.*, 2008). De la même manière, si deux mutations délétères sont liées, alors l'efficacité de la sélection négative sur chacun des sites est réduite (McVean & Charlesworth, 2000). L'ampleur et impact de tels effets d'interférence sur l'efficacité de la sélection à l'échelle du génome restent à ce jour mal connus (Slotte, 2014).

Le locus d'auto-incompatibilité est un cas d'école de la sélection balancée, et nous offre l'opportunité d'étudier plus en détails les processus de sélection liée via l'étude du polymorphisme des régions flanquantes. Plusieurs études ont montré une augmentation de polymorphisme aux abords du locus S chez *A. lyrata* (Figure 25, Kamau & Charlesworth, 2005) puis chez *A. halleri* (Figure 26, Ruggiero *et al.*, 2008; Roux *et al.*, 2013).

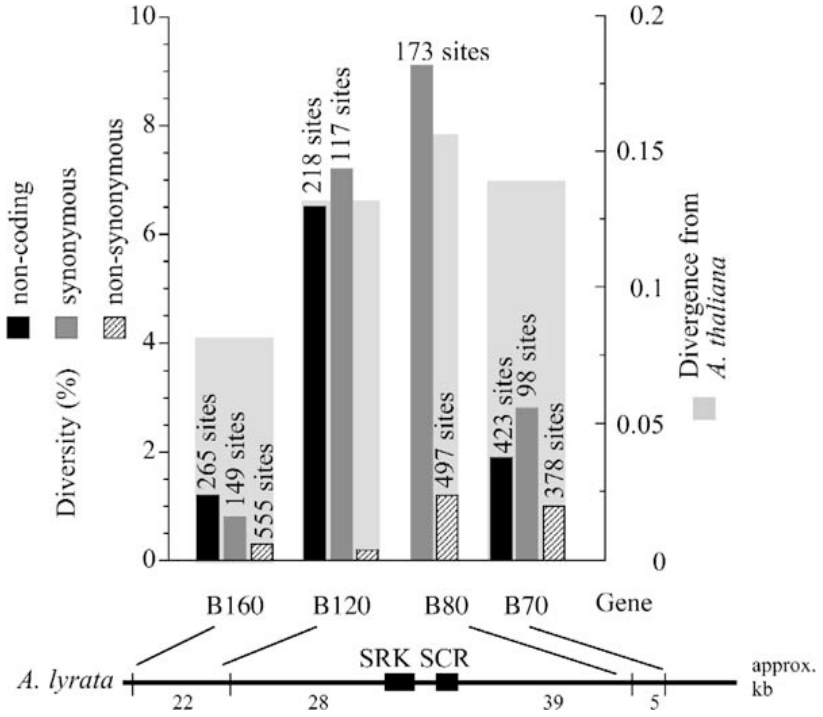


Figure 25: Diversité des sites flanquants du locus S, et divergence neutre par rapport à l'orthologue *A. thaliana*. Les distances approximatives depuis le gène du locus S le plus proche sont données en bas. Les barres représentent la diversité nucléotidique (en %, axe de gauche) synonyme, non synonyme et non codante pour chacun des quatre gènes (sauf B80 qui ne possède pas d'intron). Les nombres de sites pour chaque gène sont indiqués au-dessus des barres. La divergence est représentée par les barres gris clair en arrière-plan. Kamau & Charlesworth (2005)

On constate bien chez *A. lyrata* une augmentation de la diversité synonyme par rapport à la divergence pour B80 (ou *U-box*) et pour B120 situés respectivement à 39 et 28kb du gène du locus S le plus proche, mais ceci n'est plus vrai pour les gènes suivants B160 et B70 situé à 44 et 50kb respectivement.

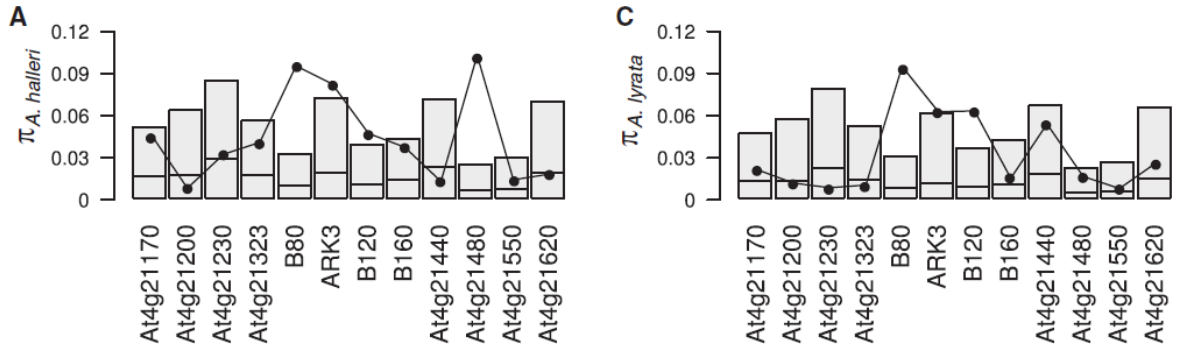


Figure 26: Histogramme de la diversité synonyme observée par paire de base pour 12 gènes flanquants du locus S chez *A. halleri* (A) et chez *A. lyrata* (C) ainsi que la distribution attendue obtenue par simulation de coalescence sous l'hypothèse nulle qu'il n'existe aucun lien avec le site sous sélection. Les boites représentent la distribution de l'intervalle de confiance à 95% obtenue par simulation, les traits verticaux représentent la médiane de chaque distribution. Les points représentent les données observées. Roux et al. (2013)

On retrouve les mêmes résultats chez *A. lyrata* sur la base d'un autre échantillonnage (Roux *et al.*, 2013 , Figure 26C) mais aussi pour *A. halleri* (Figure 26A). Les seuls gènes pour lesquels il semble y avoir un effet de la sélection au locus-S sont les plus proches : *B80*, *ARK3* et *B120*. Ces résultats nous informent que l'augmentation du polymorphisme est probablement très localisée. Toutefois, ces études ne nous permettent pas de mesurer exactement l'étendue de l'augmentation du polymorphisme, pour trois raisons. D'une part les valeurs de polymorphisme ont été obtenues sur des fragments de PCR et non sur les séquences entières des gènes. D'autre part la distance entre les gènes par exemple *B80* et *At4g21323* est grande (~20kb), et il existe des gènes présents entres eux (en l'occurrence, 4 gènes) qui n'ont pas été analysés. Enfin, le nombre d'individus analysés est relativement faible.

La question de l'ampleur de la région dont le polymorphisme est affecté par la liaison au locus S est particulièrement importante dans le cadre de l'accumulation d'un fardeau « lié » de mutations délétères (Uyenoyama, 2005). En effet, comme nous venons de le voir, les régions autour du locus-S contiennent des gènes. Par-delà la diversité neutre qu'ils devraient accumuler, il est possible que ces gènes accumulent des mutations délétères. De telles mutations sont normalement exposées à la sélection, toutefois la proximité au locus-S provoque plusieurs biais. D'une part l'excès en hétérozygotie du locus S diminue la fréquence

à laquelle ces mutations forment des combinaisons homozygotes et donc peuvent s'exprimer si elles sont récessives. D'autre part, si elles apparaissent en liaison à des allèles S qui sont en dessous de leur fréquence d'équilibre au sein d'une population, elles vont tendre à augmenter en fréquence au lieu d'être éliminées, dans un cas typique d'interférence Hill-Robertson (1966). Ce fardeau lié peut-avoir des conséquences importantes, notamment celle de réduire le taux d'émergence de nouveaux allèles (Uyenoyama, 2003), et de contribuer de façon substantielle à la dépression de consanguinité (Glémin *et al.*, 2001), affectant à leur tour les conditions de maintien de l'auto-incompatibilité. L'existence d'un tel fardeau de mutations a déjà été mise en évidence via des mesures de fitness. En forçant l'autofécondation, Stone (2004) et Llaurens *et al.*(2009) ont montré pour *Solanum carolinense* et *A. halleri* respectivement, un fort coût à l'homozygotie (cf introduction générale pour *S. carolinense* et (Figure 11 & 27).

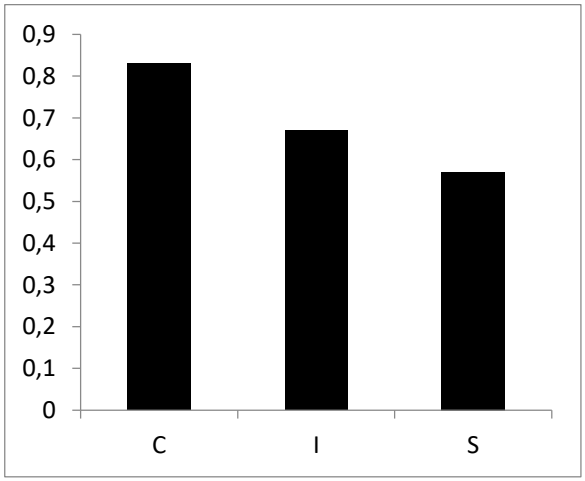


Figure 27: Histogramme représentant la largeur des feuilles (cm) de plantes d'*A. halleri* issues de plusieurs types de croisement. **C** : Compatible (croisement de contrôle) ; **I** : incompatible (descendance homozygote produite par croisement entre plantes de lignées différentes) et **S** : self (descendance homozygote issu de croisement entre clone). Plus le niveau de consanguinité en allèles S augmente, plus le coût s'en ressent sur la fitness des descendants. (Produit à partir des données de Llaurens *et al.*, 2009)

La présence d'une hiérarchie de dominance entre allèles S chez *A. halleri* introduit une asymétrie entre allèles de ce fardeau lié en fonction de la classe de dominance à laquelle ils appartiennent. En effet, on s'attend à ce que les allèles récessifs puissent se retrouver à l'état homozygote, expriment des mutations délétères, et donc soient éliminés par sélection. Au

contraire, les allèles les plus dominants sont systématiquement exprimés, et ne se retrouvent probablement jamais à l'état homozygote, ce qui fait que les mutations délétères auxquelles ils sont liés peuvent s'accumuler pendant de longues périodes sans que celles-ci soient éliminées par la sélection. On s'attend donc à ce que les allèles dominants puissent accumuler un fardeau plus important que les allèles récessifs (Figure 28).

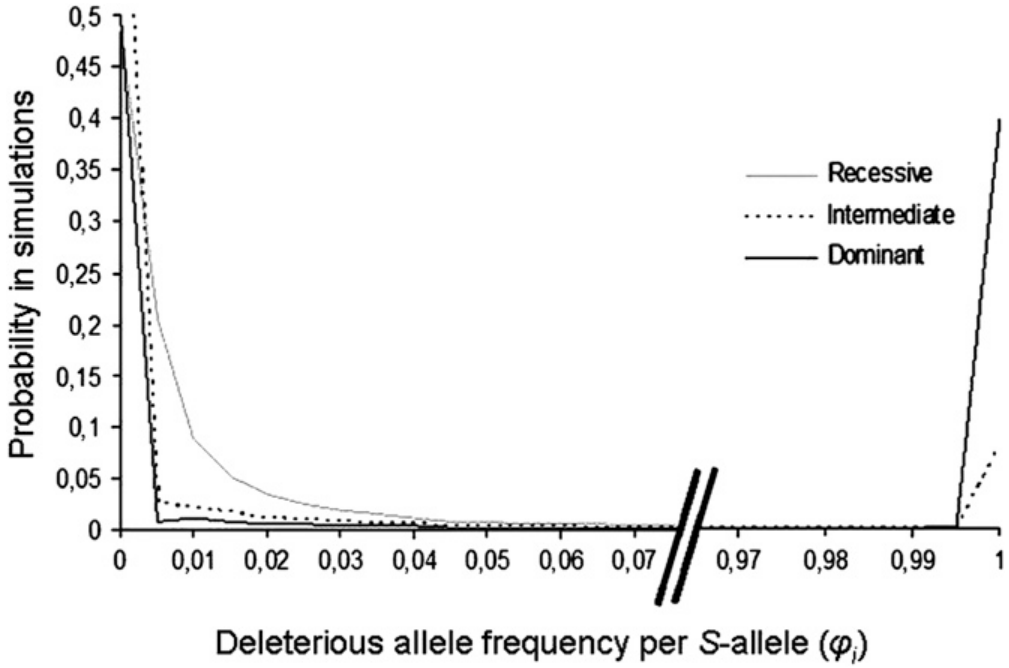


Figure 28: Distribution des fréquences des mutations délétères liées avec des allèles S appartenant à différentes classe de dominances (dominant, intermédiaire et récessif). On remarque que les allèles dominants devraient accumuler en théorie plus de mutations fortement délétères que les allèles récessifs. Llaurens *et al.* (2009)

L'existence du fardeau exige une association forte entre les allèles S et le cortège de mutations délétères qui leur sont liés. A ce jour cependant, l'étendue de la structure haplotypique de la région du locus S est mal connue. Les résultats obtenues par (Kamau & Charlesworth, 2005, Figure 29) chez *A. lyrata* montrent que les polymorphismes des gènes immédiatement aux abords du locus S sont largement structurés par l'allèle S porté, bien plus que par l'origine géographique des accessions dont ils sont issus, tandis que les gènes à des distances plus importantes montrent un patron inverse, l'association avec les allèles S devenant négligeable devant l'effet de structuration par populations pour des gènes distants voire non liés. Là

encore cependant, le nombre d'échantillons et de gènes considérés restait très faible, limitant la portée des conclusions.

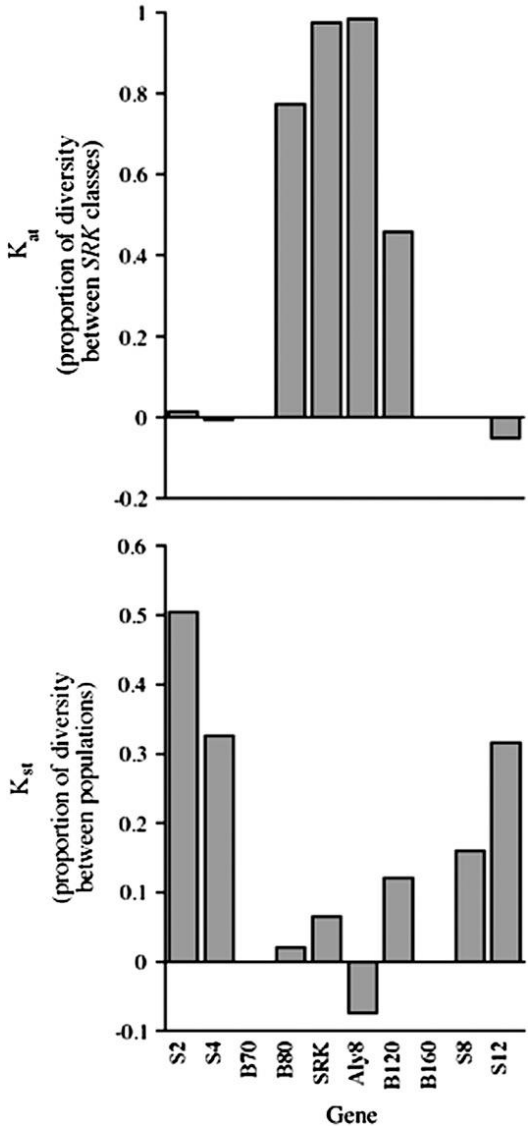


Figure 29: Proportion de la diversité entre allèles *SRK* (haut) et entre populations (bas) pour différents gènes dans et autour du locus-S. On peut constater que la proportion de diversité expliqué par les allèles est forte pour les gènes du locus S ou très proches, mais qu'ensuite la structuration se fait par population (données de B70 et B160 manquante, (Kamau *et al.*, 2007).

