POLE DE RECHERCHE ET D'ENSEIGNEMENT SUPERIEUR – Université Lille Nord de France
UNIVERSITE LILLE 1 – Sciences et Technologies
ECOLE DOCTORALE SMRE – Sciences de la Matière, du Rayonnement et de l'Environnement

THESE

SOUTENANCE : 2 décembre 2011
DISCIPLINE : Biologie Evolutive

APPORTS DES APPROCHES DE GENOMIQUE CIBLEE DANS L'ETUDE DES PATRONS D'EVOLUTION MOLECULAIRE DU LOCUS D'AUTO-INCOMPATIBILITE DANS LE GENRE *ARABIDOPSIS*

par **Pauline GOUBET**

Composition du jury :

Deborah CHARLESWORTH, Professeur – Université d'Edimbourg	Rapporteur
Maud TENAILLON, CR1 CNRS – Université Paris XI	Rapporteur
Pierre CAPY, Professeur – Université Paris XI	Examinateur
Mathieu JORON, CR1 CNRS – Museum National d'Histoire Naturelle	Examinateur
Pascal TOUZET, Professeur – Université Lille 1	Président de jury
Vincent CASTRIC, CR1 CNRS – Université Lille 1	Directeur de thèse
Xavier VEKEMANS, Professeur – Université Lille 1	Co-directeur de thèse

REMERCIEMENTS

Ces trois années de thèse, et par extension les années de M1 et de M2, passées au laboratoire GEPV (Génétique et Evolution des Populations Végétales) ont été l'occasion de nombreux échanges et rencontres. Je tiens tout d'abord à remercier Vincent pour m'avoir fait confiance en me proposant cette thèse, pour sa patience, sa gentillesse et sa présence. Je remercie également Xavier pour ses conseils toujours avisés et sa bonne humeur.

J'ai ensuite tant de personnes à remercier que j'espère ne pas en oublier. Merci tout d'abord à toutes celles qui ont fait progresser ma thèse, que ce soit par des conseils judicieux, par des discussions fructueuses, des PCR réalisées ou par des logiciels domptés : entre autres Sophie, Camillo, Anne-Catherine, Maude Pupin, Julien Dutheil… Un merci tout particulier à celles qui m'ont permis de m'intégrer au sein du laboratoire : Merci à Muriel et à toutes les personnes qui m'ont aidée pour une porte, le manteau ou l'ascenseur. La liste serait bien trop longue pour que je puisse l'établir. Merci ensuite à ceux qui pendant ces trois ans ont partagé les joies de la thèse : Aude, Benjamin, Lucy, Camilla, Emna, Romain… Merci également à Angélique et Cédric pour m'avoir régulièrement accueillie à la serre pour la petite discussion du matin. Merci enfin à tous ceux et celles qui se sont arrêtés au bureau 02 pour une pause bonne humeur : Adeline, Cécile, Sandrine, Michèle, Claire-Lise et tant d'autres… Vous avez tous contribué à votre manière, directement ou indirectement, à l'épanouissement étrange des deux thésardes de ce bureau. La liste pourrait encore s'allonger mais il faut bien l'arrêter à un moment ou à un autre…

Pour finir, comment parler de ma thèse sans parler de mes deux fidèles acolytes ? Betty tout d'abord, ou la princesse aux cheveux de feu qui jamais ne savait ce qu'elle voulait. Isabelle ensuite, sans qui cette thèse n'aurait assurément pas été la même. Ces trois ans ont scellé notre duo et j'espère bien entendre ton accent chanter encore longtemps. A toi, j'associe bien volontiers Raphaël. Merci à vous trois pour ces soirées crêpes, fous rires et autres aventures…

SOMMAIRE

LISTE DES ENCADRES

LISTE DES FIGURES

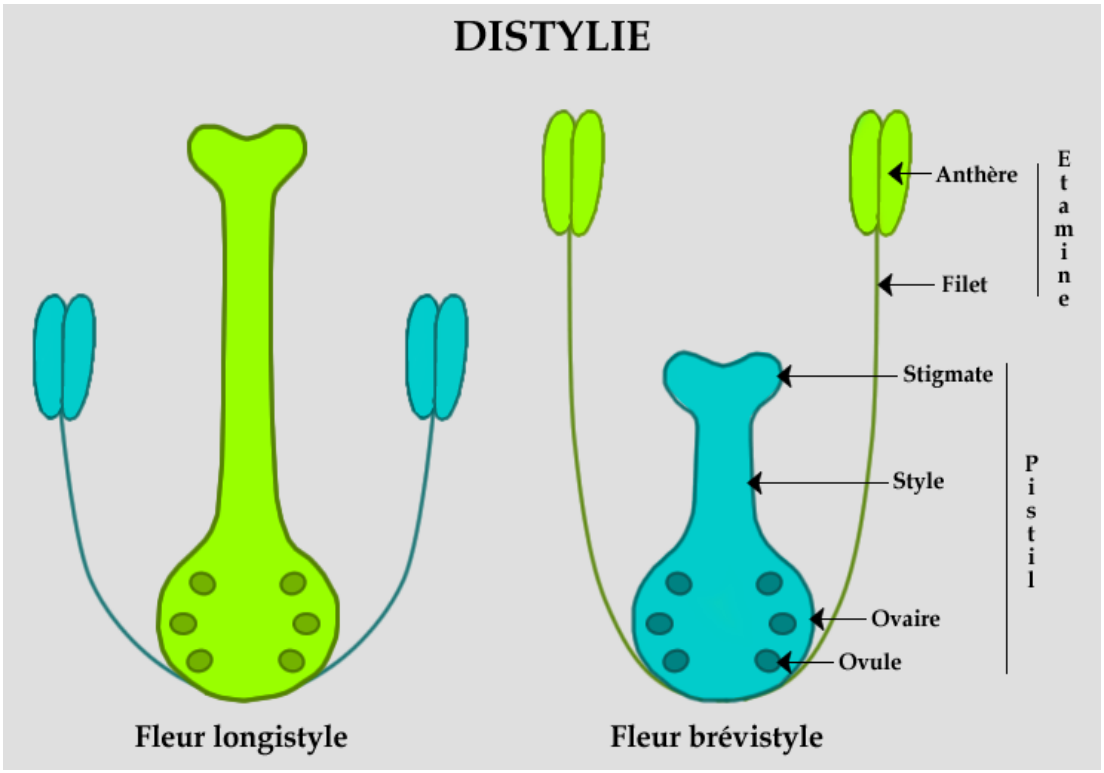
LISTE DES TABLEAUX

INTRODUCTION

Depuis le début des années 2000, la biologie traverse une véritable révolution en raison des progrès fulgurants des techniques de séquençage à haut débit : la révolution génomique. Grâce aux avancées dans ces nouvelles technologies, le nombre d'espèces dont le génome est entièrement séquencé s'accroît de manière exponentielle, et ce type de séquençage est de plus en plus commun et rapide. Ainsi, en passant de l'étude d'un gène ou d'une petite fraction du génome à celle du génome dans son ensemble, la génomique constitue un changement d'échelle majeur et offre des perspectives dans des domaines très variés, comme par exemple le domaine biomédical, l'exploration de la biodiversité, l'amélioration d'espèces à intérêt économique ou encore la compréhension de l'évolution des espèces. Néanmoins, malgré l'espoir qu'il fournit d'appréhender le fonctionnement du monde vivant dans toute sa complexité, le séquençage de génomes complets comporte également des limites. En effet, l'acquisition de quantités gigantesques de données nécessite la mise en place de nouvelles méthodes, à la fois de stockage, de gestion et d'analyse, et implique une utilisation de plus en plus intensive de l'outil informatique. De plus, les génomes ne sont pas des ensembles homogènes et des régions se révèlent plus difficiles à séquencer ou à analyser. C'est le cas par exemple des régions génomiques qui comprennent de grandes séquences répétées que les programmes d'assemblage ne peuvent pas reconstituer avec certitude, en particulier lorsqu'il s'agit d'assemblage *de novo*. C'est le cas également de certaines régions du génome qui présentent une diversité moléculaire particulièrement forte et qui mettent en échec les approches d'assemblage reposant sur l'alignement par rapport à une séquence connue et supposée unique. Ces propriétés peuvent être dues à l'action de contraintes sélectives fortes et originales, comme celles agissant au locus d'auto-incompatibilité chez les plantes à fleurs. Etant soumis à une forte sélection fréquence-dépendante et à la quasi-suppression de la recombinaison, il s'agit d'une région génomique fortement polymorphe, à la fois par le nombre d'haplotypes et par leur divergence. Ainsi, le séquençage d'un seul de ces haplotypes ne suffit pas à comprendre la dynamique du système, et il reste nécessaire d'en étudier les différentes versions. Dans ce contexte, cette présente thèse consiste en une analyse bioinformatique du locus d'auto-incompatibilité sporophytique dans le groupe d'espèces *Arabidopsis*, et illustre les intérêts de méthodes génomiques complémentaires au séquençage du génome complet des espèces dans l'analyse des régions hautement polymorphes.

I - DESCRIPTION DE L'AUTO-INCOMPATIBILITE

Chez les Angiospermes, l'hermaphrodisme est un caractère très répandu. En présentant les organes reproducteurs mâles et femelles chez les mêmes individus, les espèces hermaphrodites ont l'opportunité de pratiquer l'autofécondation, c'est-à-dire de féconder leur pistil par leur propre pollen. Cependant, parce qu'elle favorise l'homozygotie, cette stratégie de reproduction entraîne l'expression



Encadré 1. Schématisation de la distylie.