Dans ce chapitre, nous nous sommes intéressés au polymorphisme des régions flanquantes au locus-S, et ce via deux approches. Dans un premier temps, nous avons mesuré le polymorphisme des régions flanquantes à l'échelle du nucléotide sur des séquences complètes de la région du locus S produites par capture de séquences. Puis, dans un second

temps, nous avons déterminé l'intensité ainsi que l'étendue de la structure haplotypique des régions associées grâce à des données de séquences issues de clones BACs produits à partir de plantes provenant de plusieurs populations. Nous avons testé l'hypothèse selon laquelle le polymorphisme devrait être inversement corrélé à la distance du locus S, en exploitant tant les polymorphismes des régions géniques (synonymes et non synonymes) que des régions intergéniques, augmentant ainsi de façon substantielle le nombre de sites analysés. Nous avons comparé les patrons d'accumulation du polymorphisme synonyme et non-synonyme, pris comme indication de l'efficacité de la sélection. Enfin, nous avons évalué l'échelle à laquelle les polymorphismes des régions flanquantes sont associés spécifiquement aux allèles S.

Matériel & méthodes

Choix des gènes analysés

Une limitation commune aux études précédentes sur l'effet de la sélection balancée sur le polymorphisme de la région du locus S est l'échantillonnage très incomplet des séquences analysées. Dans tous les cas il s'agissait de courts fragments d'un petit nombre de gènes échantillonnés à des distances variables, souvent importantes. Pour remédier à cette limitation et atteindre une description complète de la région incluant les gènes mais également les régions intergéniques, nous avons combiné deux approches. Dans un premier temps nous avons utilisé une approche par capture de séquences, permettant d'analyser un nombre important d'échantillons. Dans un second temps, nous avons analysé la phase de ces polymorphismes en séquençant un ensemble plus limité de clones BAC issus de plusieurs populations naturelles. Une représentation de la région génomique concernée, indiquant la position physique des gènes analysés et comparant la position des gènes inclus dans les études précédentes (Kamau & Charlesworth 2005, Ruggiero et al. 2008, Roux *et al.* 2013) et la présente étude est détaillée en Figure 30.

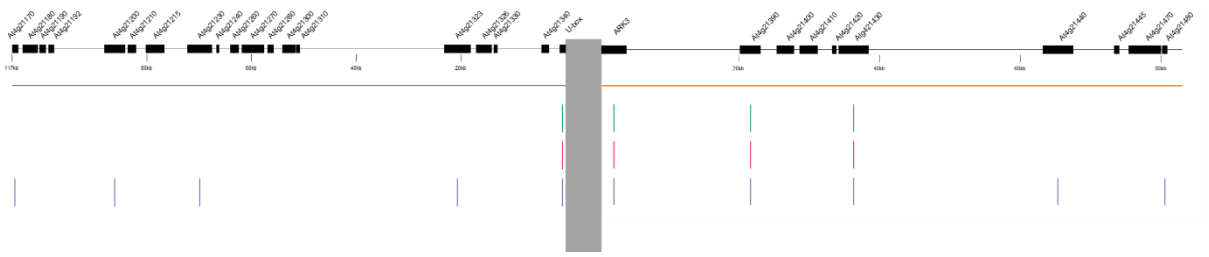


Figure 30: Comparaison de l'échantillonnage des gènes à travers les diverses études. La bande horizontale orange représente l'échantillonnage de notre étude, les barres verticales représentent les différentes régions considérées dans les études précédentes (Vert (Kamau *et al.*, 2007) ; Rose (Ruggiero *et al.*, 2008) et Mauve (Roux *et al.*, 2013)). La zone grisée représente le locus S, les positions sont calculées à partir du codon start du gène flanquant le plus proche.

Capture de séquence

De façon à décrire le polymorphisme des régions flanquantes du locus S, nous avons d'abord mis en place une approche par capture de séquences. La procédure expérimentale est décrite dans le chapitre 2. Brièvement, l'échantillonnage consiste en 72 individus d'*A. halleri* d'origines géographiques variées, sur lesquels nous avons construit des banques génomiques de courts fragments (350pb). Ces banques génomiques ont été hybridées à un ensemble de sondes synthétisées à partir de la séquence des régions flanquantes du locus S (deux clones BAC de 117 503 et 87 962 pb du côté Ubox et ARK3, respectivement). Des sondes issues des séquences de 100 régions génomiques de 25kb choisies aléatoirement comme références, dont la densité en gènes et en éléments transposables était proche (+/- 10%) de celle des régions flanquantes. Ces deux ensembles de sondes s'ajoutent à ceux détaillés dans le chapitre 2 et consistant en l'ensemble des séquences du locus S des différents allèles connus (voir chapitre 2).

Analyse du polymorphisme

La détection des variants a été réalisée via le même pipeline que dans le chapitre 2, à l'exception que nous avons considéré ici qu'il s'agissait de génotypes diploïdes. Nous avons calculé le polymorphisme total, synonyme et non-synonyme pour l'ensemble des deux régions flanquantes, puis spécifiquement pour les 10kb les plus proches du locus S. Les intervalles de

confiance à 95% ont été obtenus par une approche de bootstrap identique au chapitre 2 en rééchantillonnant les SNPs de chacun de ces ensembles. Pour les régions de contrôle, le polymorphisme total ainsi que l'intervalle de confiance lié ont été obtenus sur l'ensemble des séquences, alors que pour le polymorphisme synonyme et non synonyme, nous avons traité chaque région (ou scaffold) de manière indépendante, obtenant ainsi 100 valeurs π_S et π_N ainsi que le ratio π_N/π_S associé sur lesquelles nous avons calculé la moyenne et les intervalles de confiance représentant 95% de la distribution de ces 100 valeurs. Deux distributions ont été considérées comme distinctes lorsque les 95% des distributions ne se recouvrent pas. Les analyses de polymorphisme par fenêtres coulissantes ont été effectuées via VCFTools (Danecek *et al.*, 2011) avec une fenêtre de 10 000 nucléotides pour un pas de 1 nucléotide en excluant les fenêtres incluant des fragments annotés comme éléments transposables. Pour déterminer si les gènes proches du locus S se caractérisent par une moindre efficacité de la sélection purifiante détectable par la ségrégation d'un plus grand nombre de polymorphismes non-synonymes, nous avons comparé les valeurs de π_N/π_S de ces gènes à celles des gènes se trouvant sur les régions de contrôle d'une part et en les comparant entre eux selon leur distance au locus S d'autre part.

Clones BACs

Pour documenter de façon directe la structure haplotypique de la région du locus S sans avoir à recourir à des approches computationnelles indirectes, nous avons ensuite construit un ensemble de six banques BAC, chacune constituée à partir d'un mélange de jeunes tissus foliaires prélevés sur des individus distincts issus de graines d'une population naturelle différente (environ n=30 par population). Trois banques ont été construites à partir de populations d'*A. halleri,* dont une du nord de la France (Mortagne) et deux populations Italiennes (I9 et I13), et deux ont été construites à partir de populations d'*A. lyrata* provenant d'Islande (Ollver et ICE16, Figure 31).

La construction des banques elle-même a suivi Goubet et al. (2012) et a été réalisée au CNRGV (INRA Toulouse). Brièvement, le principe consiste à isoler de l'ADN de haut poids moléculaire, à le fragmenter en grands fragments (environ 100kb) qui sont ensuite intégrés dans des vecteurs bactériens. Les colonies bactériennes porteuses d'inserts ont été repiquées sur

plaque puis disposées sur membrane. Elles ont alors été hybridées par quatre sondes radioactives correspondant à deux fragments de chacun des deux gènes flanquants du locus S (B80/*U-box* d'une part et *ARK3* d'autre part). Les clones positifs identifiés ont alors été validés par PCR en utilisant les amorces utilisées pour la production des sondes. Nous avons considéré comme positifs l'ensemble des clones porteurs des deux gènes flanquants de chacun des deux côtés du locus S, ce qui correspond à trois configurations possibles : 1) les clones comprenant l'intégralité du locus S (possédant les portions amont et aval) 2) les clones ne comprenant que la portion amont (côté B80/*U-box*) et 3) les clones ne comprenant que la portion aval (côté *ARK3*, Figure 32). Les clones validés ont été intégralement séquencés par PACBIO et assemblés par le CNRGV, selon Bachmann et al. (2018). Chaque séquence obtenue est de ce fait une séquence complète issue d'une molécule unique d'ADN, dont l'association entre le locus S et les régions flanquantes est conservée.

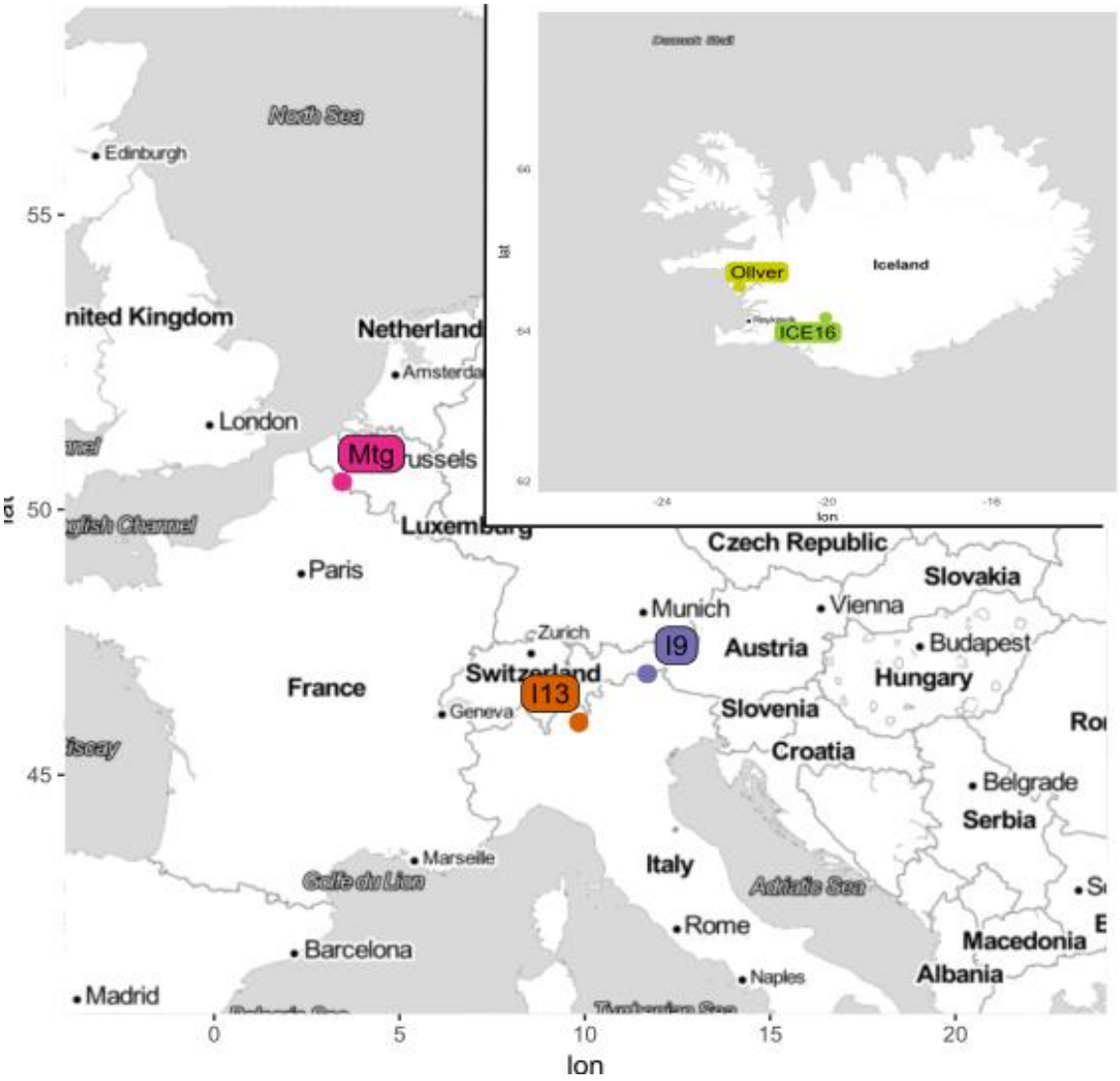


Figure 31: Localisation des populations ayant servi pour l'élaboration des banques BACs. Le cadre représentant l'Islande possède sa propre échelle.

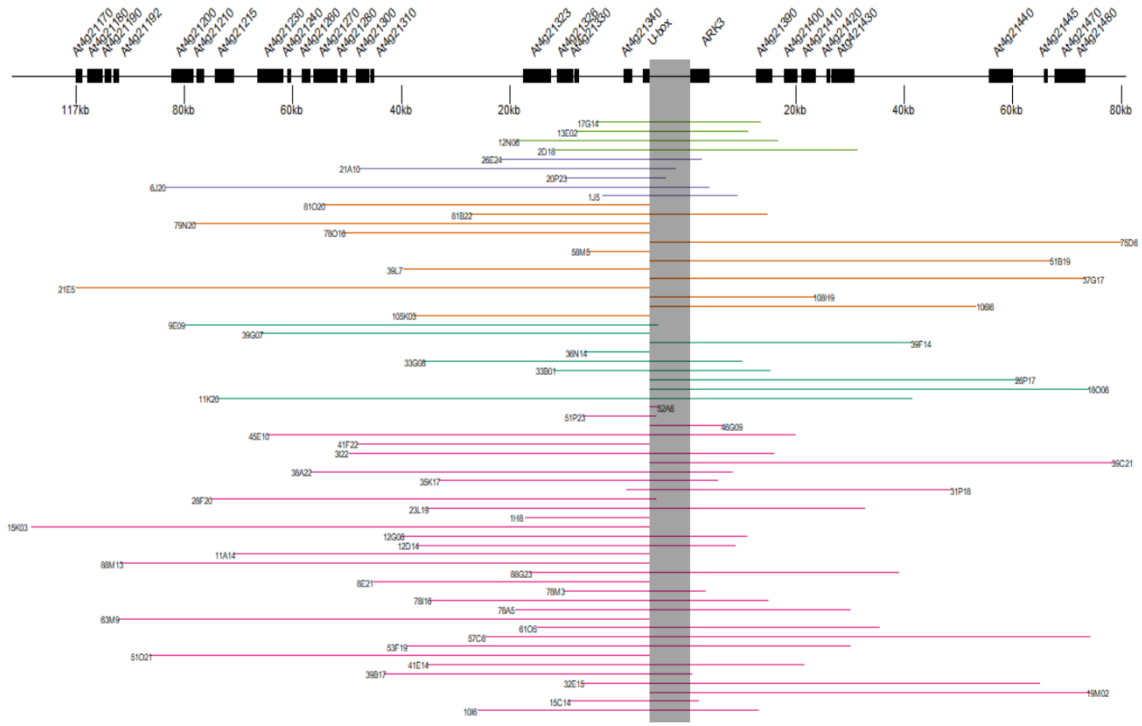


Figure 32: Disposition physique des clones BACs le long d'une séquence de référence. La zone grisée représente le locus S, dont la taille varie d'un haplotype à l'autre. Le codon start de chaque gène flanquant a servi comme point de valeur 0 pour calculer les distances relatives au locus S de chaque clone. Les barres horizontales représentent chacune un clone, dont la couleur dépend de la population d'origine *A. halleri* (Rose : Mortagne, Vert : I9, orange : I13) et *A. lyrata* (Mauve : ICE16 ; Vert clair : Öllver).

Annotation

Les séquences codant pour des gènes ont été détectées *via* Fgenesh (Solovyev *et al.*, 2006), puis nous les avons identifiées par blast des protéines qu'ils codent sur NCBI sur le génome *A. thaliana*. Pour identifier l'allèle S correspondant à chaque clone, nous avons recherché les gènes *SRK* ou *SCR* et lorsqu'ils étaient présents nous avons comparé leur séquence à notre base de données de séquences d'*A. halleri* ou *A. lyrata*. En l'absence de ces deux gènes, nous avons à ce stade considéré l'identité de l'allèle S comme indéfinie.

Description de la structure haplotypique.

L'étendue des blocs haplotypiques associés à chaque allèle S est déterminée par les échelles de temps relatives de la coalescence au sein de chaque lignée allélique d'une part et de la recombinaison d'autre part. La migration entre populations étant particulièrement efficace pour le locus S (Schierup *et al.*, 2000), on s'attend a priori à ce que la structure par populations ne soit pas un niveau majeur de l'organisation de la diversité des SNPs flanquants (Kamau *et al.*, 2007). Etant donnés la disposition variable des clones de part et d'autre du locus S (Figure 32) et l'étroitesse du pic de polymorphisme décrit par Roux *et al.*, (2013), nous nous sommes concentrés sur les CDS des cinq gènes flanquants en 3′ et 5′ du locus S pour chacun des clones d'*A. halleri et A. lyrata*, en conservant l'information de l'association avec leur allèle S respectif et de leur population d'origine. Les séquences orthologues de chaque gène chez *A. thaliana* ont été récupérées comme groupe externe. Ces séquences ont ensuite été nettoyées et alignées par MUSCLE (Edgar, 2004a,b) en utilisant le logiciel MEGA (Kumar *et al.*, 2016). Les arbres phylogénétiques ont été obtenus en appliquant la méthode du maximum de vraisemblance implémentée dans MEGA. Nous avons dans un premier temps concaténé les CDS de tous les gènes pour représenter la structure haplotypique globale de la région, puis avons ensuite traité chaque gène séparément pour examiner si la distance au locus S affectait la topologie obtenue.

Résultats

Approche par capture de séquences.

Un pic de polymorphisme d'étendue restreinte

L'approche par capture de séquence nous permet d'avoir une précision à l'échelle du nucléotide pour chaque région flanquante du locus S, nous permettant également d'obtenir des données sur les régions intergéniques. Pour les régions contrôles, nous avons obtenu un total de 1 695 198 positions ayant une couverture supérieure à 10X. Parmi ces positions, nous avons pu identifier 63 893 sites variables pour une valeur de polymorphisme moyen de 0.005. Nous avons identifié 6 321 sites variables parmi les 88 424 positions à X=10 de la région 5' (π= 0.007), et 6 407 sites variables parmi les 60 299 positions répondant au même critère pour la région 3' (π=0.01). L'analyse par fenêtre coulissante du polymorphisme total sur ces régions semble indiquer une augmentation du polymorphisme aux abords du locus S, et ce pour chaque côté (Figure 33 pour la région 5' et Figure 34 pour la région 3').

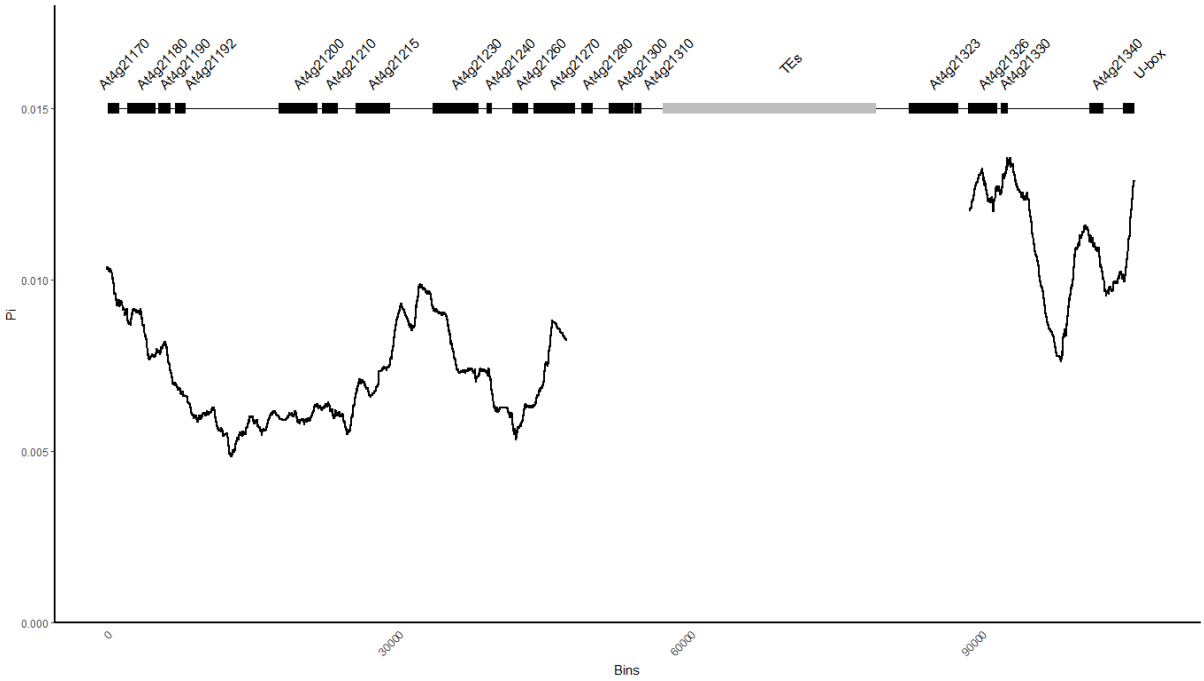


Figure 33: Analyse du polymorphisme de la région flanquante en 5' du locus S par fenêtres coulissantes de 10kb avec un pas de 1 nucléotide. Les gènes sont représentés sur la partie supérieure. La zone entre 57 307 et 79 246pb (ainsi que les fenêtres la chevauchant) a été exclue de la capture et/ou de l'analyse à cause de la présence d'un élément transposable.

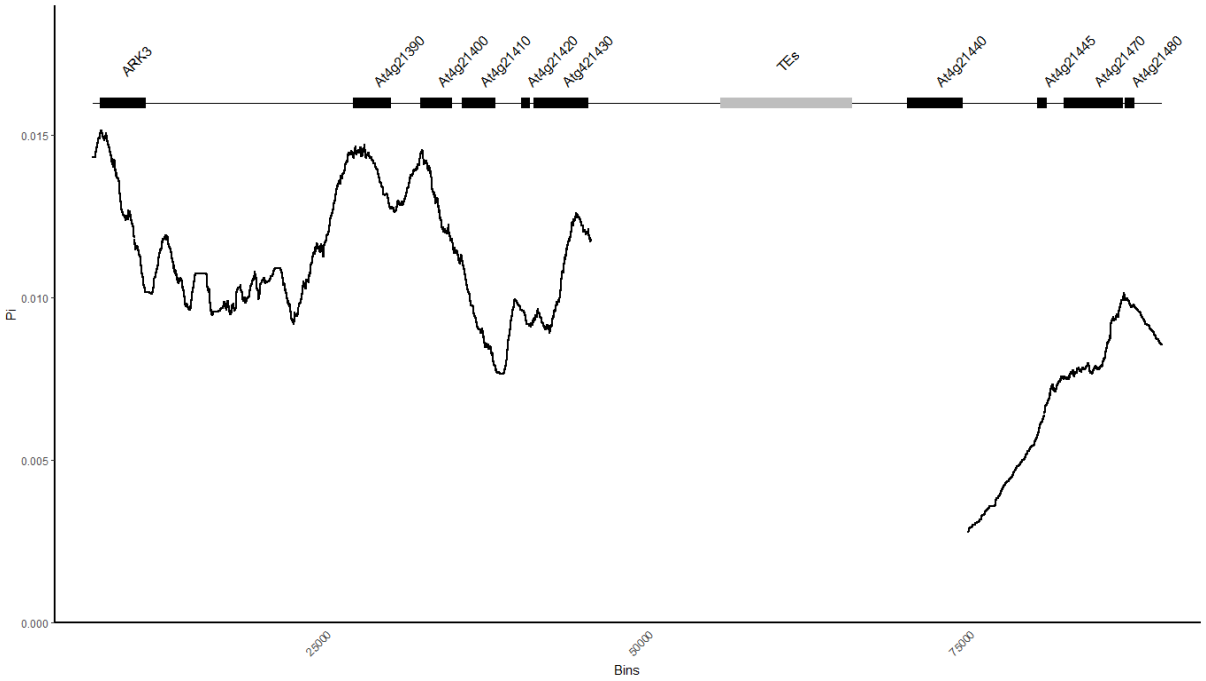


Figure 34: Analyse du polymorphisme de la région flanquante en 3' du locus S par fenêtre coulissante de 10kb avec un pas de 1 nucléotide. Les gènes sont représentés sur la partie supérieure. La zone entre 55 713 et 66 000pb (ainsi que les fenêtres la chevauchant) a été exclue de la capture et/ou de l'analyse à cause de la présence d'un élément transposable.

Pour tester formellement l'augmentation du polymorphisme aux abords du locus S, nous avons comparé par bootstrap sur les sites nucléotidiques le polymorphisme des régions de contrôle à celui des régions flanquantes entières, puis à celui des 10kb les plus proches du locus S pour chaque région flanquante. On observe d'une part que le polymorphisme des régions flanquantes s'écarte de celui des régions de contrôle pour les deux côtés du locus S (5' et 3') à une échelle de 10kb. Cependant, à une échelle plus large, cet écart n'est plus détectable pour la région flanquante 5'. A l'inverse, l'écart reste significatif du côté 3', indiquant que le retour vers des valeurs « basales » de polymorphisme est asymétrique entre les deux côtés. Ces deux régions les plus proches présentent effectivement des valeurs de polymorphisme total significativement plus élevées que celles des régions complètes (3' π_{10kb}=0.016 et 5' π_{10kb}=0.013, Figure 35).

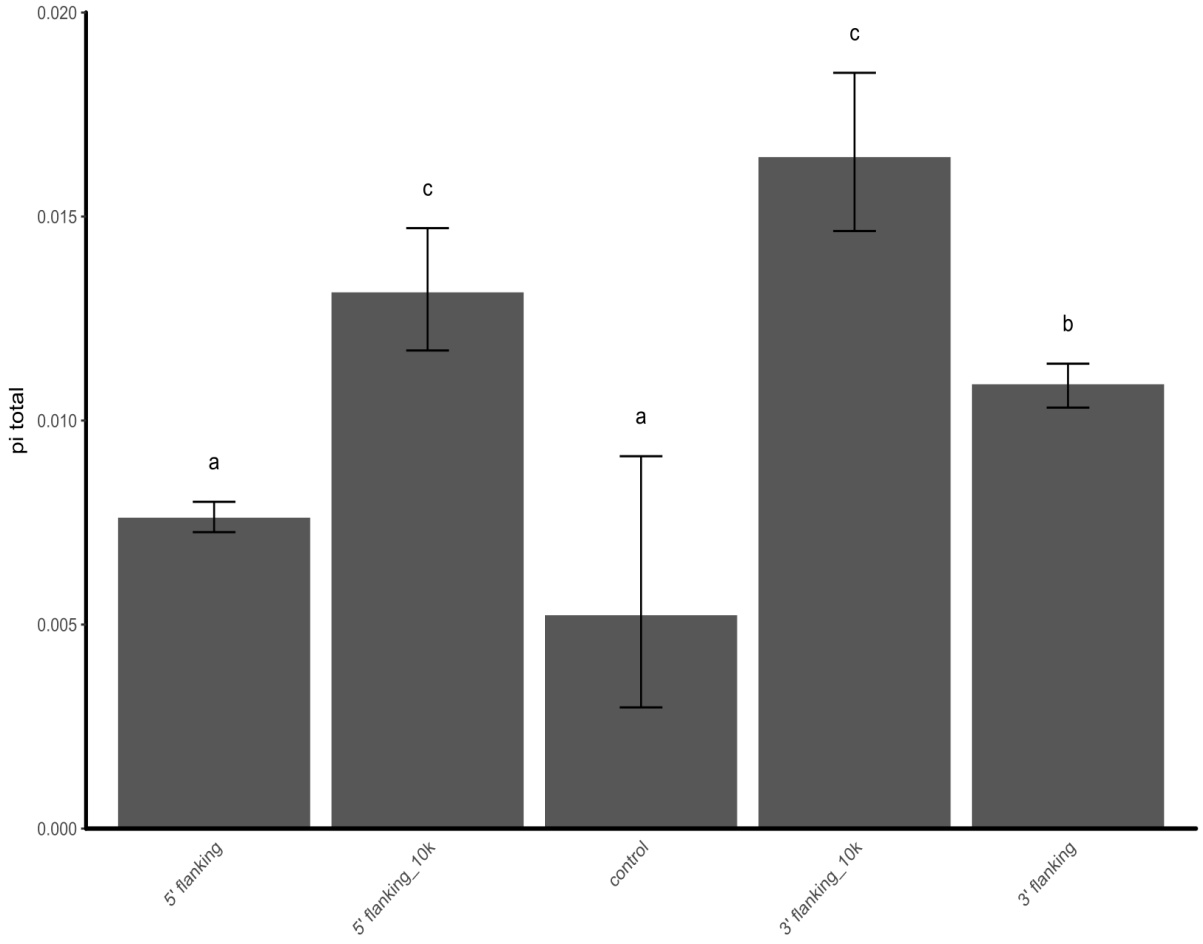


Figure 35: histogramme des valeurs de polymorphisme total pour les régions contrôles, les deux régions flanquantes 5' et 3' ainsi que les 10kb les plus proches du locus S pour ces deux régions. Les barres d'erreur représentent les intervalles de confiance à 95% obtenus par bootstrap sur les sites.