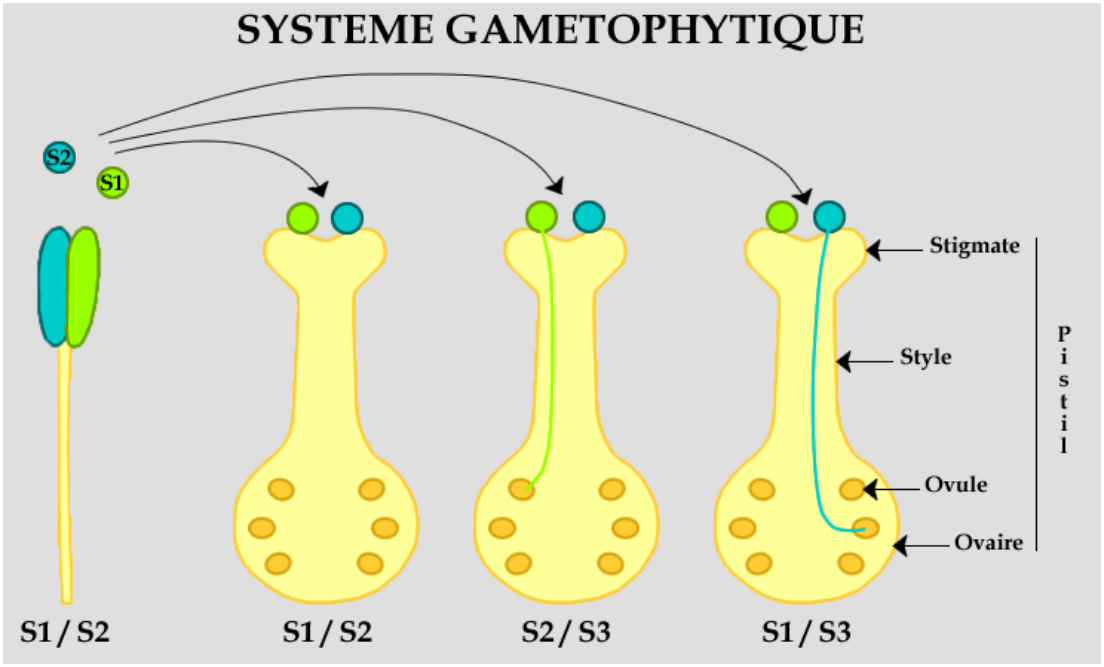
La distylie est un polymorphisme floral se caractérisant par la coexistence de deux morphes : les individus à fleurs longistyles, qui possèdent des styles plus longs que les étamines, et les individus à fleurs brévistyles, qui possèdent des styles plus courts que les étamines. Le pollen est transporté par les insectes d'une étamine vers un style de même longueur.

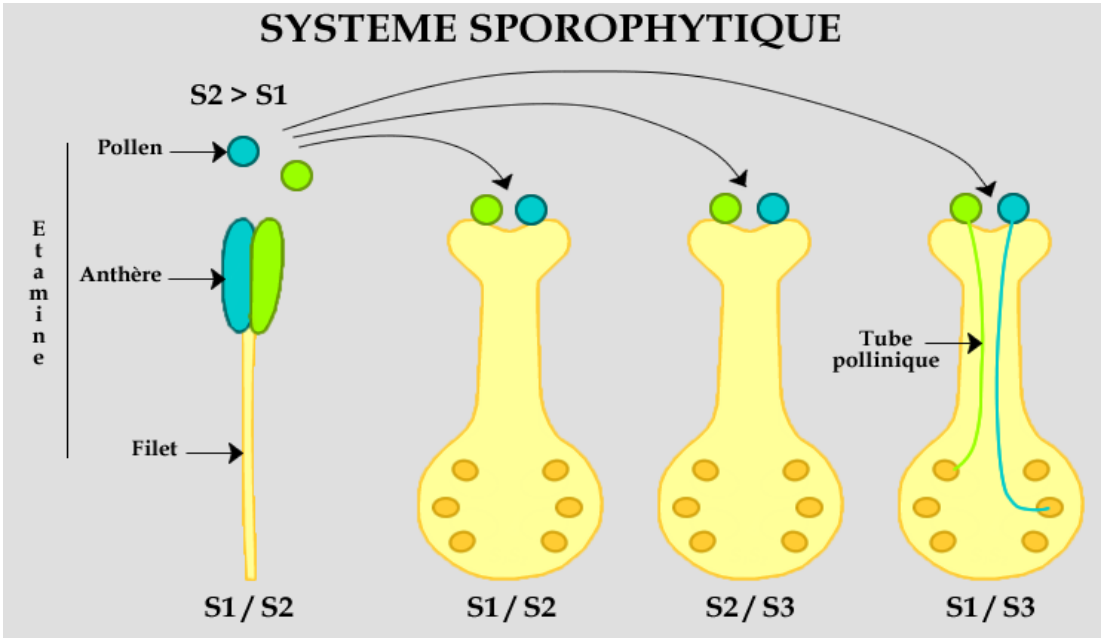
de mutations délétères récessives. C'est ce que l'on appelle la dépression de consanguinité. Divers mécanismes ont été mis en place par les plantes pour éviter cette diminution de valeur sélective. La dichogamie permet par exemple un décalage dans le temps de la période de maturité des organes mâles et femelles d'un même individu. On parle de protandrie lorsqu'il y a maturation des organes mâles avant les organes femelles, et de protogynie dans le cas contraire. D'une autre manière, les systèmes d'auto-incompatibilité hétéromorphes sont caractérisés par la coexistence, au sein des populations naturelles, de plusieurs morphes. La fécondation n'étant possible qu'entre morphes différents, l'autofécondation est ainsi évitée. Ce type de système comprend notamment la distylie, dont les deux morphes se différencient par la taille de leurs styles et de leurs étamines (voir l'encadré 1).

Présents chez environ 40% des Angiospermes (IGIC and LANDE 2008), les systèmes d'auto-incompatibilité homomorphes sont néanmoins le moyen le plus répandu de limiter la dépression de consanguinité (DE NETTANCOURT 2001; REA and NASRALLAH 2008). Ces systèmes, apparus plusieurs fois de manière indépendante au cours de l'évolution des Angiospermes (STEINBACHS and HOLSINGER 2002), reposent sur l'existence de catégories de phénotypes constituant des groupes d'incompatibilité (ou spécificités) : un croisement n'est possible qu'entre individus de groupes phénotypiques différents. Basés sur une reconnaissance hautement spécifique de type clé-serrure entre protéines du pistil et du pollen, ils permettent d'éviter l'autofécondation. On distingue parmi ces systèmes l'auto-incompatibilité gamétophytique et l'auto-incompatibilité sporophytique.

A - SYSTEME D'AUTO-INCOMPATIBILITE GAMETOPHYTIQUE

Documenté chez de nombreuses familles, c'est le plus commun des systèmes d'auto-incompatibilité mais ses bases moléculaires restent dans la majorité des cas encore très mal connues (FRANKLIN-TONG and FRANKLIN 2003). Contrairement au pistil dont le phénotype d'auto-incompatibilité dépend de ses deux allèles (tous les individus sont en principe hétérozygotes au locus S), le phénotype du pollen dépend ici de son génotype haploïde. La fécondation est donc inhibée si un grain de pollen possède un allèle identique à l'un des deux allèles du pistil. Deux mécanismes différents d'auto-incompatibilité gamétophytique ont été caractérisés au niveau moléculaire. Le premier de ces systèmes est partagé par les Plantaginaceae, les Rosaceae, les Solanaceae et les Rubiaceae. Le pistil y sécrète une protéine de type S-RNase (HUA *et al.* 2008) qui pénètre dans le tube pollinique de manière non spécifique, alors que le phénotype du pollen est déterminé par une ou plusieurs protéines de la famille des SLF (S-locus F-box (KUBO *et al.* 2010; LAI *et al.* 2002)). En l'absence de reconnaissance entre les protéines du pollen et du pistil, la S-RNase est dégradée. Au contraire, si les protéines du pistil et du pollen expriment le même phénotype, la S-RNase entraîne l'auto-incompatibilité en bloquant la croissance du tube pollinique par dégradation de ses ARN. Le second système a été caractérisé chez *Papaver rhoeas*, une espèce de la famille des Papaveraceae. L'interaction entre une protéine sécrétée par le stigma, PrsS (*Papaver rhoeas* stigma S determinant), et une protéine transmembranaire du pollen, PrpS (*Papaver*





Encadré 2. Exemples d'interactions entre les allèles du pollen et du pistil dans le cas où ceux du pistil sont codominants.

Dans le cas du système gamétophytique, le phénotype du pollen est déterminé par son génotype haploïde. Un parent S1 / S2 produit donc des grains de pollen exprimant un phénotype S1 et des grains de pollen exprimant un phénotype S2.

Dans le cas du système sporophytique, le phénotype du pollen est déterminé par le génotype diploïde du parent. Si S2 est dominant sur S1, alors un parent S1 / S2 ne produit que des grains de pollen exprimant un phénotype S2.

rhoeas pollen S determinant), y entraîne une cascade de réactions intracellulaires dans le pollen incompatible (McCLURE and FRANKLIN-TONG 2006; WHEELER *et al.* 2009). Ces réactions impliquant les ions Ca^{2+} stoppent l'élongation du tube pollinique, et sont suivies par la mise en place d'un mécanisme d'apoptose (THOMAS and FRANKLIN-TONG 2004).

B - SYSTEME D'AUTO-INCOMPATIBILITE SPOROPHYTIQUE

Les systèmes d'auto-incompatibilité sporophytique sont également retrouvés dans de nombreuses familles parmi lesquelles les Brassicaceae, les Convolvulaceae ou les Asteraceae. Le phénotype du pollen y est déterminé par le génotype diploïde de son parent. Des relations de dominance complexes, avec présence de classes de dominance, peuvent alors intervenir entre les différents allèles d'auto-incompatibilité (KUSABA *et al.* 2002). L'encadré 2 présente quelques exemples d'interactions entre pollen et pistil selon le type de système d'auto-incompatibilité.

A ce jour, les bases moléculaires de l'auto-incompatibilité sporophytique n'ont été caractérisées que chez les Brassicaceae.

1 - Le système d'auto-incompatibilité sporophytique chez les Brassicaceae

Chez les Brassicaceae, l'auto-incompatibilité sporophytique est le système de reproduction ancestral (NASRALLAH *et al.* 2004), et est contrôlée par une région génomique que l'on appelle le locus S. Ce locus comprend au moins deux gènes :

- Le gène *SCR* (S-Locus Cystein Rich, (SCHOPFER *et al.* 1999)), également parfois nommé *SP11* dans le genre *Brassica* (SUZUKI *et al.* 1999), code une protéine de petite taille. Bien qu'extrêmement divergente, cette protéine présente des résidus de cystéine conservés qui forment des ponts disulfures par paires, et sont importants pour sa structure tertiaire (WATANABE *et al.* 2000). Elle est exprimée dans le pollen mais aussi de manière sporophytique dans le tapetum des anthères, et se retrouve ainsi déposée à la surface du pollen.

- Le gène *SRK* (S-Locus Receptor Kinase, (STEIN *et al.* 1991; TAKASAKI *et al.* 2000)) code une protéine transmembranaire du pistil. Cette protéine comprend trois domaines distincts : le domaine S est un domaine extracellulaire agissant comme récepteur de la protéine pollinique ; le domaine transmembranaire sert à ancrer la protéine à la surface du pistil ; le domaine kinase est un domaine intracellulaire chargé de transmettre le signal aboutissant au rejet du pollen. Dans le genre *Brassica*, un second gène est parfois présent, *SLG* (S-Locus Glycoprotein (NASRALLAH *et al.* 1985)). Très proche de *SRK* et également exprimé dans le stigma, il aurait pour rôle d'amplifier la réaction d'auto-incompatibilité (BOYES and NASRALLAH 1993). Parce que les allèles du gène *SLG* ségrégent avec les phénotypes d'auto-incompatibilité, ce gène a longtemps été considéré comme un bon candidat pour être le déterminant femelle

(NASRALLAH *et al.* 1985). Ce fonctionnement en famille de gènes souligne l'importance d'accéder à la séquence génomique d'un système complexe tel que le système d'auto-incompatibilité pour en comprendre le mécanisme.