Un patron très similaire est observé lorsqu'on restreint l'analyse aux seules régions intergéniques, pour lesquelles on constate une différence significative entre les 10kb les plus proches du locus S et le reste des régions flanquantes (5'π=0.006 vs 5' π_{10kb}=0.011 et 3' π=0.010 vs 3' π_{10kb}=0.013). Ces valeur étant elles-mêmes significativement différentes du polymorphisme des régions intergéniques mesuré sur les régions de contrôle ($\pi_{control}$= 0,004 ; 5' π_{10kb}=0.011 ; 3' π_{10kb}= 0.013, Figure 36). La valeur de polymorphisme de la totalité de la région 3' est également significativement différente de celle de nos régions de contrôles, ce qui n'est pas le cas pour la région 5'.

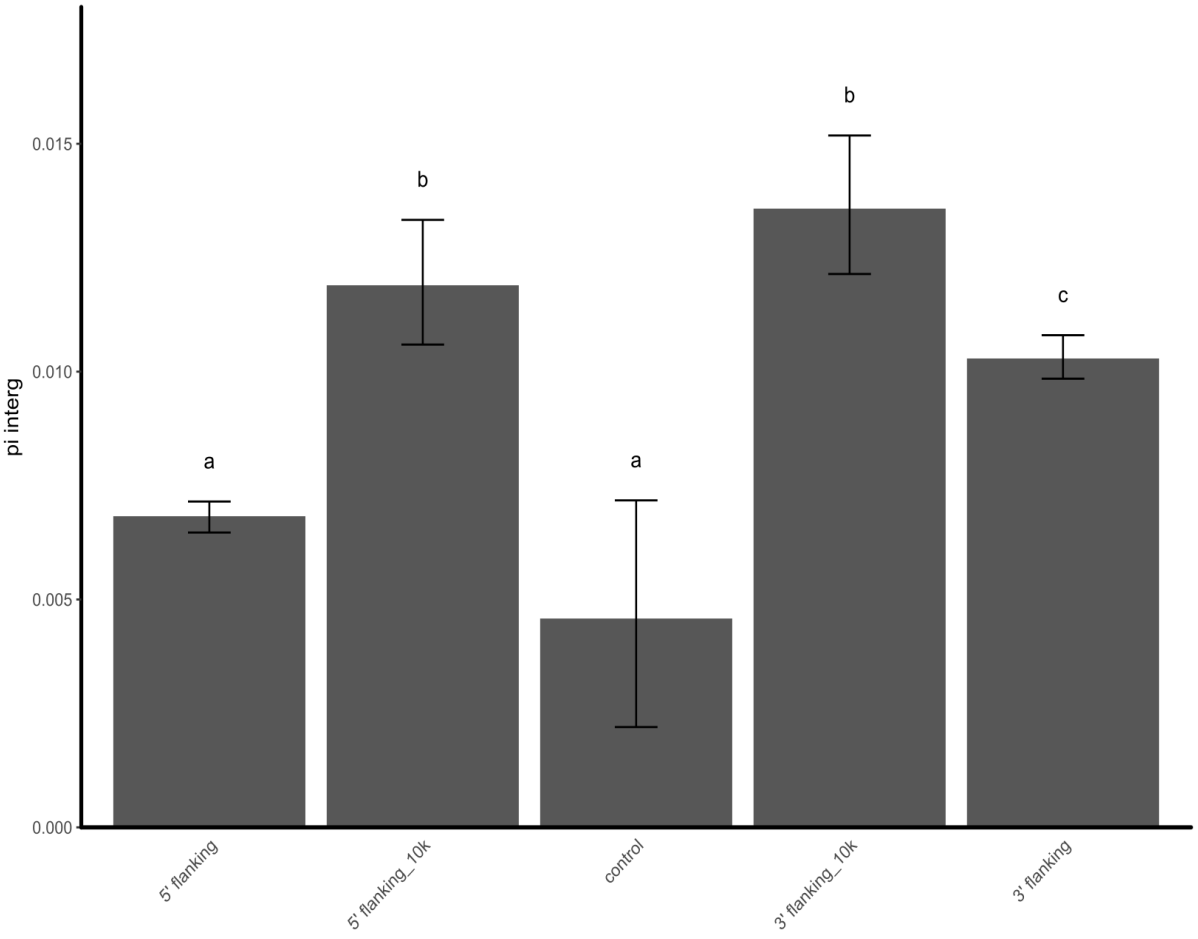


Figure 36: histogramme des valeurs de polymorphisme des régions intergéniques pour les régions contrôles, les deux régions flanquantes 5' et 3' ainsi que les 10kb les plus proches du locus S pour ces deux régions. Les barres d'erreur représentent les intervalles de confiance à 95% obtenus par bootstrap sur les sites.

Globalement, nos résultats confirment donc une augmentation du polymorphisme aux abords du locus S, en accord avec les attendus (Charlesworth, 2006; Kamau *et al.*, 2007; Roux *et al.*, 2013; DeGiorgio *et al.*, 2014), et cette augmentation est au moins détectable dans les 10 kb autour du locus S, avant de retourner très rapidement à des niveaux basaux de polymorphisme. Cette augmentation présente par ailleurs une asymétrie, et est plus nette du côté 3' (*U-box*) que du côté 5' (*ARK3*).

Détection du fardeau lié

L'analyse du polymorphisme synonyme sur les gènes proches du locus S confirme également l'augmentation de polymorphisme observé précédemment. On constate que *U-box*, *ARK3* et *At4g21330* possèdent des valeurs de π_S plus élevées que pour les régions contrôles (0.10 ; 0.09 ; 0.08 et 0.02, respectivement) alors que sur leur intégralité, les deux régions flanquantes ne se distinguent pas significativement des régions contrôles (Figure 37).

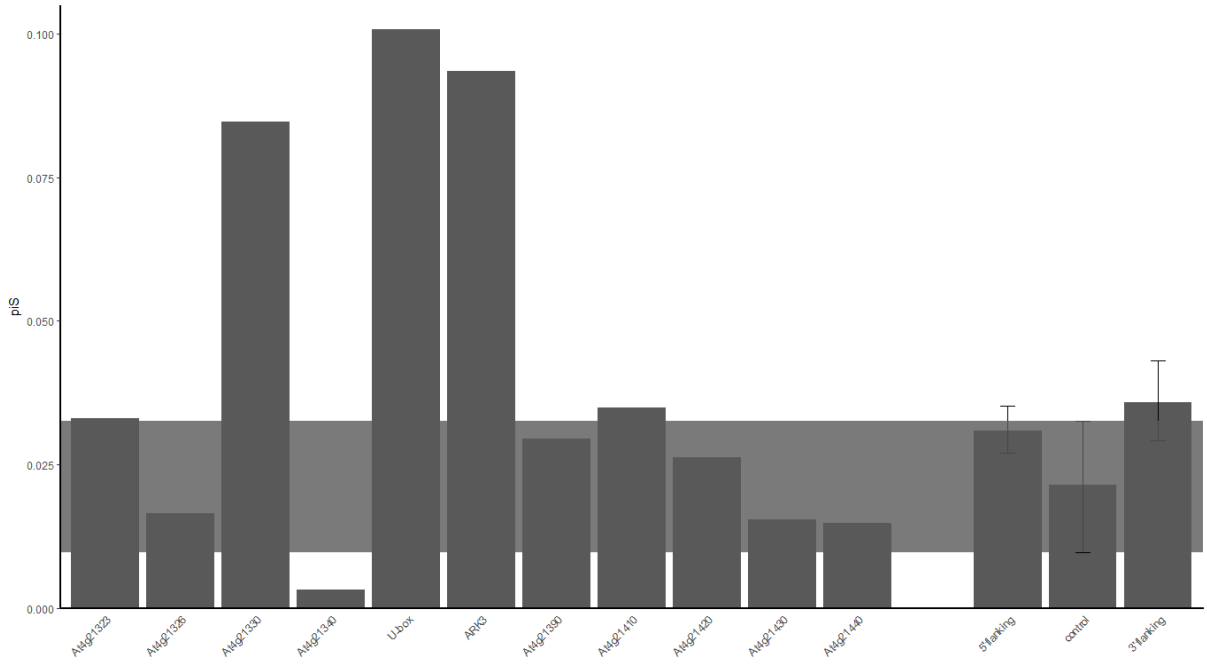


Figure 37: Histogramme des valeurs de polymorphisme synonyme pour les gènes les plus proches du locus S pour chaque région flanquante et des régions de contrôles, ainsi que la valeur moyenne de π_S pour l'intégralité des régions flanquantes. Les barres verticales représentent les intervalles de confiance à 95%, la zone grisée représente l'intervalle de confiance à 95% des valeurs π_S pour les 100 régions contrôles.

A l'inverse, le polymorphisme non-synonyme est de manière générale significativement plus élevée sur l'ensemble des régions flanquantes par rapport aux régions de contrôle (région 5' : 0.011 ; région 3' : 0.014 vs 0.004 pour les régions contrôle, Figure 38). En effet, la plupart des gènes possèdent une valeur de π_N significativement plus élevée que pour les régions contrôles. On y retrouve les trois gènes cités précédemment (*U-box*, *ARK3* et *At4g21330* avec

respectivement un π_N de 0.012, 0.027 et 0.011) mais également les gènes *At4g21390* et *At4g21410* pour des π_N de 0.014 et 0.022 respectivement (Figure 38).

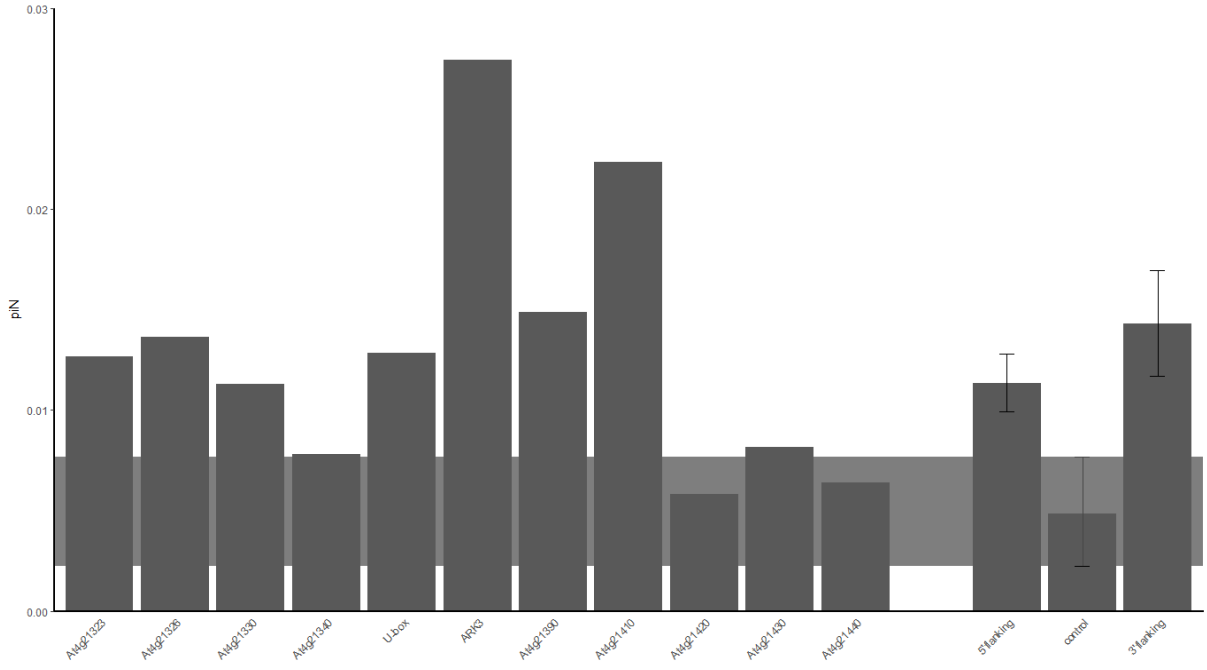


Figure 38: Histogramme des valeurs de polymorphisme non-synonyme pour les gènes les plus proches du locus S pour chaque région flanquante et des régions de contrôles, ainsi que la valeur moyenne de π_N pour l'intégralité des régions flanquantes. Les barres verticales représentent les intervalles de confiance à 95%, la zone grisée représente l'intervalle de confiance à 95% des valeurs π_N pour les 100 régions contrôles.

Ces résultats suggèrent que l'augmentation du niveau de polymorphisme observé aux abords du locus S concerne aussi bien le polymorphisme synonyme que non-synonyme, et que cette augmentation reste limitée aux régions les plus proches pour le polymorphisme synonyme, mais et plus marquée et s'étend plus largement pour le polymorphisme non-synonyme.

Toutefois, cette augmentation générale de polymorphisme ne semble pas indiquer une diminution de l'intensité de la sélection aux abords du locus S. En effet, le ratio de π_N/π_S pour les gènes flanquant *U-box* et *ARK3* n'est significativement pas plus élevé que pour les régions de contrôle (0.013 et 0.029 vs 0.022 respectivement, l'intervalle de confiance supérieur étant de 0.036). Le gène *At4g21340* possède quant à lui un ratio très différent des autres gènes de cette région avec une valeur de 2.40, majoritairement expliqué par le très faible niveau de π_S

(0.003, soit environ 10 fois moins que le reste de la région flanquante 5') en comparaison avec sa valeur de π_N (0.007). Il ne semble pas y avoir de corrélation entre ratio de π_N/π_S et la proximité du locus S (π_N/π_S = 0.367 et 0.398 pour les régions 5' et 3' respectivement, contre une limite supérieur de 0.361 pour les régions contrôles, Figure 39).

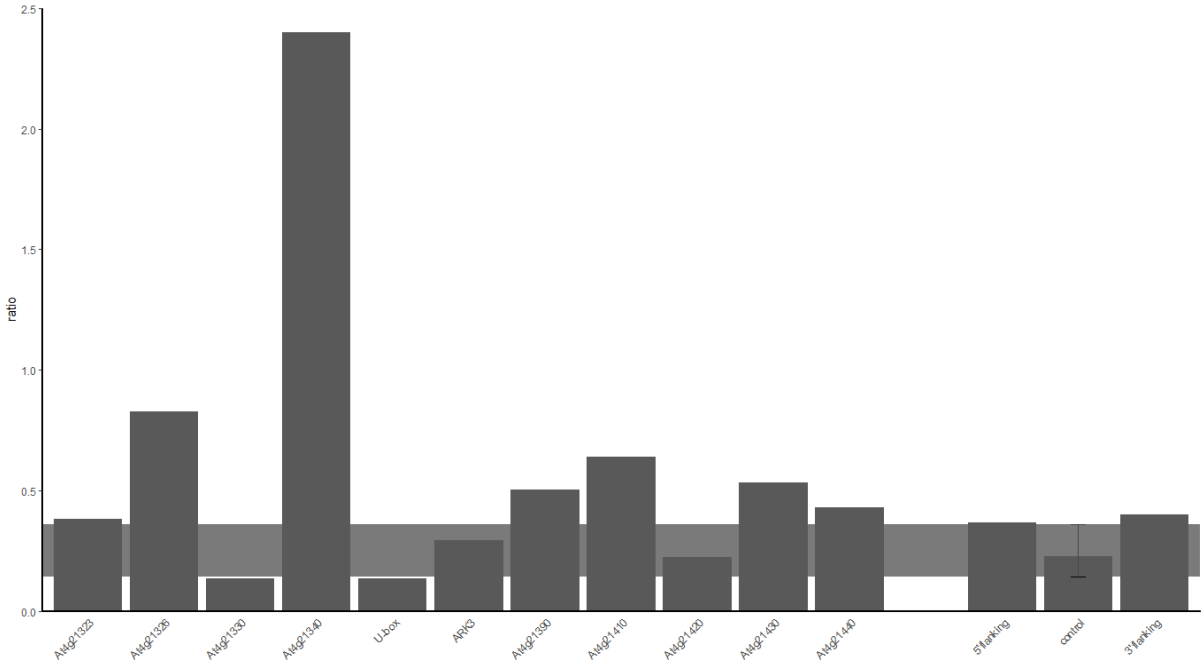


Figure 39: Histogramme des ratios de polymorphisme non-synonyme sur synonyme pour les gènes les plus proches du locus S pour chaque région flanquante et des régions de contrôles, ainsi que la valeur moyenne de π_N/π_S pour l'intégralité des régions flanquante. Les barres verticales et la zone grisée représente l'intervalle de confiance à 95% des valeurs π_N/π_S pour les 100 régions contrôles.

A ce stade, nous ne sommes pas en mesure de connaître l'impact exact de ces mutations non-synonymes en ségrégation, que seule une étude approfondie sur la fonction des gènes pourrait déterminer. Il est cependant frappant de constater que le pic de polymorphisme est restreint en taille, ne concerne qu'une poignée de gènes, et qu'il ne semble pas associé à une diminution de l'intensité de la sélection purifiante, les polymorphismes synonymes et non-synonymes s'accumulant à un taux quasi proportionnel.

Structure haplotypique.

Pour déterminer comment la liaison au locus S distord les patrons de polymorphisme des régions qui lui sont associées, nous avons caractérisé en détail la structure haplotypique des SNPs dans les régions flanquantes en construisant et exploitant un ensemble de banques BACs dédiées. A l'issue des étapes de construction des banques BAC, de screening et de séquençage des clones positifs, nous avons obtenu un total de 72 séquences intégrales, issues des populations Mortagne (Mtg, n=35), I9 (n=13) et I13 (n=10). Les deux populations d'*A. lyrata* provenant d'Islande sont représentées par n=4 (Ollver) et n=5 séquences (ICE16). Les séquences obtenues ont une longueur moyenne de 97 516 pb Figure 40. Les variations du nombre de séquences obtenus entre les différentes banques reflètent les variations de nombres de clones obtenus, cette étape étant la plus limitante au niveau technique.

Génotype des clones BACs.

Nous avons été en mesure d'identifier l'allèle S de 52 des 72 clones obtenus, les 20 derniers ne présentant pas de séquence des gènes *SRK* ou *SCR* nous permettant de les attribuer à un allèle S spécifique. En accord avec la très grande diversité allélique en populations naturelles, la plupart des séquences portent des allèles présents en fréquence faible. L'allèle le plus fréquent dans notre échantillonnage est l'allèle Ah03 (14 copies), suivi de Ah01 (5 copies), puis de Ah24 et Ah25 (4 copies chacun), puis Ah12 (3 copies), et enfin Ah04 et Al01 (2 copies

chacun). Les autres allèles sont présents en unique copie et ne permettent donc pas de déterminer si les SNPs flanquants qu'ils portent leur sont spécifiques (Figure 40).

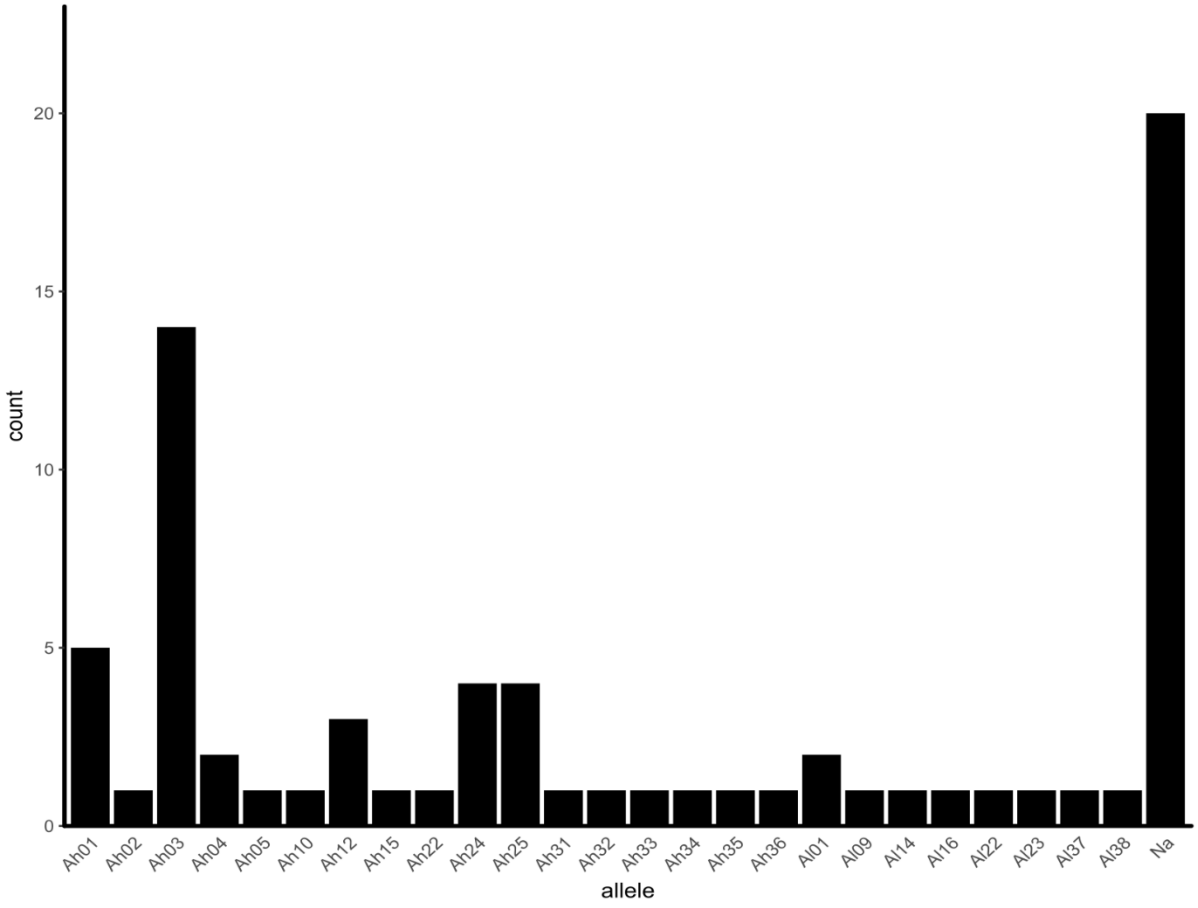


Figure 40: Distribution des fréquences des allèles pour les 72 clones BACs analysés. Nous avons beaucoup d'allèles pour lesquels une seule copie est présente. Nous n'avons pas été en mesure d'identifier le génotype S de 20 clones.

Les deux premiers allèles (Ah03 et Ah01) sont chacun présents dans deux populations différentes (I13, et Mtg ; I9 et Mtg, respectivement). Spécifiquement, nous avons obtenu 9 copies de l'allèle Ah03 dans la population I9, et 5 dans la population Mtg. Bien que les fréquences des allèles S soient en général fortement négativement corrélées à leur niveau de dominance à l'échelle de l'espèce, l'allèle Ah03 est le plus fréquent dans notre échantillonnage, en accord avec la forte représentation des clones de la population I9, au sein de laquelle cet allèle est particulièrement abondant.

Analyse de la structure haplotypique.

Nous avons dans un premier temps considéré les clones comme des haplotypes entiers, en concaténant l'ensemble des séquences codantes de chaque clone et en établissant une phylogénie de ces séquences concaténées. Cette première analyse globale révèle d'une part un groupement des haplotypes par espèces, les séquences d'*A. lyrata* formant un groupe de séquences monophylétique. On observe ensuite globalement une structure par populations, l'ensemble des haplotypes tendant à se regrouper par population d'origine plutôt que par allèle S auquel ils sont associés (Figure 18). Ainsi, les haplotypes associés aux différentes copies de l'allèle Ah03 de la population I9 se regroupent avec les haplotypes associés à d'autres allèles S de cette même population (et des autres populations italiennes), plutôt qu'avec ceux associés aux autres copies de l'allèle Ah03 de la population de Mortagne. De la même façon, les copies de l'allèle Ah01 de la population I13 se regroupent avec d'autres haplotypes de la population I13 plutôt qu'avec les haplotypes associés à l'allèle Ah01 de la population Mortagne. A l'échelle de l'espèce les SNPs associés aux allèles S sont donc globalement structurés par population d'origine plutôt que par allèle S, en contradiction avec les résultats obtenus par (Kamau *et al.*, 2007) chez *A. lyrata*. Au sein des populations, on observe à l'inverse une très faible divergence entre haplotypes associés à un même allèle S, suggérant un très faible niveau de polymorphisme à ce niveau d'organisation.

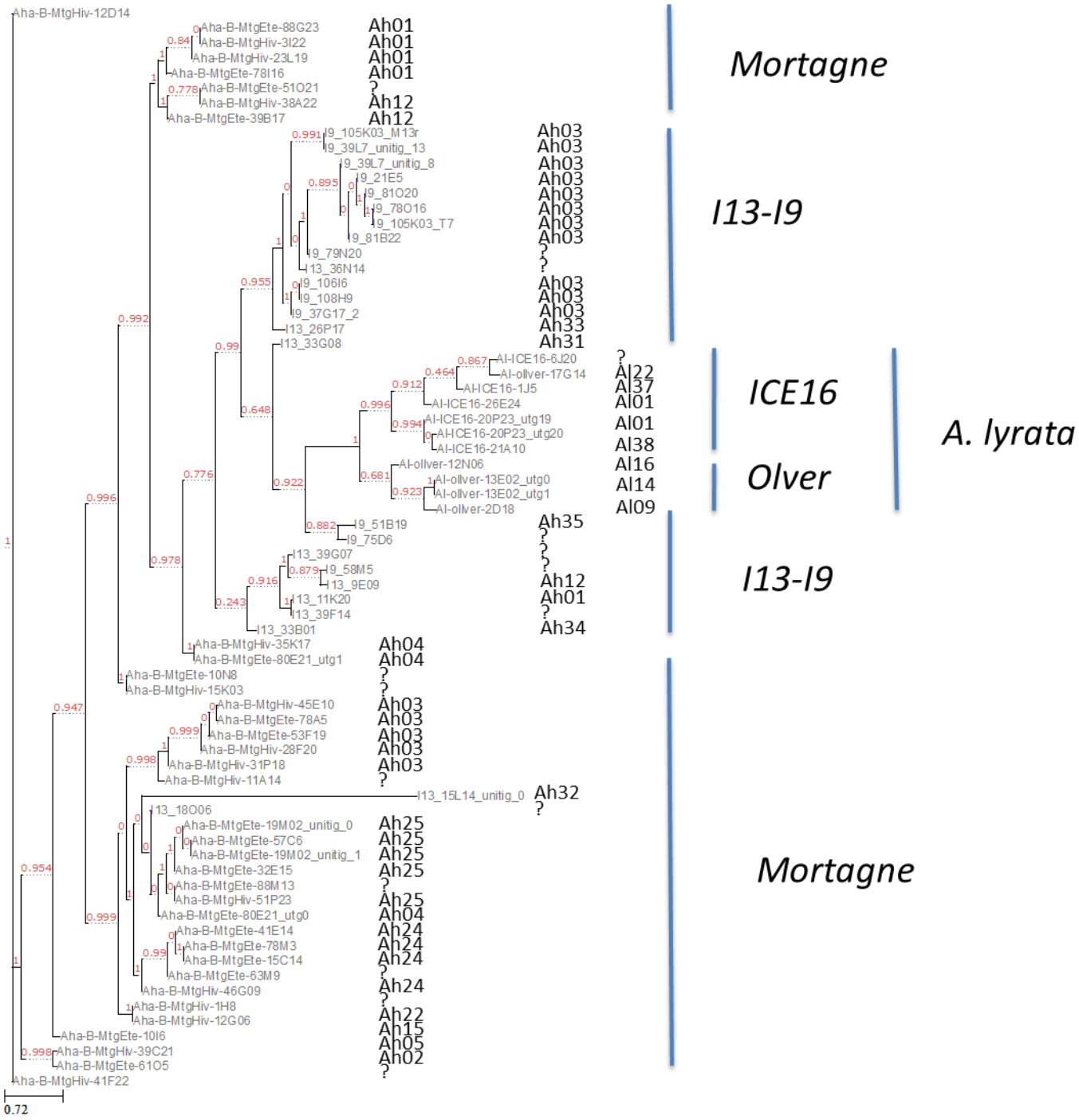


Figure 41: Arbre phylogénique représentant l'ensemble des séquences concaténées des gènes échantillonnés autour des deux côté du locus S. Les valeurs de bootstrap sont indiquées en rouge.

Nous avons dans un second temps examiné en détail la structure de la phylogénie pour chacun des gènes flanquants. Le gène *U-box* (immédiatement flanquant le locus S) montre déjà une

structure par populations plutôt que par allèles (Figure S9), indiquant que de ce côté l'association est immédiatement rompue à l'échelle de l'espèce. A l'inverse cependant, on observe que les allèles du gène *ARK3* se regroupent selon l'allèle S auquel ils sont liés plutôt que par population d'origine (Figure S10). Ce signal est partiellement préservé pour le gène suivant (*At4g21390*, Figure S11). Seul *ARK3* présente donc un déséquilibre de liaison avec le locus S assez fort pour prédominer sur la structure par population, ce signal d'association étant déjà un peu plus diffus pour *At4g21390*, et complètement perdu pour les gènes suivants le long du chromosome.

A l'échelle intra-populationnelle, la longueur des haplotypes associés aux allèles S reste très importante. En effet, du côté 5', et sur une longueur d'au moins 15kb, l'association entre l'allèle S et les régions flanquantes est parfaitement conservée. Nous remarquons sur le gène *At4g21323*, que tous les allèles Ah03 de Mortagne sont encore regroupés ensembles, tandis que les allèles Ah03 des populations italiennes forment un autre groupe (Figure S12). Côté 3', au niveau du gène *At4g21430* situé en environ 30kb en aval du locus S, les allèles Ah03 de Mortagne forment un groupe, tandis que les allèles Ah01 de la même population forment un autre groupe distinct (Figure S13). Ceci signifie que l'association entre l'allèle S et la région flanquante est encore présente à cette distance. Au sein des populations, le déséquilibre de liaison entre le locus S et les régions flanquantes s'étend donc au moins sur une région de 45kb, sans compter la taille du locus S lui-même. Ceci indique un fort effet de l'échelle géographique à laquelle cette association est étudiée, qui n'avait pas été noté dans les études précédentes.