Lorsque le grain de pollen est déposé sur le pistil, si les protéines SCR et SRK expriment la même spécificité, c'est-à-dire le même phénotype, leur interaction moléculaire entraîne une cascade de réactions qui se traduit par l'inhibition de la croissance du tube pollinique.

S'agissant d'un système sporophytique, des relations de dominance existent entre les différents allèles d'auto-incompatibilité. Ainsi, deux classes de dominance coexistent dans le genre *Brassica* (HATAKEYAMA *et al.* 1998; UYENOYAMA 1995), la classe I étant dominante sur la classe II dans l'expression du phénotype pollen, et les allèles d'une même classe étant codominants entre eux. Sur la base de données phylogénétiques et de croisements, les allèles d'auto-incompatibilité du genre *Arabidopsis* ont quant à eux été répartis dans quatre classes de dominance (PRIGODA *et al.* 2005; SCHIERUP *et al.* 2006), la classe I étant la plus récessive, et la classe IV la plus dominante. Cependant, en raison du grand nombre d'allèles au locus d'auto-incompatibilité, les interactions possibles deux à deux sont très nombreuses. Les expériences permettant de les vérifier étant par conséquent importantes en termes de croisements à effectuer, seule une faible fraction de ces interactions par paires ont été validées expérimentalement (LLAURENS *et al.* 2008; MABLE *et al.* 2003).

II - PROCESSUS EVOLUTIFS ET PARTICULARITES LIES AU LOCUS D'AUTO-INCOMPATIBILITE

Le locus d'auto-incompatibilité est soumis à des contraintes sélectives intenses, à savoir la sélection fréquence-dépendante, une forme de sélection naturelle qui confère aux phénotypes des valeurs sélectives différentes selon leur fréquence, et la restriction importante du mécanisme de recombinaison. Comme décrit ci-dessous, ces contraintes ont un impact fort sur la région génomique concernée et influent sur son évolution.

A - SELECTION FREQUENCE-DEPENDANTE

1 - Divergence inter- et intra-spécifique

La sélection fréquence-dépendante semble être la force évolutive principale agissant au locus d'auto-incompatibilité (WRIGHT 1939) et une validation empirique de ce type de sélection entre deux générations a par ailleurs été publiée récemment (STOECKEL *et al.* 2011). En effet, moins un allèle d'auto-incompatibilité est fréquent, plus il profitera de l'avantage du rare conféré par la sélection fréquence-dépendante, son pourcentage de partenaires compatibles étant plus élevé. Ainsi moins soumis à l'action de la dérive génétique, les allèles tendent à être maintenus sur de plus longues périodes de temps (VEKEMANS and SLATKIN 1994). Il a d'ailleurs été montré que leur divergence était antérieure à la divergence des espèces (DWYER *et al.* 1991). C'est ainsi que des groupes d'espèces

proches peuvent présenter un polymorphisme trans-spécifique, c'est-à-dire qu'elles partagent une partie de leurs allèles d'auto-incompatibilité. C'est le cas par exemple des espèces du genre *Arabidopsis* (CASTRIC *et al.* 2008) ou du genre *Brassica* (SATO *et al.* 2003).

De plus, parce que les allèles d'auto-incompatibilité ont un temps de résidence plus important que les allèles d'un locus neutre, ils tendent également à être extrêmement divergents au sein d'une même espèce. Cette forte divergence a été mise en évidence chez de nombreuses espèces telles que *Solanum carolinense* (RICHMAN *et al.* 1995), *Crataegus monogyna, Sorbus aucuparia* (RASPE and KOHN 2002) dans le cas d'un système gamétophytique, ou encore *Arabidopsis lyrata* (MABLE *et al.* 2003) et *Brassica rapa* (NOU *et al.* 1993) dans le cas d'un système sporophytique.

2 - Diversité allélique

En favorisant les allèles présents en faibles fréquences, il est attendu que la sélection permette la coexistence d'un nombre important d'allèles au locus d'auto-incompatibilité (WRIGHT 1939). Cette prédiction a été vérifiée de manière empirique chez de nombreuses espèces (CASTRIC and VEKEMANS 2004). Ainsi, avant même l'avènement de la biologie moléculaire, les premières études portant sur l'auto-incompatibilité gamétophytique décrivaient 34 allèles différents se maintenant en populations naturelles chez *Oenothera organensis* (EMERSON 1939). Des résultats similaires ont été trouvés chez diverses autres espèces comme par exemple *Papaver rhoeas*, avec 31 allèles détectés (CAMPBELL and LAWRENCE 1981) ou *Sorbus aucuparia*, avec 32 allèles décelés (RASPE and KOHN 2007). Chez les Brassicaceae, le système d'auto-incompatibilité sporophytique compte au moins 20 allèles différents chez *B. insularis* (GLEMIN *et al.* 2005), 30 allèles chez *B. oleracea* (OCKENDON 2000), 50 allèles chez *B. rapa* (NOU *et al.* 1993), 30 allèles chez *A. halleri* et 38 allèles chez *A. lyrata* (CASTRIC *et al.* 2008).

3 - Effets de la dominance

Dans le cas d'un système gamétophytique, les allèles du locus d'auto-incompatibilité sont considérés comme sélectivement égaux, et ils tendent par conséquent à atteindre des fréquences équivalentes (WRIGHT 1939). Le cas du système sporophytique est plus complexe, car les relations de dominance engendrent une sélection asymétrique parmi les allèles (SCHIERUP *et al.* 1997), et des fréquences différentes sont attendues selon leur niveau de dominance. Les allèles récessifs ayant la possibilité de se trouver dissimulés par des allèles plus dominants, ils peuvent augmenter en fréquence et apparaître à l'état homozygote. C'est ce que l'on appelle l'effet récessif (SAMPSON 1974). Au sein d'une même classe de dominance, la sélection s'exerce de manière uniforme sur les différents allèles et leurs fréquences tendent à être homogènes (UYENOYAMA 2000). Par ailleurs, on s'attend à observer à l'équilibre un allèle unique dans la classe la plus récessive, puis un nombre d'allèles d'autant plus élevé que la classe à laquelle ils appartiennent est dominante par rapport aux autres (BILLIARD *et al.* 2007).

Plusieurs études ont testé cette répartition particulière des fréquences alléliques en populations naturelles dans le cas d'un système sporophytique. Glémin *et al*. (2005) ont par exemple analysé la répartition des allèles d'auto-incompatibilité dans cinq populations corses de *B. insularis* et ont montré que les allèles de la classe dominante présentaient des fréquences globales plus faibles que les allèles de la classe récessive. La classe dominante compte par ailleurs dix-huit allèles au total alors que la classe récessive n'en compte que deux. De la même manière, Edh *et al.* (2009b) ont étudié quatre populations crétoises de *B. cretica* et ont également montré que les allèles récessifs étaient plus fréquents et moins nombreux que les allèles dominants. Schierup *et al.* (2008) ont quant à eux étudié douze populations islandaises d'*A. lyrata*. Contrairement au genre *Brassica* qui présente seulement deux classes de dominance distinctes (NASRALLAH *et al.* 1991), le genre *Arabidopsis* en présente au moins quatre (PRIGODA *et al.* 2005). La répartition des allèles de ces quatre classes a mis en évidence la présence d'un seul allèle à forte fréquence dans la classe récessive, ainsi que d'allèles d'autant plus nombreux et peu fréquents que leur classe était dominante. De plus, l'étude de Llaurens *et al.* (2008) dans une population d'*A. halleri* semble suggérer que les relations de dominance peuvent être légèrement différentes en ce qui concerne les phénotypes du pollen et du pistil.

B - RESTRICTION DE LA RECOMBINAISON

La recombinaison est également un processus important lié au locus d'auto-incompatibilité. Il s'agit d'un mécanisme génétique fondamental car il peut potentiellement permettre à la sélection naturelle d'agir indépendamment sur chaque gène (BARTON and CHARLESWORTH 1998). Son absence ou sa suppression dans une région génomique entraîne la transmission de cette même région comme une seule entité à la descendance. Parce que le maintien à long terme de l'auto-incompatibilité nécessite la conservation de la reconnaissance spécifique des protéines du pistil et du pollen, le système impose une liaison génétique forte des gènes impliqués et on s'attend à ce que la recombinaison soit restreinte dans cette région génomique (CHARLESWORTH and AWADALLA 1998). On peut en effet imaginer qu'un évènement de recombinaison entre *SCR* et *SRK* empêcherait la reconnaissance spécifique de leurs protéines et provoquerait une rupture de l'auto-incompatibilité. Le recombinant produit étant alors soumis à une forte dépression de consanguinité, il serait fortement contre-sélectionné. *SCR* et *SRK* étant donc transmis comme une seule entité, on parle d'haplotypes d'auto-incompatibilité, chacun de ces haplotypes étant constitué d'un gène *SCR*, de son gène *SRK* associé et de leur région génomique.

Cette restriction de la recombinaison entre haplotypes différents du locus d'auto-incompatibilité a été vérifiée chez les Brassicaceae par plusieurs études. Chez *Brassica*, aucun évènement de recombinaison n'a été détecté à l'intérieur du locus S (CASSELMAN *et al.* 2000). Takuno *et al.* (2007) ont quant à eux suggéré des évènements de recombinaison au sein du domaine kinase de *SRK*, domaine qui n'est pas impliqué dans la reconnaissance des protéines, mais pas entre les gènes *SCR* et *SRK*. Chez *Arabidopsis*, seuls de rares évènements de recombinaison dans une région limitée autour du locus S (KAWABE *et al.*

2006) ou entre des haplotypes phylogénétiquement proches (CASTRIC *et al.* 2010) ont été mis en évidence.

Par ailleurs, de la recombinaison entre différentes copies d'un même haplotype fonctionnel a été détectée dans le cas des haplotypes les plus récessifs (CASTRIC *et al.* 2010). En effet, comme expliqué précédemment, ces haplotypes peuvent former des combinaisons homozygotes, par croisement entre deux génotypes hétérozygotes partageant un même haplotype récessif. En l'absence d'une barrière à la recombinaison formée par la divergence entre haplotypes au sein de ces génotypes homozygotes, ils peuvent alors recombiner (CASTRIC *et al.* 2010). On ne s'attend donc pas à ce que la recombinaison soit totalement exclue du locus d'auto-incompatibilité, mais à ce qu'elle y soit fortement restreinte, et ce de manière différente selon le niveau de dominance. Ainsi, plus un haplotype est dominant, plus la restriction de la recombinaison qui s'y appliquera sera forte.

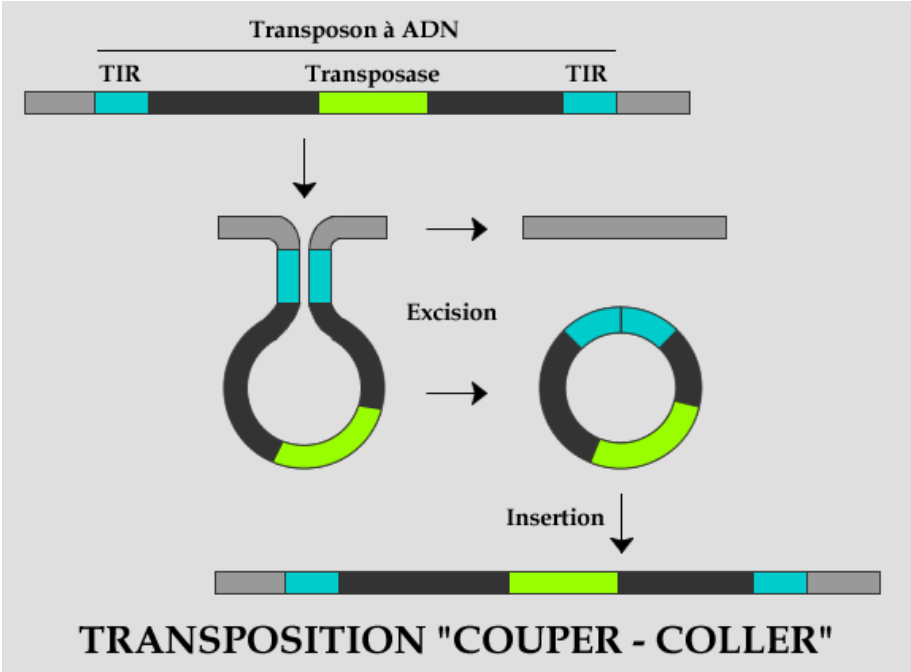
C - ELEMENTS TRANSPOSABLES

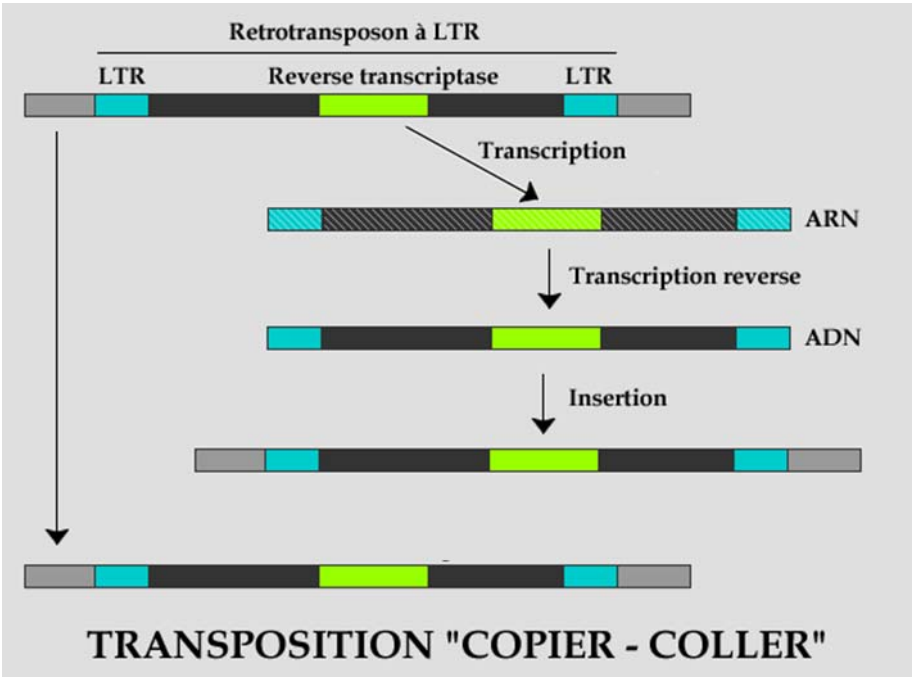
Il existe divers cas de régions génomiques qui, tout comme le locus S, se distinguent par une forte restriction de la recombinaison. Parmi ces régions, les chromosomes sexuels ont évolué plusieurs fois de manière indépendante dans différents groupes d'organismes, comme les plantes, les oiseaux ou les mammifères. Il s'agissait au début de leur évolution de chromosomes autosomaux (LAHN and PAGE 1999). Alors que le chromosome Y a progressivement cessé toute recombinaison à l'exception de petites régions pseudo-autosomales, le chromosome X a continué de recombiner chez les individus femelles. Le chromosome Y est donc un excellent cas d'étude d'une région privée de recombinaison. La comparaison des chromosomes X et Y a notamment révélé que, depuis l'arrêt de la recombinaison, le chromosome Y avait subi de nombreux réarrangements chromosomiques accompagnés d'une forte accumulation d'éléments transposables (BACHTROG 2005; SKALETSKY *et al.* 2005).

Ces éléments, découverts en 1950 par Barbara McClintock, sont des séquences d'ADN de taille variable. Capables de se déplacer et de se multiplier de manière autonome dans les génomes, ils sont omniprésents dans le monde vivant (WESSLER 2006) et sont l'un des composants majeurs des génomes eucaryotes, et en particulier de ceux des plantes (BENNETZEN 2000). On les y retrouve cependant dans des proportions très diverses. Ils représentent par exemple moins de 20 % du génome d'*A. thaliana* (THE *ARABIDOPSIS* GENOME INITIATIVE 2000), et près de 70 % du génome de *Zea mays* (SANMIGUEL *et al.* 1996).

1 - Classification

Les éléments transposables peuvent être divisés en deux grandes catégories (voir Wicker *et al.* (2007) pour une classification complète).





Encadré 3. Comparaison des modes de transposition des éléments transposables.

La plupart des transposons à ADN se transposent par un mécanisme de type « couper – coller ». Après avoir été excisés, ils s'insèrent à une nouvelle position du génome. Les retrotransposons se transposent quant à eux par un mécanisme de type « copier – coller ». Après transcription, leur ARN génère une nouvelle copie grâce à l'action d'une reverse transcriptase. Une copie supplémentaire du retrotransposon est donc créée à chaque transposition.

Abréviations : TIR - Terminal Inverted Repeat LTR - Long Terminal Repeat

- Les retrotransposons, ou éléments transposables de type I, se transposent par un mécanisme de « copier - coller » en utilisant un ARN comme intermédiaire (voir l'encadré 3 pour une description des modes de transposition). Ces éléments comprennent les retrotransposons à LTR (Long Terminal Repeats), caractérisés par des séquences répétées à leurs extrémités, et les retrotransposons sans LTR tels que les LINE ou les SINE.

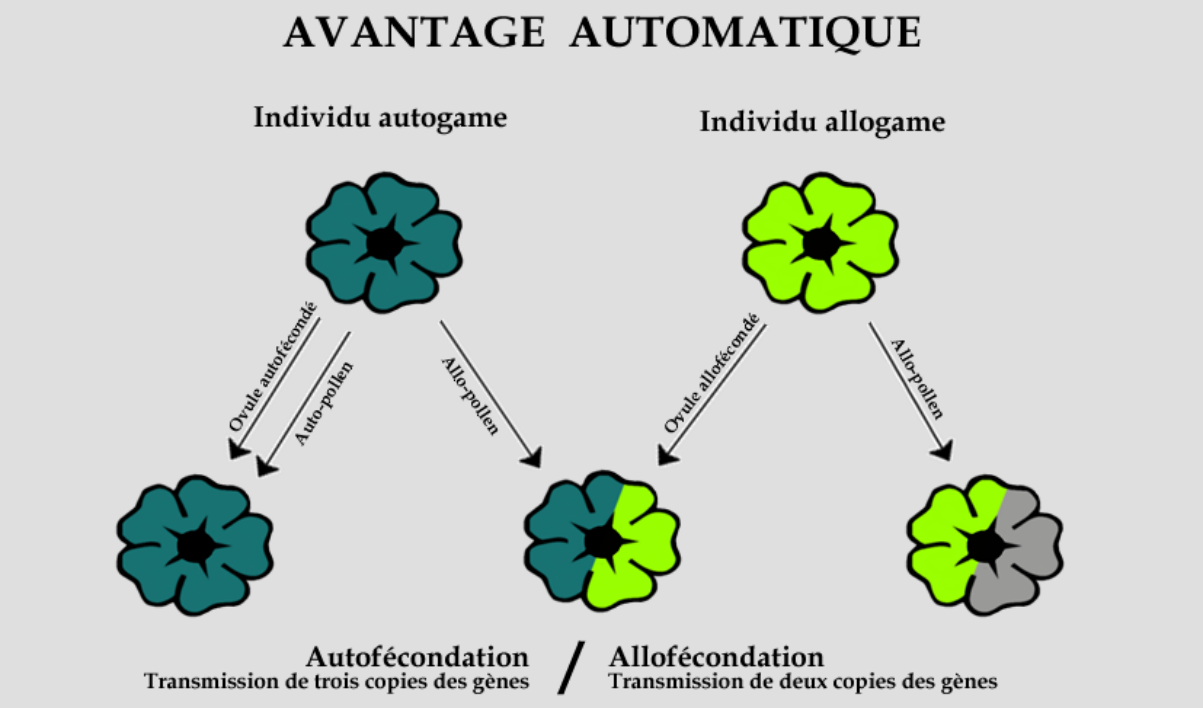
- Les transposons à ADN, ou éléments transposables de type II, rassemblent diverses familles telles que les Helitrons, les Mariners ou encore les Harbingers. A l'exception des Helitrons, ils se transposent par un mécanisme de « couper - coller ».

2 - Implications dans l'évolution des génomes

Par leur mode de transposition, les éléments transposables peuvent provoquer des mutations dans les génomes. S'ils s'insèrent dans la région codante d'un gène, ils peuvent par exemple interrompre sa traduction. Une insertion dans une région régulatrice peut quant à elle entraîner une modification de l'expression génique. A une échelle plus large, les éléments transposables peuvent également être responsables de réarrangements chromosomiques ou de variation dans la taille des génomes (KIDWELL 2002; VITTE and PANAUD 2005). On peut s'attendre à ce que les insertions délétères qu'ils peuvent provoquer soient fortement contre-sélectionnées, comme lorsqu'elles provoquent des maladies (VORECHOVSKY 2010). Des mécanismes épigénétiques entrent par ailleurs en compte pour limiter leur prolifération et les éléments transposables sont souvent ciblés par des petits ARN qui provoquent leur méthylation (ALMEIDA and ALLSHIRE 2005). Ce mécanisme serait moins fréquent pour les éléments transposables situés dans des régions riches en gènes, en raison de l'effet délétère de la méthylation de l'ADN sur leur expression (HOLLISTER and GAUT 2009).