Discussion

L'étude de la diversité en lien avec un site sous sélection balancée présente plusieurs défis. Le haut niveau de polymorphisme attendu, l'accumulation d'éléments transposables et la présence de nombreux haplotypes très divergents rendent ces régions difficiles à séquencer. C'est pourquoi, jusqu'à maintenant, la plupart des résultats disponibles ont été obtenus sur des séquences partielles de gènes dont la plupart étaient relativement éloignés du locus S (Kamau *et al.*, 2007; Ruggiero *et al.*, 2008; Roux *et al.*, 2013). L'approche par capture de séquence s'est révélée puissante, nous permettant d'avoir accès aux séquences tout en y

capturant la diversité de manière exhaustive sans avoir à choisir des fragments de façon arbitraire. Grâce à cette approche, nous avons pu décrire de manière précise l'accumulation de polymorphisme autour du locus S, et aussi d'estimer l'ampleur du fardeau lié au locus S, et ce sur un échantillonnage issu de plusieurs populations naturelles. Une limite de cette approche est que contrairement aux clones BACs, l'association entre les SNPs et l'allèle S est perdue, de même que toute variation structurale des allèles-S et des régions flanquantes. C'est pourquoi les deux approches sont complémentaires et permettent d'étudier sous plusieurs aspects l'association entre le locus S et ses régions associées.

Longueur des haplotypes associés aux allèles S

Les modèles théoriques sur la sélection balancée prédisent que la diversité des sites neutres associés au locus S devrait être principalement structurée par les différences entre lignées alléliques et pas entre populations (Charlesworth *et al.*, 2003) et que différents allèles fonctionnels ségrégeant au sein des dèmes chez une espèce présentant des populations fragmentées devraient présenter une structuration par populations plus faible que des variant neutres (Schierup *et al.*, 2000; Muirhead, 2001). Tous ces modèles théoriques s'appliquent cependant au système gamétophytique, et la différence de fréquences entre allèles du système sporophytique rend ces prédictions plus difficiles à interpréter (Schierup *et al.*, 1998; Uyenoyama, 2000). Grâce à l'analyse des clones BACs, nous avons pu mettre en évidence que, chez *A. halleri,* la variabilité présente aux régions liées était avant tout structurée par populations mais que l'association entre le locus S et les variant des gènes liés était visible au sein des populations et pouvait s'étendre sur une région faisant au moins 66kb de long, sans compter le locus S, dont la taille varie fortement d'un allèle à l'autre (Goubet *et al.*, 2012). Ces résultats seraient en accord avec ceux obtenus chez *A. lyrata* (Kamau et *al.* 2007) si l'on considère que les populations Islandaises ne forment en fait qu'une seule et même population essentiellement continue.

Sélection balancée et niveau de diversité neutre

L'analyse du polymorphisme des séquences obtenues par capture montre clairement un lien entre niveau de polymorphisme et proximité du locus S (Figure 33 & Figure 34). La présence

du pic de polymorphisme est un attendu théorique, et il a été montré que la profondeur de la généalogie des gènes au sein des régions flanquantes liées à un locus sous sélection balancée augmentait le temps au cours duquel les événements de recombinaison pouvaient se produire (Schierup *et al.*, 2001), résultant en une région limitée au sein de laquelle le pic de polymorphisme est détectable. Les précédentes études empiriques ont suggéré que l'étendue de ce pic était limitée à quelques kb (Kamau *et al.*, 2007; Roux *et al.*, 2013), mais sans pouvoir décrire avec précision son étendue. Nos résultats montrent qu'il existe bien un pic de polymorphisme autour du locus S, et que celui-ci s'étend sur une distance qui ne dépasse pas 10kb de part et d'autre. Des résultats obtenus chez *A.thaliana*, ont également montré qu'il était possible de détecter un niveau de diversité plus élevé aux abords du gène RPS5 impliqué dans la résistance au pathogène et présentant des traces de régime de sélection balancée, mais que ce signal ne s'étendait pas au-delà de 10kb. Chez l'humain, les études visant à détecter la signature de sélection indirecte autour de locus sous sélection balancée ont montré la présence d'un pic restreint (Akey *et al.*, 2002; Bubb *et al.*, 2006; Andrés *et al.*, 2009). Ces résultats présentent un défi pour les approches visant à détecter les locus sous sélection balancée sur la base de la description du polymorphisme à l'échelle du génome.

Accumulation de polymorphisme non-synonyme et fardeau lié

Les régions soumises à des interférences de sélection avec un locus sous sélection balancée sont censées accumuler des mutations délétères récessives. L'excès en hétérozygotie peut aider à créer un fardeau abrité en protégeant les mutations délétères de la purge, comme suggéré pour la région MHC (van Oosterhout, 2009) ou le locus S (Llaurens *et al.*, 2009). La forte fréquence de maladie associée aux gènes *HLA* chez l'humain suggère l'accumulation de mutations délétères à ce locus sous sélection balancée (De Bakker *et al.*, 2006; Shiina *et al.*, 2006), tout comme les effets significatifs montrés sur des homozygotes issus de croissement forcés chez *A. halleri* (Llaurens *et al.*, 2009) et *A. lyrata* (Stift *et al.*, 2013) confirment l'existence de ce fardeau lié. Il existe chez l'oiseau *Philomachus pugnax* (le Combattant varié) trois morphes sexuels, dont l'un (appelé satellite) est préférentiellement sélectionné par les femelles. Or ce morphe est causé par une mutation récessive délétère. Il y a donc une forme de sélection balancée issue d'une sélection fréquence dépendante, où le morphe est

maintenu dans les populations par la sélection de femelles, mais qui ne peut pas se fixer car il est létal à l'état homozygote. L'étude de ce fardeau a permis de mettre en avant une possible inversion dans un gène essentiel codant pour une protéine du centromère (*CENP-N*) (Küpper *et al.*, 2015). La plupart du temps cependant, les bases génétiques de ce fardeau sont inconnues. Nous avons montré que le polymorphisme des régions à moins de 10kb du locus S était significativement impacté par la présence du locus S, les gènes s'y trouvant semblant accumuler de nombreuses mutations non-synonymes. Certains de ces polymorphismes pourraient constituer des candidats pour expliquer les effets de fardeau lié mesurés sur les plantes issue de croisements forcés chez *A. halleri* (Llaurens *et al.*, 2009). Toutefois, nos résultats indiquent que ces mutations non-synonymes s'accumulent à un taux proportionnel à celui des mutations synonymes, indiquant que la simple profondeur des généalogies plutôt qu'une diminution de l'efficacité de la sélection est responsable de ce patron. De la même façon, à l'échelle de l'espèce, l'association entre allèles S et variants liés est très rapidement perdue, n'étant essentiellement détectable qu'au sein des populations. Ceci indique que l'échelle de temps à laquelle ces mutations délétères se maintiennent est inférieure à celle de la migration des allèles S entre les différentes populations, de sorte que des allèles S de populations distantes telles que nous les avons analysées ici portent des mutations délétères liées essentiellement distinctes.

Une des limites de l'approche par capture de séquence est la perte de la phase entre les régions flanquantes et l'allèle S associé, ce qui ne nous permet pas de tester si les allèles dominants possèdent en moyenne un fardeau lié plus important que les allèles récessifs. Habituellement, une analyse de trio de parenté est nécessaire pour pouvoir phaser les haplotypes afin d'étudier ce phénomène. Cela dit, la forte association entre l'allèle S et son cortège de SNPs découverte au sein des populations via l'étude des clones BACs ouvre peut-être l'opportunité de phaser plus simplement les individus en analysant deux séquences d'un même allèle issu d'une même population. De façon à décrire l'accumulation du fardeau et son hétérogénéité entre allèles le long de la hiérarchie de dominance tel que prédit par Llaurens et al. (2009), il serait maintenant utile de concentrer les efforts de séquençage par capture de séquence sur un ensemble de populations plus proches, tel que par exemple le système

constitué des différentes populations italiennes dont seules deux ont été incluses dans la présente étude.

Bibliographie

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. **2002**. Interrogating a high-desity SNP map for signatures of natural selection. *Genome Research* **12**: 1805–1814.

Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, *et al.* 2009. Targets of balancing selection in the human genome. *Molecular Biology and Evolution* **26**: 2755–2764.

De Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, *et al.* 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics* **38**: 1166–1172.

Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, Palmieri A, Subramanian S, Zhou Y, Kaul R, *et al.* 2006. Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173**: 2165–2177.

Charlesworth D. **2006**. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2**: 379–384.

Charlesworth B, Charlesworth D, Barton H. **2003**. The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution, and Systematics* **34**: 99–125.

Charlesworth B, Morgan MT, Charlesworth D. **1993**. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.

Comeron JM, Williford A, Kliman RM. **2008**. The Hill-Robertson effect: Evolutionary consequences of weak selection and linkage in finite populations. *Heredity* **100**: 19–31.

Cutter AD, Payseur BA. **2013**. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Review Genetics* **14**: 262–274.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

DeGiorgio M, Lohmueller KE, Nielsen R. **2014**. A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genetics* **10**: e1004561.

Edgar RC. **2004a**. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

Edgar RC. **2004b**. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 1–19.

Glémin S, Bataillon T, Ronfort J, Mignot A, Olivieri I. **2001**. Inbreeding depression in small populations of self-incompatible plants. *Genetics* **159**: 1217–1229.

Goubet PM, Bergès H, Bellec A, Prat E, Helmstetter N, Mangenot S, Gallina S, Holl A-C, Fobis-Loisy I, Vekemans X, *et al.* 2012. Contrasted pattern of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. *PLoS genetics* **8**.

Hill WG, Robertson A. **1966**. The effect of linkage on limits to artificial selection. *Genetics Research* **89**: 268–294.

Kamau E, Charlesworth D. **2005**. Balancing selection and low recombination affect diversity near the self-incompatibility loci of the plant Arabidopsis lyrata. *Current Biology* **15**: 1773–1778.

Kamau E, Charlesworth B, Charlesworth D. **2007**. Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of Arabidopsis lyrata. *Genetics* **176**: 2357–2369.

Kern AD, Hahn MW. **2018**. The neutral theory in light of natural selection. *Molecular biology and evolution*: 1–6.

Kumar S, Stecher G, Tamura K. **2016**. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33**: 1870–1874.

Küpper C, Stocks M, Risse JE, Dos Remedios N, Farrell LL, McRae SB, Morgan TC, Karlionova N, Pinchuk P, Verkuil YI, *et al.* 2015. A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics* **48**: 79–83.

Llaurens V, Gonthier L, Billiard S. **2009**. The sheltered genetic load linked to the S locus in plants: New insights from theoretical and empirical approaches in sporophytic self-incompatibility. *Genetics* **183**: 1105–1118.

Maynard Smith J, Haigh J. **1974**. The hitch-hiking effect of a favourable gene. *Genetics Research* **23**: 23–35.

McVean GAT, Charlesworth B. **2000**. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**: 929–944.

Muirhead CA. **2001**. Consequences of population structure on genes under balancing selection. *Evolution* **55**: 1532–41.

NCBI. https://blast.ncbi.nlm.nih.gov/

van Oosterhout C. **2009**. A new theory of MHC evolution: beyond selection on the immune genes. *Proceedings of the Royal Society B: Biological Sciences* **276**: 657–665.

Roux C, Pauwels M, Ruggiero M-V, Charlesworth D, Castric V, Vekemans X. **2013**. Recent and Ancient Signature of Balancing Selection around the S-Locus in Arabidopsis halleri and A. lyrata. *Molecular Biology and Evolution* **30**: 435–447.

Ruggiero MV, Jacquemin B, Castric V, Vekemans X. **2008**. Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in Arabidopsis halleri. *Genetics Research* **90**.

Schierup MH, Mikkelsen AM, Hein J. **2001**. Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. *Genetics* **159**: 1833–1844.

Schierup MH, Vekemans X, Charlesworth D. **2000**. The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetical Research* **76**: 51–62.

Schierup MH, Vekemans X, Christiansen FB. **1998**. Allelic genealogies in sporophytic self-incompatibility systems in plants. *Genetics* **150**: 1187–1198.

Shiina T, Ota M, Shimizu S, Katsuyama Y, Hashimoto N, Takasu M, Anzai T, Kulski JK, Kikkawa E, Naruse T, *et al.* 2006. Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* **173**: 1555–1570.

Slotte T. **2014**. The impact of linked selection on plant genomic variation. *Briefings in Functional Genomics and Proteomics* **13**: 268–275.

Solovyev V, Kosarev P, Seledsov I, Vorobyev D. **2006**. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**: 10.1-10.12.

Stift M, Hunter BD, Shaw B, Adam A, Hoebe PN, Mable BK. **2013**. Inbreeding depression in self-incompatible North-American Arabidopsis lyrata: Disentangling genomic and S-locus-specific genetic load. *Heredity* **110**: 19–28.

Stone JL. **2004**. Sheltered load associated with S-alleles in Solanum carolinense. *Heredity* **92**: 335–342.

Uyenoyama MK. **2000**. Evolutionary dynamics of self-incompatibility alleles in Brassica. *Genetics* **156**: 351–359.

Uyenoyama MK. **2003**. Genealogy-dependent variation in viability among self-incompatibility genotypes. *Theoretical Population Biology* **63**: 281–293.

Uyenoyama MK. **2005**. Evolution under tight linkage to mating type. *New Phytologist* **165**: 63–70.