En dépit des effets délétères que les éléments transposables peuvent représenter, ils possèdent un fort potentiel évolutif. Outre les réarrangements génomiques à grande échelle qui peuvent influencer l'évolution des espèces (KIM *et al.* 1998; LIM and SIMMONS 1994), leur mode de transposition peut par exemple entraîner des duplications de gènes qui pourront être suivies d'une diversification des duplicats. Par des mécanismes d' « exon shuffling » ou d'épissage alternatif, ils peuvent également conduire à l'émergence de nouvelles fonctions protéiques.

Au niveau du locus d'auto-incompatibilité, la présence d'éléments transposables a déjà été mentionnée dans différents taxa, que ce soit dans le cas d'un système gamétophytique, comme chez *Papaver rhoeas* (WHEELER *et al.* 2003), ou dans le cas d'un système sporophytique, comme chez *Ipomoea trifida* (TOMITA *et al.* 2004), *B. rapa*, *B. oleraceae* (FUJIMOTO *et al.* 2006), *B. napus* (CUI *et al.* 1999) ou *A. lyrata* (GUO *et al.* 2011). Leur abondance par rapport au reste du génome n'a cependant pas été documentée. On ne sait donc pas si la restriction de la recombinaison au locus S favorise leur accumulation, comme c'est le cas pour le chromosome Y (BACHTROG 2005; SKALETSKY *et al.* 2005). De



Encadré 4. Principe de l'avantage automatique (FISHER 1941 ; GOODWILLIE *et al.* 2005)

Dans le cas de l'autogamie, un individu peut transmettre ses gènes par la voie femelle (ovule) et la voie mâle (pollen) au travers de l'autofécondation, mais son pollen peut également féconder d'autres individus. Trois copies de ses gènes sont ainsi transmises. Au contraire, un individu allogame ne peut transmettre que deux copies de ses gènes. L'autogamie confère donc un avantage de 50 % en termes de transmission des gènes à la descendance.

plus, cette restriction n'étant pas équivalente au sein des différentes classes de dominance dans le cas d'un système sporophytique, des patrons différents pourraient y être observés.

III - RUPTURE DU SYSTEME D'AUTO-INCOMPATIBILITE

Malgré les bénéfices liés à l'allogamie en termes de réduction de la dépression de consanguinité, l'évolution d'un système d'auto-incompatibilité vers l'autogamie est l'une des transitions les plus fréquentes et les plus documentées chez les Angiospermes (BARRETT 2002). L'auto-compatibilité a en effet évolué de manière indépendante à de nombreuses reprises. Par exemple, une analyse phylogénétique d'un groupe d'espèces annuelles, *Leptosiphon* (Polemoniaceae), a montré l'existence de quatre transitions indépendantes de l'auto-incompatibilité vers ce système de reproduction au sein de ce genre (GOODWILLIE 1999).

A - SOUS QUELLES CONDITIONS ?

La transition vers l'auto-compatibilité n'est possible que lorsque les avantages liés à l'autofécondation sont supérieurs aux effets délétères de la dépression de consanguinité qui l'accompagne. Deux principaux éléments sont connus pour favoriser l'autogamie par rapport à l'allogamie. Tout d'abord, l'autofécondation confère une transmission des gènes plus importante que dans le cas de l'allogamie (voir l'encadré 4). C'est ce que l'on appelle l'avantage automatique (FISHER 1941; GOODWILLIE *et al.* 2005). L'autofécondation présente également des avantages indéniables en termes d'assurance reproductive (BAKER 1967). La possibilité d'assurer seul sa descendance peut en effet se révéler cruciale lorsque les partenaires potentiels ou les pollinisateurs sont rares, ainsi que dans le cas des plantes pionnières qui colonisent de nouveaux milieux (BUSCH and SCHOEN 2008). Par ailleurs, sous ces conditions qui favorisent l'autogamie, il a été montré qu'une mutation rompant le système d'auto-incompatibilité était plus à même de se répandre si elle touchait le déterminant mâle (UYENOYAMA *et al.* 2001). En effet, un individu porteur d'une mutation inactivant la fonction mâle peut échapper au contrôle de l'auto-incompatibilité et féconder un individu de même haplotype. Cet individu ne peut quant à lui pas féconder le mutant. A l'inverse, un individu porteur d'une mutation femelle ne peut pas féconder un individu de même haplotype mais peut être fécondé par ce dernier. Il subit donc un désavantage non réciproque.

B - CONSEQUENCES EVOLUTIVES

Chez une espèce ayant expérimenté une transition récente vers l'autogamie, l'effet de la dépression de consanguinité tend à diminuer avec le temps. En effet, si l'augmentation de l'homozygotie entraîne l'expression de mutations délétères récessives, la sélection naturelle sera plus apte à les éliminer. La purge de ces mutations étant donc favorisée (LANDE and SCHEMSKE 1985), le retour à un état allogame à partir de l'autogamie est très improbable car les avantages de l'allogamie en termes de diminution de la dépression de consanguinité sont alors amoindris. A plus long terme, il est supposé que la

ARABIDOPSIS THALIANA

Nom commun :	Arabette des Dames
Répartition :	Mondiale
Cycle :	Annuelle
Système de reproduction sexuée :	Auto-compatible
Nombre de chromosomes :	2n = 10
Signes particuliers :	Organisme modèle Espèce pionnière

ARABIDOPSIS LYRATA

Nom commun :	Arabette lyrée
Répartition :	Majeure partie de l'Europe, Russie, Amérique du Nord
Cycle :	Pérenne
Système de reproduction sexuée :	Auto-incompatible
Nombre de chromosomes :	2n = 16
Signes particuliers :	Génome séquencé

ARABIDOPSIS HALLERI

Nom commun :	Arabette de Haller
Répartition :	Majeure partie de l'Europe centrale et de l'Europe de l'Est
Cycle :	Pérenne
Système de reproduction sexuée :	Auto-incompatible
Nombre de chromosomes :	2n = 16
Signes particuliers :	Pseudo-métallophyte

Encadré 5. Caractéristiques des trois espèces étudiées.

réduction de taille efficace entraîne une augmentation de la dérive génétique et rende la sélection moins efficace, suggérant que l'autogamie est une impasse évolutive (TAKEBAYASHI and MORRELL 2001). On s'attend en outre à ce que l'efficacité amoindrie de la sélection naturelle à éliminer les mutations délétères entraîne un ratio dN/dS (rapport du taux de substitution non-synonyme sur le taux de substitution synonyme) plus élevé chez les espèces autogames que chez les espèces allogames qui leur sont proches (GLEMIN 2007). Cette différence n'étant cependant pas observée lorsqu'un grand nombre d'espèces est analysé (GLEMIN *et al.* 2006), il peut être supposé que les transitions de l'allogamie vers l'autogamie sont récentes, comme c'est le cas par exemple chez *A. thaliana* (BECHSGAARD *et al.* 2006; CHARLESWORTH and VEKEMANS 2005), et que par conséquent les espèces autogames ne se maintiennent pas dans le temps. De plus, il a récemment été démontré dans la famille des Solanaceae que le taux d'extinction des espèces auto-compatibles autogames était supérieur à celui des espèces possédant un système d'auto-incompatibilité gamétophytique (GOLDBERG *et al.* 2010).

IV - *ARABIDOPSIS* COMME MODELE D'ETUDE

Arabidopsis est un genre de la famille des Brassicaceae comprenant une dizaine d'espèces originaires d'Eurasie (O'KANE and AL-SHEHBAZ 1997). On s'intéresse ici à trois de ces espèces : *A. thaliana, A. lyrata* et *A. halleri* (voir l'encadré 5).

A - *ARABIDOPSIS LYRATA*

L'arabette lyrée est une proche parente de l'espèce modèle *A. thaliana* (KOCH *et al.* 1999). Elle est également proche du groupe d'espèces *Brassica*, utilisé comme modèle de l'auto-incompatibilité sporophytique (FUJIMOTO and NISHIO 2007), et de nombreux outils moléculaires définis chez *Brassica* peuvent être transférés chez *Arabidopsis*. Ainsi, si les espèces étudiées du genre *Brassica* sont des espèces cultivées et ont par conséquent subi l'influence de la domestication, *A. lyrata* s'est imposée depuis quelques années comme l'espèce modèle de l'auto-incompatibilité sporophytique en populations naturelles (BOGGS *et al.* 2009a; GUO *et al.* 2011; KUSABA *et al.* 2001). De plus, son génome a récemment été entièrement séquencé (HU *et al.* 2011), et des comparaisons entre le locus d'auto-incompatibilité et le reste du génome peuvent par conséquent être envisagées.

B - *ARABIDOPSIS HALLERI*

Découverte en Allemagne au XVIIIe siècle, l'Arabette de Haller est principalement étudiée pour sa capacité à tolérer et accumuler des métaux lourds tels que le zinc ou le cadmium (BERT 2000; KRÄMER 2010; PAUWELS *et al.* 2006). Elle est particulièrement proche d'*A. lyrata*, dont elle aurait divergé il y a environ 350 000 ans (ROUX *et al.* 2011), contrairement aux estimations précédentes qui dataient leur divergence à environ deux millions d'années (KOCH and MATSCHINGER 2007). Les deux espèces sœurs possèdent le même système d'auto-incompatibilité sporophytique et présentent un polymorphisme

trans-spécifique (CASTRIC *et al.* 2008), c'est-à-dire qu'elles partagent une partie de leurs haplotypes d'auto-incompatibilité.