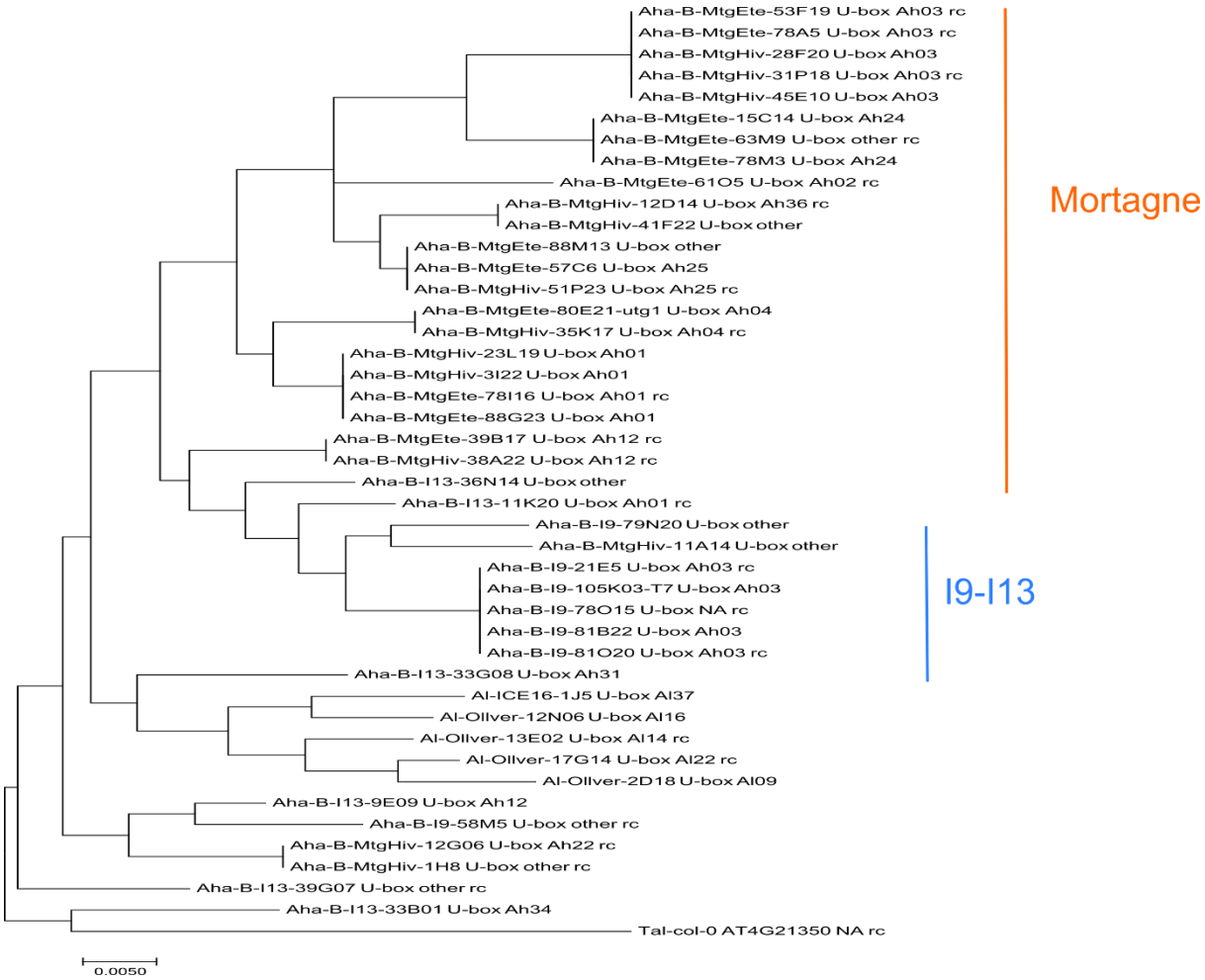


Figure S9: Phylogénie du gène *U-box* obtenue par maximum de vraisemblance. On constate une organisation par population, et ensuite par allèles au sein des populations.

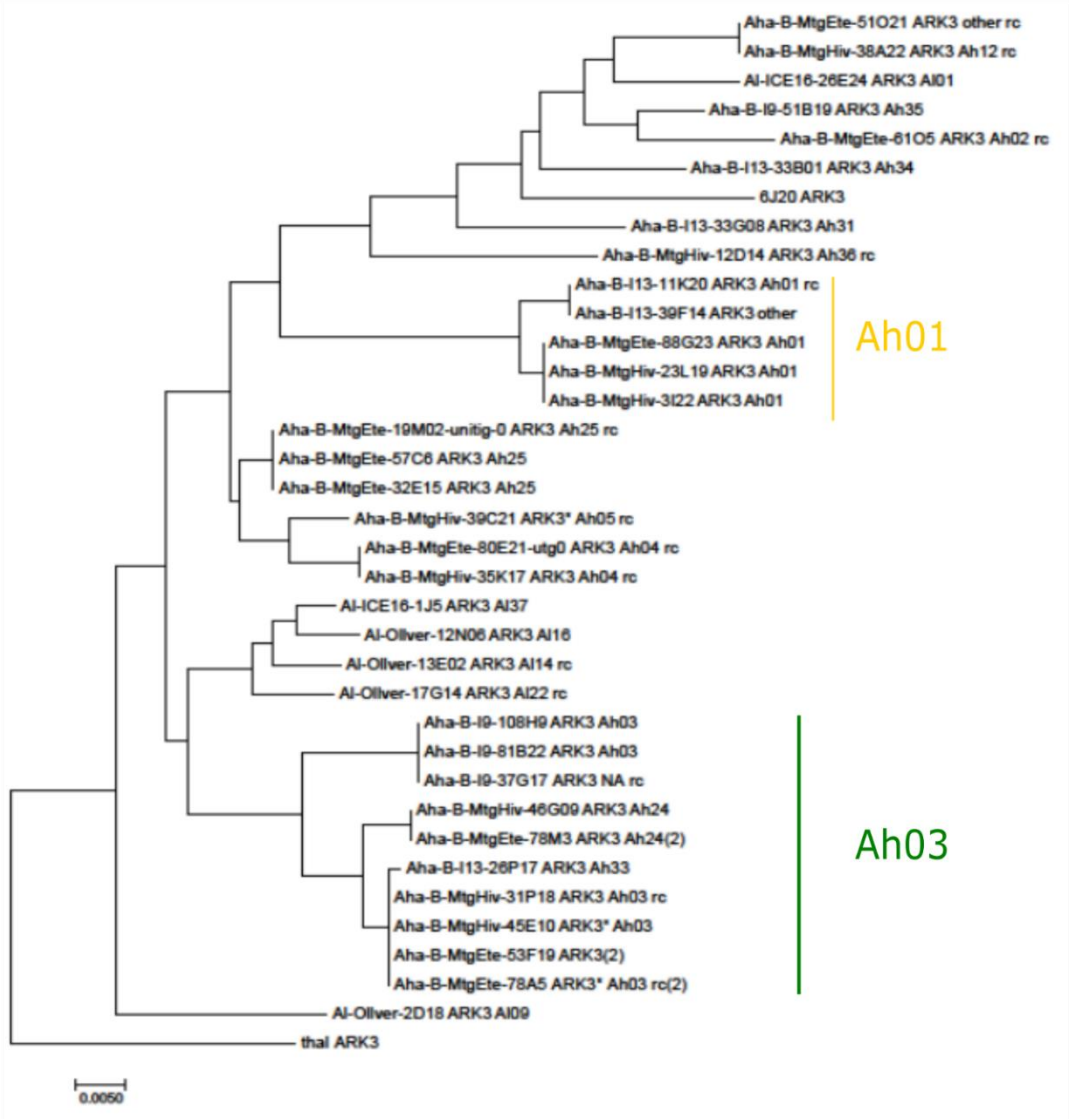


Figure S10: Phylogénie du gène *ARK3* obtenue par maximum de vraisemblance. On constate une organisation par allèle et ensuite par population.

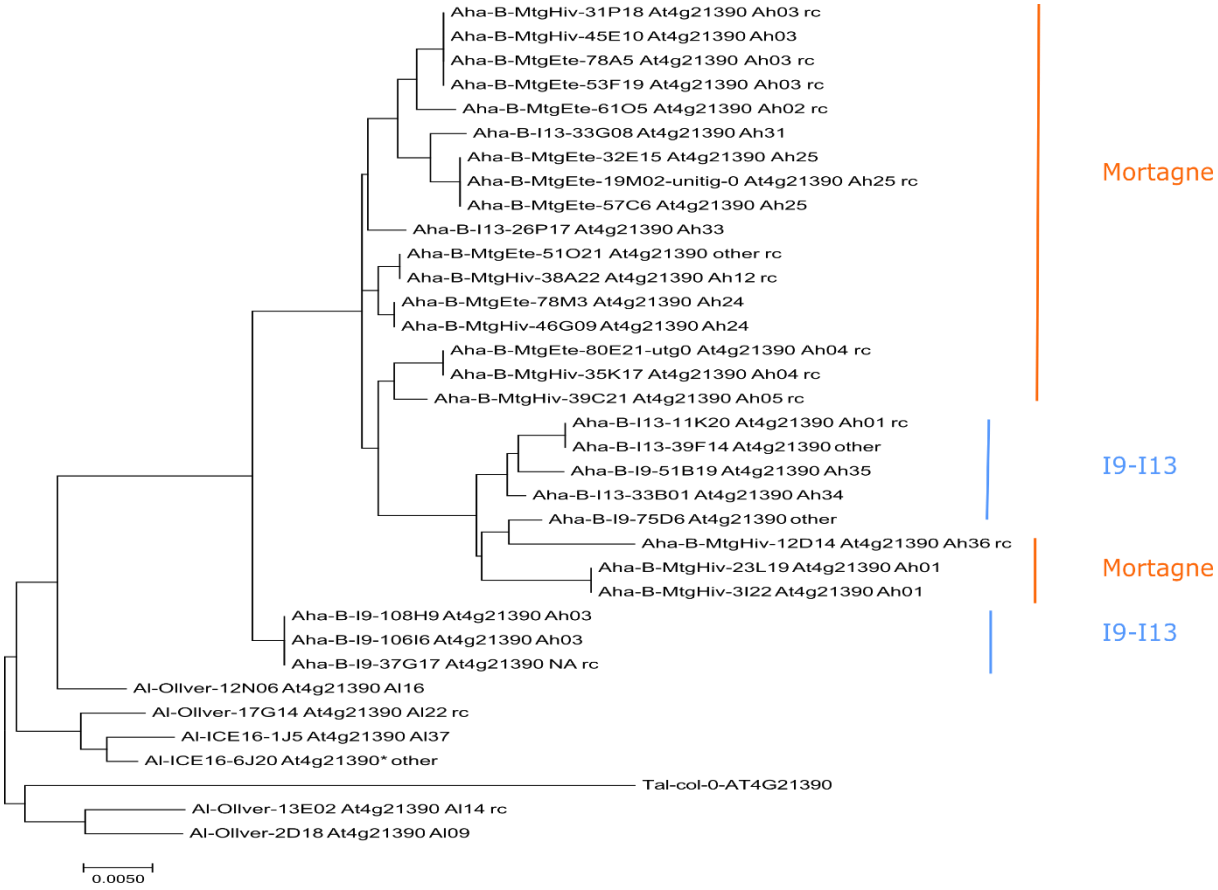


Figure S11: Phylogénie du gène *At4g21390* obtenue par maximum de vraisemblance. On constate une organisation par population, et ensuite par allèles au sein des populations, il perdure toutefois une structure des allèles avec un groupe d'individus de la population Mortagne possédant notamment l'allèle Ah01 qui sont placés au sein du groupe italiens.

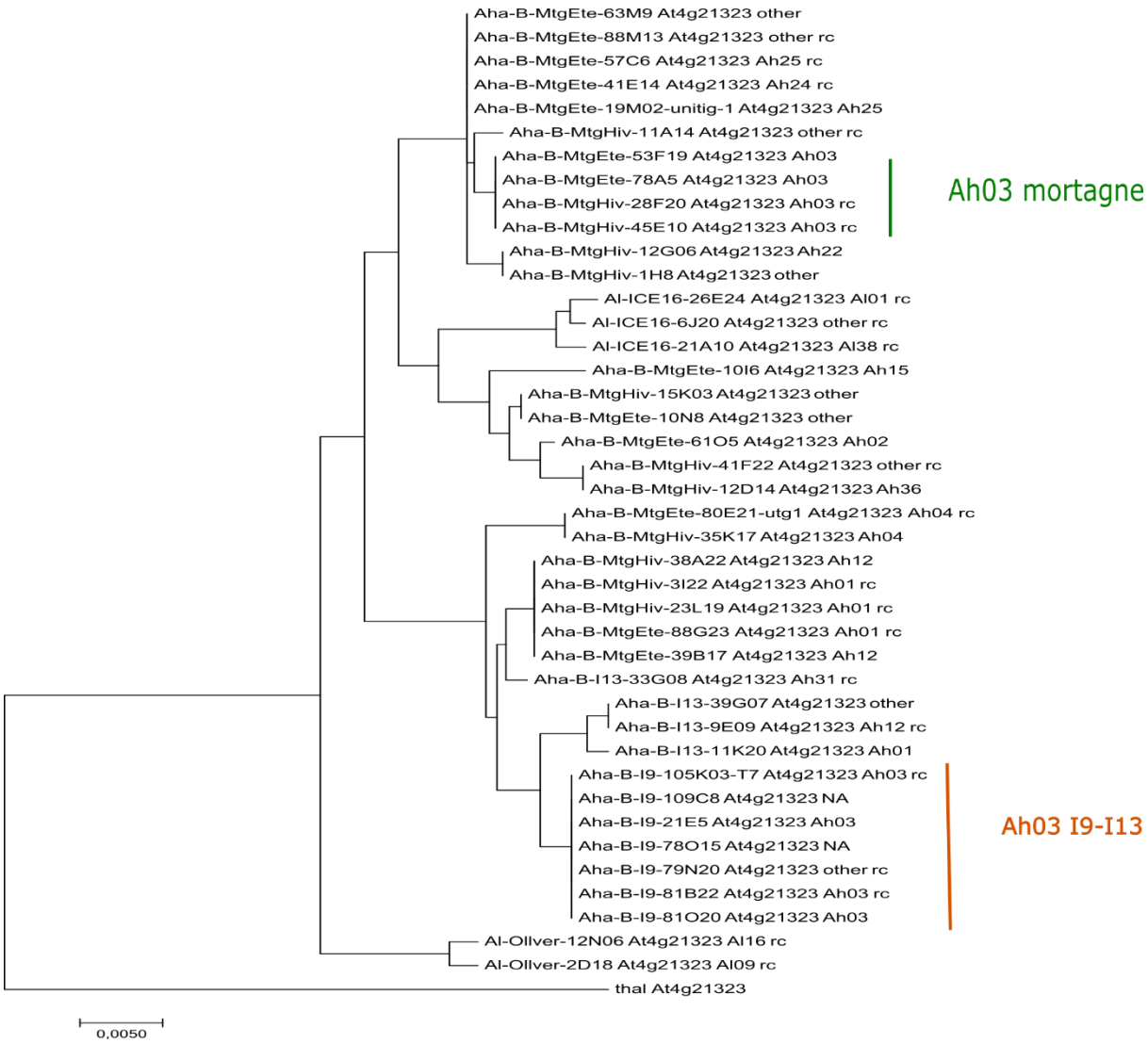


Figure S12: Phylogénie du gène *At4g21323* obtenue par maximum de vraisemblance. Au sein des populations, les clones portant les mêmes allèles sont regroupés entre eux. Ce qui veut dire que le déséquilibre de liaison au locus S est encore présent à cette distance.