C - *ARABIDOPSIS THALIANA*

L'arabette des dames est la première plante supérieure à avoir vu son génome entièrement séquencé (THE *ARABIDOPSIS* GENOME INITIATIVE 2000). Cette espèce annuelle présente en effet de nombreux intérêts dont un cycle de développement court, un nombre de graines important et une taille de génome relativement petite dans le monde végétal (BENNETT *et al.* 2003). Contrairement à *A. lyrata* et *A. halleri* dont elle a divergé il y a environ trois à six millions d'années (CLAUSS and KOCH 2006), *A. thaliana* a perdu son système d'auto-incompatibilité. Cette transition vers l'autogamie, relativement récente (BECHSGAARD *et al.* 2006; CHARLESWORTH and VEKEMANS 2005), se traduit notamment par le fait que seuls trois haplogroupes du locus d'auto-incompatibilité inactivé ont été maintenus chez cette espèce : l'haplogroupe *B*, dont la distribution est restreinte aux îles situées autour du Cap Vert, et les haplogroupes *A* et *C*, que l'on trouve largement répartis en Europe, Asie, Amérique du Nord et Afrique du Nord. La rupture du système d'auto-incompatibilité pourrait être en relation avec la recolonisation de l'Europe au Pléistocène, il y a de cela environ 30 000 ans, à partir des refuges glaciaires asiatique et méditerranéen (SHERMAN-BROYLES *et al.* 2007). Selon cette hypothèse, l'haplogroupe *B* aurait été maintenu sur son aire de répartition actuelle. L'haplogroupe *A* aurait quant à lui été abrité dans le refuge glaciaire asiatique et l'haplogroupe *C* dans le refuge méditerranéen. Plusieurs mutations d'inactivation du locus d'auto-incompatibilité auraient alors eu lieu indépendamment, soit dans les refuges, favorisées par de petites tailles de populations, soit pour les haplogroupes *A* et *C* pendant la recolonisation, avec un avantage sélectif fort des individus capables de s'autoféconder. Des évènements de recombinaison entre les haplogroupes *A* et *C* auraient en outre été rendus possibles dans les zones de contact entre leurs aires de répartition.

Parce qu'*A. thaliana* est une espèce modèle en génétique, de nombreuses données sont disponibles à son sujet. En particulier, le projet 1001 génomes (WEIGEL and MOTT 2009) vise à décrire la diversité de l'espèce à l'échelle du génome entier au travers du séquençage d'un millier d'accessions. Ces données constituent un atout majeur dans la compréhension de la perte du système d'auto-incompatibilité. Cependant, en raison de la forte divergence du locus S, l'assemblage ne permet pas d'étudier directement cette région génomique car il est en partie basé sur la séquence de référence, qui ne représente qu'un haplogroupe particulier. Des assemblages *de novo* ou basés sur les différents haplogroupes identifiés seraient donc nécessaires afin d'étudier la diversité du locus d'auto-incompatibilité chez ce grand nombre d'accessions.

V - OBJECTIFS DE LA THESE

Dans le contexte décrit précédemment, cette thèse consiste en une analyse bioinformatique du locus d'auto-incompatibilité sporophytique chez les Brassicaceae, et plus précisément dans le genre *Arabidopsis*. A travers l'obtention et l'analyse d'une douzaine d'haplotypes du locus S chez *A. halleri, A. lyrata* et *A. thaliana*, elle permet en particulier d'explorer les apports et les limites des approches génomiques dans l'étude du système d'auto-incompatibilité

A - INTERETS DE LA METHODE GENOMIQUE

La majorité des études concernant le système d'auto-incompatibilité sont basées sur des approches destinées à l'analyse de fragments d'ADN très ciblés et de petite taille (pouvant aller jusqu'à quelques kilobases). Ces approches ne s'appuient donc pas sur la totalité du locus S, et la diversité de la région génomique qui le comprend est encore très mal connue. Elles supposent en outre de connaître au moins partiellement la séquence étudiée. Ainsi, dans le genre *Arabidopsis*, si le polymorphisme du déterminant femelle *SRK* a été largement étudié (CASTRIC and VEKEMANS 2007; MABLE *et al.* 2003), peu de séquences du déterminant mâle *SCR* ont pu être obtenues (BOGGS *et al.* 2009a; BOGGS *et al.* 2009b; GUO *et al.* 2011; KUSABA *et al.* 2001; TSUCHIMATSU *et al.* 2010). Ce gène est en effet fortement divergent, et ne comprend pas de régions qui soient assez conservées pour ancrer des amorces et effectuer des PCR (Polymerase Chain Reaction). La méthode génomique consistant à séquencer la totalité du locus d'auto-incompatibilité a donc l'intérêt de permettre d'accéder à la fois à l'environnement génomique du système et au gène pollen *SCR*.

De manière générale, les données de séquences comprenant la région génomique complète du locus d'auto-incompatibilité sont relativement restreintes dans la littérature. Dans le cadre d'un système gamétophytique, six séquences génomiques sont par exemple disponibles chez *Antirrhinum hispanicum*, mais seule une de ces séquences comprend à la fois le gène codant la protéine du pistil, une S-RNase, et les gènes codant l'ensemble de protéines SLF (S-Locus F-box) qui contrôlent la spécificité du pollen (ZHOU *et al.* 2003). Une séquence comprenant une partie du locus d'auto-incompatibilité est également disponible chez *Papaver rhoeas* (WHEELER *et al.* 2003). Dans le cadre d'un système sporophytique, un haplotype du locus d'auto-incompatibilité est disponible chez *Ipomoea trifida* (TOMITA *et al.* 2004) mais n'a pas encore permis de mettre en évidence les gènes contrôlant le système. Chez les Brassicaceae, douze haplotypes d'auto-incompatibilité, répartis dans les deux classes de dominance, sont disponibles chez les espèces cultivées du genre *Brassica* (CUI *et al.* 1999; FUJIMOTO *et al.* 2006; FUKAI *et al.* 2003; KIMURA *et al.* 2002; SHIBA *et al.* 2003; SUZUKI *et al.* 1999; TAKUNO *et al.* 2007). Néanmoins, certaines de ces séquences ne comprennent pas les régions flanquantes du locus S, et il est possible que la domestication de ces espèces ait influencé la diversité observée. Dans le genre *Arabidopsis*, deux haplotypes parmi les trois identifiés chez *A. thaliana* ont été séquencés (TANG *et al.* 2007; THE *ARABIDOPSIS* GENOME INITIATIVE 2000), ainsi qu'un recombinant entre deux de ces

haplotypes (SHERMAN-BROYLES *et al.* 2007). Enfin, sept haplotypes, dont deux qui seraient non-fonctionnels, ont été séquencés chez *A. lyrata* (BOGGS *et al.* 2009a; GUO *et al.* 2011; KUSABA *et al.* 2001). Le nombre d'haplotypes fonctionnels provenant de populations naturelles est donc très limité dans le groupe d'espèces *Arabidopsis*, à savoir cinq haplotypes, et ne couvre pas les quatre classes de dominance. Les données de cette thèse comprennent onze haplotypes fonctionnels, répartis dans les quatre classes de dominance, chez *A. lyrata* et *A. halleri* ainsi que le troisième haplotype identifié chez *A. thaliana*. Leur annotation et leur analyse ont permis d'aborder diverses pistes d'étude déclinées en trois chapitres.

B - ANALYSE DES PATRONS D'EVOLUTION MOLECULAIRE DU LOCUS D'AUTO-INCOMPATIBILITE DANS LE GENRE *ARABIDOPSIS*

A travers l'analyse d'une douzaine d'haplotypes S fonctionnels chez *A. lyrata* et *A. halleri*, le premier chapitre s'intéresse aux patrons d'évolution moléculaire d'un locus d'auto-incompatibilité fonctionnel. En effet, si le système d'auto-incompatibilité en lui-même est relativement bien décrit chez les Brassicaceae, peu de données relatives à la diversité et la dynamique de sa région génomique sont à ce jour disponibles. Le jeu de données présenté dans cette partie a permis dans un premier temps de délimiter avec précision le locus d'auto-incompatibilité, et d'analyser dans quelle mesure les patrons d'évolution moléculaire sont contrastés entre le locus S et ses régions flanquantes. Pour ce faire, la conservation de la séquence entre les différents haplotypes et les phylogénies des gènes de la région du locus d'auto-incompatibilité a été étudiée. Dans un second temps, la question de la diversité structurale du locus d'auto-incompatibilité a été posée, que ce soit au niveau de la taille des haplotypes, de l'organisation de leurs gènes ou de leur contenu en éléments transposables. Cette diversité structurale a en outre été mise en rapport avec le caractère dominant ou récessif des différents haplotypes. Enfin, la méthode génomique permettant d'obtenir un nombre important de couples *SCR-SRK*, les patrons de coévolution de ces deux gènes ont été abordés.

C - RUPTURE DU SYSTEME D'AUTO-INCOMPATIBILITE CHEZ *ARABIDOPSIS THALIANA*

Le second chapitre se concentre quant à lui sur la perte du système d'auto-incompatibilité chez *A. thaliana*. Chez cette espèce, le locus d'auto-incompatibilité a été inactivé, et seules trois spécificités témoignent du polymorphisme ancestral. Le séquençage de la seule de ces spécificités dont la séquence n'était pas encore connue a permis de poser la question de la rupture du système d'auto-incompatibilité dans cet haplogroupe. De manière plus générale, l'obtention de cette séquence a permis d'étudier la répartition des différentes spécificités en populations naturelles en s'appuyant sur les données du projet 1001 génomes. Des réarrangements génomiques et des évènements de recombinaison entre haplogroupes différents ont ainsi pu être identifiés, et un scénario décrivant les patrons de dégénérescence du locus d'auto-incompatibilité d'*A. thaliana* après inactivation a été construit et analysé.

D - ANALYSE PRELIMINAIRE DE LA COEVOLUTION ENTRE LES PROTEINES DU PISTIL ET DU POLLEN

Parce qu'ils présentent une reconnaissance de type clé-serrure, les protéines du pollen et du pistil ont besoin d'évoluer de manière conjointe, dans le cadre du maintien de cette reconnaissance ou de l'évolution vers une nouvelle spécificité. Ils constituent par conséquent un modèle pertinent pour une analyse de la coévolution. Dans cette troisième partie, une analyse préliminaire des patrons de coévolution des protéines SCR et SRK a été effectuée grâce à une méthode basée sur des cartes de substitution et visant à détecter des groupes de sites se trouvant potentiellement en situation de coévolution (DUTHEIL and GALTIER 2007; DUTHEIL *et al.* 2005). Ainsi, un signal fort de coévolution a pu être mis en évidence entre la protéine du pollen et le domaine extracellulaire de la protéine du pistil.

Encadré 6. Photographie d'une plante de l'espèce *A. halleri* en culture à la serre.

ACQUISITION DES DONNEES

Dans le cadre de cette thèse, un jeu de séquences constitué de sept haplotypes d'*A. halleri*, quatre haplotypes d'*A. lyrata* et un haplotype d'*A. thaliana* a été séquencé, annoté et étudié. L'acquisition de ces données, entreprise avant le début de la thèse par l'équipe d'auto-incompatibilité, a été faite en collaboration avec le Génoscope et le Centre National de Ressources de Génomique Végétale de l'INRA (CNRGV).

I - DE LA PLANTE A LA SEQUENCE

A - MATERIEL VEGETAL

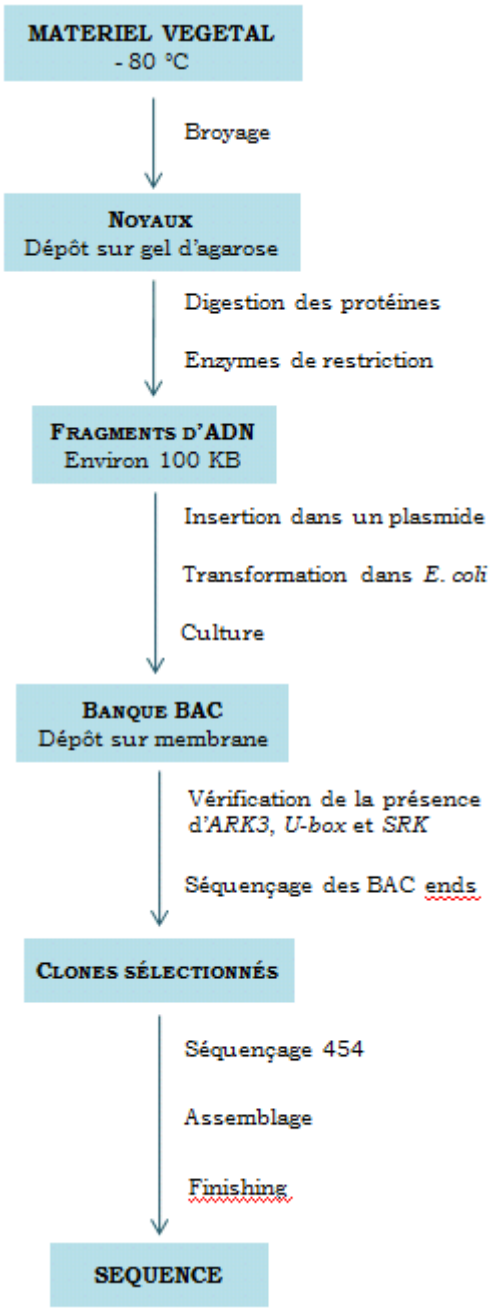
Sur la base de leur génotype d'auto-incompatibilité, des graines d'*A. thaliana*, *A. lyrata* et *A. halleri* ont été sélectionnées afin de constituer des banques BAC et de permettre le séquençage de la région génomique du locus S. Ces graines sont originaires de divers pays : Belgique, France, Islande, Italie, Maroc et Pologne. Elles ont été semées en serre et, après culture, ont été bouturées afin d'obtenir les 100 g de jeunes feuilles nécessaires à la constitution des banques (encadré 6). Environ six mois après le semis, les jeunes plantes ont été placées cinq jours à l'obscurité, afin que l'amidon soit digéré et n'interfère pas lors de l'isolement des noyaux, et les feuilles ont ensuite été récoltées. Les étapes principales pour obtenir la séquence de la région génomique du locus d'auto-incompatibilité à partir de ce matériel végétal sont décrites ci-dessous et schématisées dans l'encadré 7.

B - OBTENTION DES BANQUES

Les banques BACs (Bacterial Artificial Chromosome) ont été constituées par le CNRGV. Pour ce faire, le matériel végétal a été congelé à -80°C, avant d'être broyé. Les noyaux restés intacts ont ensuite été isolés et placés dans du gel d'agarose. Toutes les protéines ont alors été digérées par des enzymes afin de ne laisser que l'ADN. Cet ADN a été mis en présence d'enzymes de restriction qui l'ont coupé en fragments d'environ 100 kb. Les fragments obtenus ont été insérés dans des plasmides puis transformés chez *Escherichia coli*. Après culture, les colonies d'*E. coli* contenant chacune un fragment d'ADN ont été placées sur une plaque, ce qui constitue la banque BAC.

C - CRIBLAGE DES BANQUES ET VALIDATION DES CLONES POSITIFS

Les banques BACs ont été déposées sur des membranes qui ont été hybridées avec les sondes des gènes recherchés, en l'occurrence *ARK3* et la *U-box* qui bordent le locus S, ainsi que *SRK*. La présence de ces gènes a ensuite été vérifiée par une PCR classique pour *ARK3* et la *U-box*, et une PCR allèle-spécifique pour *SRK* en raison de la forte divergence de ses allèles.



Encadré 7. Succession d'étapes ayant permis d'obtenir la séquence de la région génomique du locus d'auto-incompatibilité, à partir de feuilles d'*Arabidopsis*.

Les BAC ends, c'est-à-dire les extrémités de séquences, de 96 clones montrant la présence d'au moins un des gènes en question ont été séquencés par la compagnie Eurofins / MWG. Ces courtes séquences (500 à 900 bp) ont alors pu être alignées sur le génome d'*A. lyrata* et *A. thaliana* afin d'évaluer leur proximité au locus S, et d'en déduire la longueur de la séquence totale des clones. Pour chaque haplotype choisi, les clones les plus longs et les plus centrés sur le locus S ont ainsi pu être sélectionnés pour être séquencés. Il est à noter que deux clones ont été nécessaires dans trois cas afin de couvrir la totalité du locus S. D'autre part, le locus S étant une région fortement divergente, certains BAC ends n'ont pu être alignés, et les limites précises de trois clones séquencés n'étaient donc pas connues.

D - SEQUENÇAGE ET ASSEMBLAGE

Douze clones couvrant neuf haplotypes ont été séquencés par le Génoscope. Les données brutes du séquençage 454 ont été assemblées, selon une méthode d'assemblage *de novo*, par le Génoscope à l'aide du programme NEWBLER (www.roche.com) en des séquences totalisant de 84 à 115 kb selon les haplotypes. Cet assemblage n'a cependant pas pu reconstituer des séquences complètes, probablement en raison de la présence de longues séquences répétées. Les séquences ainsi obtenues sont donc constituées de deux à neuf contigs, avec dans certains cas une suggestion d'orientation des différents contigs fournie par l'assemblage. Cependant, dans la plupart des cas, seuls les contigs extrêmes sont orientés, et on ne connait ni l'orientation, ni l'ordre des contigs intermédiaires.

Deux clones ont quant à eux été séquencés par le CNRGV, en collaboration avec les Plateformes Génomique et Bioinformatique de Toulouse, et un dernier clone au laboratoire de Reproduction et Développement des Plantes (ENS Lyon). Trois séquences supplémentaires, de 90 à 118 kb, et constituées de trois à huit contigs non orientés, ont ainsi pu être ajoutées aux données.

Les données obtenues après séquençage et assemblage sont décrites dans le tableau 1.

E - FINISHING DES SEQUENCES

Dans quatre cas, les exons d'un même gène (*SCR* ou *SRK*) se trouvaient sur deux contigs différents : *AlSRK01, AhSRK15, AlSCR39* et *AhSCR03*. Des amorces ont alors été définies aux extrémités des deux contigs concernés afin de réaliser une PCR (Polymerase Chain Reaction). Une amplification ayant été observée dans chaque cas, il a ainsi été vérifié que les contigs portant les différents exons du gène étaient bien consécutifs, et que le gène n'avait par conséquent pas subi de réarrangement.

L'organisation des gènes du locus d'auto-incompatibilité étant un point important de notre étude, la méthode de finishing a également été appliquée aux contigs portant les gènes *AlSCR14* et *AhSCR20.* Des PCR ont donc été réalisées en définissant des amorces aux extrémités des contigs pour chaque hypothèse d'orientation. Néanmoins, en raison de la présence de séquences répétées aux extrémités

Tableau 1. Description des clones BAC séquencés et caractéristiques des séquences obtenues.

Espèce	Banque	Provenance	BAC	Taille (bp)	Nombre de lectures	Taille moyenne des lectures	Couverture	Contigs	Lieu de séquençage
A. lyrata	05B17	Islande	Al01	100 937	10 900	331	35.74	4	Génoscope
A. lyrata	05B17	Islande	Al14	117 539	54 361	344	159.09	3	ENS Lyon
A. lyrata	05B37	Islande	Al18 BAC1	88 062	15 066	372	63.64	3	Génoscope
A. lyrata	05B37	Islande	Al18 BAC2	96 061	5 126	346	18.46	4	Génoscope
A. halleri	I9	Italie	Ah03	84 197	6 848	359	29.20	8	Génoscope
A. halleri	HF11	Belgique	Ah28 BAC1	101 609	9 569	362	34.09	2	Génoscope
A. halleri	HF11	Belgique	Ah28 BAC2	94 078	18 635	349	69.13	3	Génoscope
A. lyrata	05B8	Islande	Al39	90 060	17 367	376	75.22	5	CNRGV
A. halleri	HF11	Belgique	Ah13	88 292	11 931	348	47.03	5	Génoscope
A. halleri	L406	France	Ah15 BAC1	109 343	8 877	351	28.50	9	Génoscope
A. halleri	L406	France	Ah15 BAC2	85 357	10 636	345	42.99	4	Génoscope
A. halleri	L406	France	Ah20	105 142	27 578	350	91.80	5	Génoscope
A. halleri	PL22	Pologne	Ah32	115 243	15 378	351	46.84	5	Génoscope
A. halleri	I9	Italie	Ah43	95 096	12 656	378	47.54	8	CNRGV
A. thaliana	Ita-0	Maroc	HapC	116959	7789	333	22.18	5	Génoscope

des différents contigs, les PCR classiques se sont révélées inefficaces dans certains cas. Ce sont alors des PCR longs fragments qui ont été effectuées, avec des amorces ancrées dans des régions ne contenant pas de répétitions ou d'éléments transposables. L'incertitude sur l'orientation de ces gènes par rapport aux autres gènes du locus d'incompatibilité a ainsi pu être levée. Cette méthode n'a cependant pas pu être appliquée aux haplotypes *Ah32* et *Ah43*, en raison d'un nombre trop important d'hypothèses d'orientation à tester.

Dans un dernier temps, la même méthode a été appliquée à l'haplotype *C* d'*A. thaliana*, afin de tester l'ordre et l'orientation de ses contigs. La suggestion d'orientation fournie avec l'assemblage a ainsi été confirmée.

II - ANNOTATION DES SEQUENCES

A - GENES

L'annotation de la région génomique du locus S étant une part essentielle de la thèse, mon travail a commencé, après avoir appris à maîtriser les outils le permettant, par effectuer une nouvelle annotation des séquences connues. Cette étape m'a permis de vérifier la pertinence de mes résultats en les comparant à une annotation déjà réalisée. J'ai ainsi refait l'annotation des haplotypes connus et disponibles chez *A. thaliana*, *A. lyrata*, *B. rapa* et *B. oleracea*. Deux programmes de prédiction statistique ont pour cela été utilisés : FGENESH (SALAMOV and SOLOVYEV 2000) et GENSCAN (BURGE and KARLIN 1997). D'un point de vue général, FGENESH tend à être plus précis et GENSCAN plus sensible dans la détection de gènes chez *Arabidopsis*. Leur association permet donc de détecter la majorité des gènes. Les ORF (Open Reading Frame) détectés par ces programmes ont ensuite été blastés à l'aide de BLASTX (GISH and STATES 1993) ou de NUCLEOTIDE BLAST (ALTSCHUL *et al.* 1990). Les protéines ou les ARN_m ainsi trouvés ont alors été alignés sur les séquences génomiques à l'aide de divers programmes : SPALN (GOTOH 2008), FGENESH+ (SALAMOV and SOLOVYEV 2000) et EST2GENOME (MOTT 1997). L'annotation des gènes a été déduite de la comparaison des résultats de ces différents programmes.