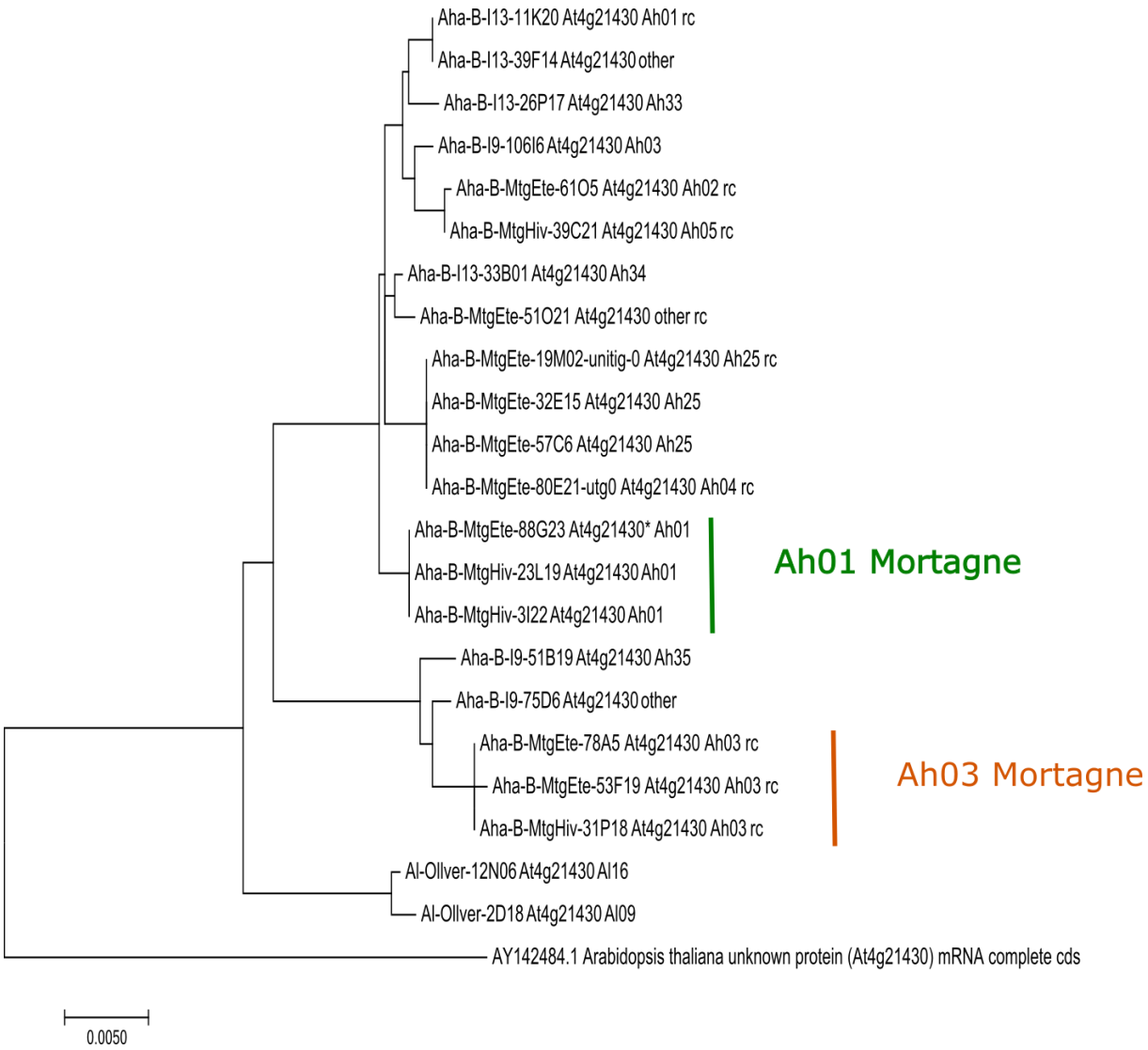


Figure S13: Phylogénie du gène *At4g21430* obtenue par maximum de vraisemblance. Au sein des populations, les clones portant les mêmes allèles sont regroupés entre eux. Ce qui veut dire que le déséquilibre de liaison au locus S est encore présent à cette distance.

Discussion & Perspectives

Le locus d'auto-incompatibilité offre l'opportunité de travailler sur des questions de sélection naturelle et d'évolution dans le contexte bien particulier qu'est celui de la sélection balancée. Cette forme de sélection, bien que connue depuis très longtemps, possède toujours des aspects qui restent difficiles à étudier. Parce qu'elle augmente la diversité des régions soumises à son régime de sélection, elle rend l'accès aux données génomiques difficile car elle nécessite de prendre en compte toute cette diversité, ce qui n'était pas possible sans grande peine avant l'émergence des NGS. Le système d'auto-incompatibilité sporophytique chez *A. halleri* présente la caractéristique de présenter de nombreux allèles organisés selon une hiérarchie de dominance (Llaurens *et al.*, 2008). Lors de ce projet, je me suis intéressé d'une part aux causes de cette hiérarchie de dominance, en contribuant à caractériser les éléments génétiques qui la contrôlent, et d'autre part aux conséquences qu'elle peut avoir, en termes d'accumulation du polymorphisme, tant sur les allèles S eux-mêmes que dans la région génomique à laquelle ils sont liés.

La hiérarchie de dominance est sous contrôle d'un réseau d'interactions d'un ensemble de miRs et de cibles (Durand *et al.*, 2014) mais les bases moléculaires qui régissent l'interaction entre ces miRS et leurs cibles restent incomplètement comprises. Mon travail a permis de préciser les critères selon lesquels se met en place une interaction de dominance entre deux allèles. Nous avons validé la généralité du contrôle transcriptionnel de cette dominance et défini un critère d'appariement basé sur le score d'alignement entre le miR et sa cible, et montré que le contrôle transcriptionnel repose sur un système à seuil au-delà duquel l'interaction entre les deux éléments de la machinerie se traduit de manière phénotypique. S'il est probable que ces miR agissent de la même manière que chez *Brassica* (avec possiblement une valeur du seuil qui peut être différente en lien avec des machineries de silencing qui pourraient différer entre *Arabidopsis* et *Brassica*), nous ne connaissons toujours pas la nature exacte de ces miRs ainsi que leur mécanisme d'action et les voies métaboliques impliqués dans cette régulation. La production et l'étude de mutant KO pour les différents éléments présent dans les potentielles voies métabolique de genèse et d'action de ces miRs

(en particulier en lien avec la voie RdDM) est en cours au sein de l'équipe et nous renseignera pour savoir quelles voies métaboliques ces miRs modificateurs de dominance utilisent pour réaliser leur fonction de silencing transcriptionnel. En particulier, une hypothèse intrigante pourrait être que différents miRs utilisent différentes voies métaboliques.

Si l'existence des modificateurs de dominance prédite par Fisher (1928) est généralement admise, nous ne connaissons pas l'intensité de la sélection agissant sur ces motifs. Wright, en 1929 prédisait que cette sélection serait de l'ordre de grandeur du taux de mutation, soit très faible et donc peu susceptible d'évoluer suffisamment pour induire des relations de dominance entre allèles rapidement après diversification. C'est pourquoi, nous nous sommes intéressés dans un second temps au polymorphisme de ces modificateurs de dominance au sein des lignées alléliques. Nous avons pu mettre en avant le fait que la pression de sélection sur ces modificateurs de dominance était du même ordre de grandeur que pour les régions codantes, soit beaucoup plus que ce que préconisait Wright. Si nos résultats confirment la présence de sélection purifiante sur les motifs (Jovelin & Cutter, 2014), nous n'avons pas pu trouver de trace de mutations compensatrices traduisant un processus d'évolution en cours au sein des différentes populations. Ceci peut être à mettre en lien avec l'existence du système à seuil, la plupart des mutations observées ne modifiant pas le contrôle transcriptionnel, et ne causant donc pas de pression de sélection pour « compenser » un éventuel écart. Augmenter à la fois le nombre de copies d'allèles mais aussi le nombre d'allèles différents, permettra peut-être de découvrir des traces de telles mutations compensatrices. Par ailleurs, l'étude des processus de divergence entre espèces (plutôt que de polymorphisme au sein d'une espèce, comme présenté ici), pourrait permettrait de mieux caractériser les processus de coévolution entre les deux éléments de régulation que sont les miRs et leurs cibles. Une analyse comparée des allèles S d'Arabidopsis avec ceux présents chez Capsella est en cours en collaboration avec l'équipe de Tanja Slotte (Université de Stockholm, Suède).

Par ailleurs, la présence de cette hiérarchie de dominance modifie les prédictions en termes de polymorphisme à plusieurs échelles. La sélection balancée maintient sur de très longues périodes de temps les allèles fonctionnels (Vekemans & Slatkin, 1994). Le temps de coalescence de ces allèles fonctionnels varie en fonction de leur position dans la hiérarchie de dominance (Castric *et al.*, 2010). En effet, les allèles récessifs, parce qu'ils ne sont que très

rarement exprimés, sont transmis de manière passive. Ils sont donc présents en plus grand nombre à l'échelle des populations et donc sont maintenus pendant une plus longue période temps. On s'attend donc à une asymétrie dans l'accumulation de polymorphisme intra-allélique entre allèles dominant accumulant peu de polymorphisme et récessifs plus polymorphes. Nous avons cependant confirmé l'absence de corrélation entre dominance et polymorphisme intra-allélique précédemment décrite par Castric *et al.*, 2010. Cette absence est surprenante et indique probablement que des processus sélectifs non pris en compte dans les modèles sont à l'œuvre, possiblement en lien avec la vitesse de turnover allélique qui pourrait différer entre allèles dominants et récessifs. Les modèles de diversification actuellement disponibles n'ont étudié que des systèmes gamétophytiques (Uyenoyama *et al.*, 2001; Gervais *et al.*, 2011). Une perspective intéressante à mon travail serait de développer des modèles de diversification qui prennent explicitement en compte les relations de dominance.

A une autre échelle, chaque allèle S va peu à peu acquérir dans ses régions flanquantes un cortège de SNPs associé qui lui est propre. A cette échelle on peut comparer les lignées alléliques à des dèmes entre lesquels les polymorphismes liés peuvent « migrer » par recombinaison (Kaplan *et al.*, 1988; Hudson, 1990; Nordborg, 1997; Takahata & Satta, 1998). La recombinaison aura tendance à casser cette association entre l'allèle S et la diversité associée, agissant sur les haplotypes de la même manière que la migration entre différentes populations. On s'attend en conséquence à trouver une augmentation du polymorphisme aux abords des locus soumis à la sélection balancée (DeGiorgio *et al.*, 2014), sur une étendue plus ou moins vaste, dépendante du temps de coalescence des lignées alléliques et du taux de recombinaison. L'étude des régions flanquantes nous a permis de montrer que l'augmentation du pic de polymorphisme était détectable, visible dans les 10kb immédiatement aux abords du locus S, mais qu'il ne semblait pas être associé à un ratio π_N/π_S plus élevé. La question de l'architecture du fardeau lié reste donc entière : si le locus S lui-même ne contient pas d'autres gènes que ceux qui contrôlent l'auto-incompatibilité elle-même, si les gènes flanquants ne montrent pas d'élévation du π_N/π_S et que les SNPs qu'ils portent ne forment pas de cortèges associés aux différents allèles S, sur quelle base le fardeau peut-il se constituer ? Une possibilité est que ce fardeau se constitue essentiellement au sein

des populations, mais ne soit pas maintenu à l'échelle de l'espèce. Tester cette hypothèse nécessiterait de comparer le fardeau lié en prenant en compte l'échelle des populations, en forçant l'homozygotie pour des allèles S identiques mais issus soit de la même population soit de populations différentes. Cette échelle a généralement été négligée dans les études sur l'auto-incompatibilité étant donnés les temps de coalescence considérables entre lignées alléliques (mais voir Schierup *et al.*, 2000), bien supérieurs à l'échelle de la coalescence neutre dans une population même fortement subdivisée. Nos résultats montrent cependant que l'échelle spatiale à laquelle on étudie le polymorphisme des régions liées est déterminante.

Bibliographie

Castric V, Bechsgaard JS, Grenier S, Noureddine R, Schierup MH, Vekemans X. **2010**. Molecular evolution within and between self-incompatibility specificities. *Molecular Biology and Evolution* **27**: 11–20.

DeGiorgio M, Lohmueller KE, Nielsen R. **2014**. A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genetics* **10**: e1004561.

Durand E, Méheust R, Soucaze M, Goubet PM, Gallina S, Poux C, Fobis-loisy I, Guillon E, Gaude T, Sarazin A, *et al.* 2014. Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. *Science* **346**: 1200–1205.

Fisher RA. **1928**. The Possible Modification of the Response of the Wild Type to Recurrent Mutations. *The American Naturalist* **62**: 115–126.

Gervais CE, Castric V, Ressayre A, Billiard S. **2011**. Origin and diversification dynamics of self-incompatibility haplotypes. *Genetics* **188**: 625–636.

Hudson RR. **1990**. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**: 1–44.

Jovelin R, Cutter AD. **2014**. Microevolution of nematode miRNAs reveals diverse modes of selection. *Genome biology and evolution* **6**: 3049–63.

Kaplan NL, Darden T, Hudson RR. **1988**. The coalescent process in models with selection. *Genetics* **120**: 819–829.

Llaurens V, Billiard S, Leducq JB, Castric V, Klein EK, Vekemans X. **2008**. Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in *Arabidopsis halleri*? *Evolution* **62**: 2545–2557.

Nordborg M. **1997**. Structured Coalescent Processes on Different Time Scales.

Schierup MH, Vekemans X, Charlesworth D. **2000**. The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetical Research* **76**: 51–62.

Takahata N, Satta Y. **1998**. Footprints of intragenic recombination at HLA loci. *Immunogenetics* **47**: 430–441.

Uyenoyama MK, Zhang Y, Newbigin E. **2001**. On the origin of self-incompatibility haplotypes: Transition through self-compatible intermediates. *Genetics* **157**: 1805–1817.

Vekemans X, Slatkin M. **1994**. Allelic Genealogies at a Gametophytic Self-Incompatibility Locus. *Genetics* **137**: 1157–1165.

Wright S. **1929**. Fisher's Theory of Dominance. *The American Naturalist* **63**: 274–279.

Résumé :

La dominance entre allèles S chez les Brassicaceae est contrôlée par un ensemble de petits ARNs non codants et de leurs séquences cibles. Les relations de dominance qu'ils établissent ont un ensemble de conséquences sur l'accumulation du polymorphisme au locus S lui-même mais également dans les régions immédiatement liées. L'idée du projet a été d'étudier 1) les critères d'appariement selon lesquels les petits ARNs non-codants provoquent le silencing transcriptionnel, 2) la diversité des petits ARNs, de leurs précurseurs et de leurs cibles en populations naturelles afin de déterminer si les contraintes fonctionnelles qu'ils subissent s'apparentent à celles qui pèsent en général sur les miRNAs et enfin 3) la diversité des séquences flanquantes, afin de déterminer d'une part l'ampleur du pic de polymorphisme attendu en raison de la sélection balancée et d'autre part si on peut détecter un fardeau de mutations délétères spécifique à chaque allèle le long de la hiérarchie de dominance.

Abstract :

The dominance between S-alleles in the Brassicaceae is controlled by a set of small non-coding RNAs and their cognate targets. These dominance relationships have important consequences on the polymorphism accumulated at the S-locus itself but also at the flanking regions. The aim of the project was to study 1) the base-pairing criteria by which the small non-coding RNAs transcriptionally silence their target gene, 2) the diversity of these small RNAs, of their precursors and targets in natural populations in order to determine if the selective constraint they undergo is similar to what we know for other miRNAs genes in the genome, and finally, 3) the diversity of the flanking regions, to determine the size of the predicted peak of polymorphism peak caused by balancing selection, test whether genes in these regions show evidence of the predicted sheltered load and whether polymorphisms at these genes are specifically associated with S-alleles.