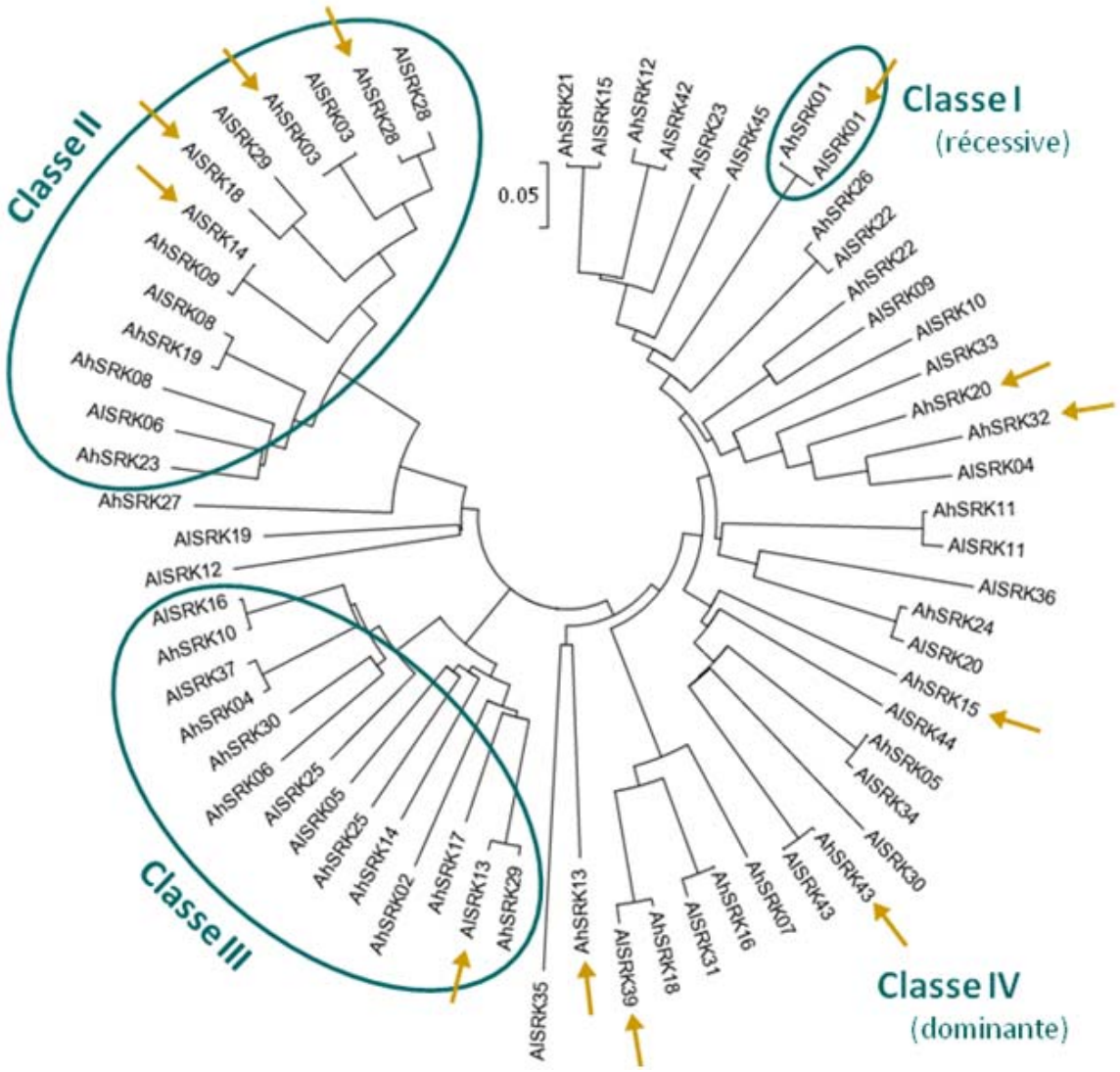
Les gènes du locus d'auto-incompatibilité étant maintenus sur de longues périodes de temps, ils sont caractérisés par une extrême divergence de leurs allèles (IOERGER *et al.* 1990). De plus, le gène pollen *SCR* se distingue par ses deux exons de petite taille (approximativement 70 et 180 pb) séparés par un intron dont la taille peut atteindre plusieurs kilobases. De par cette configuration, ce gène n'est que très rarement détecté par la plupart des programmes de prédiction statistique. Les protéines SCR connues ont donc été comparées aux séquences génomiques avec ALN (GOTOH 1982), un programme d'alignement particulièrement sensible. Les résultats de ce programme ont ensuite été analysés afin de localiser le gène *SCR*, en partie grâce aux résidus de cystéine caractéristiques de sa protéine. Cette

procédure nous a notamment permis d'identifier un second exon fonctionnel pour deux haplotypes séquencés récemment chez *A. lyrata* (GUO *et al.* 2011), alors que ces haplotypes étaient considérés comme non-fonctionnels par défaut d'identification de cet exon.

B - ELEMENTS TRANSPOSABLES

Les éléments transposables ont été annotés à l'aide du programme CENSOR (KOHANY *et al.* 2006), utilisant la base de données REPBASE (JURKA *et al.* 2005). Les résultats obtenus ont ensuite été filtrés grâce à PLOTREP (TOTH *et al.* 2006), un programme qui permet également de réunir les différents fragments d'un même élément transposable qui aurait été morcelé. Les zones de faible complexité et les répétitions simples ont quant à elles été détectées par le programme REPEATMASKER (SMIT *et al.* 1996-2006).

Tout au long des phases d'annotation et d'analyse des séquences, le programme ARGO genome (www.broadinstitute.org/annotation/argo) a été utilisé afin de visualiser les différentes couches d'annotation.



Encadré 8. Phylogénie des allèles *SRK* chez *A. halleri* et *A. lyrata* (CASTRIC *et al.* 2008).

Les allèles analysés dans ce chapitre sont indiqués par une flèche.

CHAPITRE I - ANALYSE DES PATRONS D'EVOLUTION MOLECULAIRE

Des données génomiques de la région du locus S chez les Brassicaceae sont disponibles essentiellement pour les espèces cultivées du genre *Brassica* qui ont servi de modèles d'étude de l'auto-incompatibilité : *B. rapa*, *B. napus* et *B. oleracea* (CUI *et al.* 1999; FUJIMOTO *et al.* 2006; FUKAI *et al.* 2003; KIMURA *et al.* 2002; SHIBA *et al.* 2003; SUZUKI *et al.* 1999; TAKUNO *et al.* 2007). Plus récemment, le genre *Arabidopsis* a émergé comme modèle sur ce sujet (BOMBLIES and WEIGEL 2010), permettant notamment des avancées sur le polymorphisme de ce locus en populations naturelles (SCHIERUP *et al.* 2008). Par ailleurs, ce modèle présente l'intérêt de posséder un plus grand nombre de classes de dominance que *Brassica* (NASRALLAH *et al.* 1991; PRIGODA *et al.* 2005), multipliant ainsi le nombre d'interactions deux à deux possibles entre allèles dominants et récessifs. Cependant, peu de données sont disponibles dans la littérature en ce qui concerne cette région génomique chez *A. lyrata* et *A. halleri* (BOGGS *et al.* 2009a; GUO *et al.* 2011; KUSABA *et al.* 2001).

Par le séquençage de onze haplotypes du locus d'auto-incompatibilité chez *A. lyrata* et *A. halleri*, cette partie propose l'étude d'un grand jeu de données (Encadré 8), incluant l'haplotype provenant du séquençage de l'espèce *A. lyrata* (HU *et al.* 2011). Les patrons d'évolution moléculaire du locus d'auto-incompatibilité et de ses régions flanquantes y sont étudiés à travers notamment l'analyse de la densité en gènes, des phylogénies, de l'orientation respective et de la distance séparant les gènes, ainsi que de la teneur en éléments transposables. Le locus S y est en outre comparé à d'autres régions génomiques impliquées dans le déterminisme du sexe : les chromosomes sexuels chez les plantes ou les animaux et le locus de type sexuel chez les algues ou les champignons.

Ce chapitre fait l'objet d'un article accepté dans le journal PLoS Genetics : « Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis » par Pauline Goubet, Hélène Bergès, Arnaud Bellec, Elisa Prat, Nicolas Helmstetter, Sophie Gallina, Anne-Catherine Holl, Isabelle Fobis-Loisy, Xavier Vekemans et Vincent Castric.

TITLE: CONTRASTED PATTERNS OF MOLECULAR EVOLUTION IN DOMINANT AND RECESSIVE SELF-INCOMPATIBILITY HAPLOTYPES IN *ARABIDOPSIS*

SHORT TITLE: MOLECULAR EVOLUTION OF THE S-LOCUS REGION

Authors: Pauline Goubet[1], Hélène Bergès[2], Arnaud Bellec[2], Elisa Prat[2], Nicolas Helmstetter[2], Sophie Mangenot[3], Sophie Gallina[1], Anne-Catherine Holl[1], Isabelle Fobis-Loisy[4], Xavier Vekemans[1], Vincent Castric[1*]

Affiliations:

[1] Laboratoire de Génétique et Evolution des Populations Végétales, CNRS FRE 3268, Bâtiment SN2, Université Lille Nord de France, Cité scientifique, F-59655 Villeneuve d'Ascq Cedex, France

[2] Centre National des Ressources Génomiques Végétales, INRA UPR 1258, Chemin de Borde-Rouge, BP 52627, F-31326 Castanet Tolosan Cedex, France

[3] Genoscope, Commissariat à l'Energie Atomique (CEA), Direction des Sciences du Vivant, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057 Evry Cedex, Evry cedex F-91006, France

[4] 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, 46 allée d'Italie, F-69364 Lyon Cedex 07, France

Target journal: PLoS Genetics

* Corresponding author

I - ABSTRACT

Self-incompatibility has been considered by geneticists a model system for reproductive biology and balancing selection, but our understanding of the genetic basis and evolution of this molecular lock-and-key system has remained limited by the extreme level of sequence divergence among haplotypes, resulting in a lack of appropriate genomic sequences. In particular, the "key" partner of the system has proven especially challenging to sequence. In this manuscript, we report and analyze the full sequence of eleven distinct haplotypes of the self-incompatibility locus (S-locus) in two closely related *Arabidopsis* species, obtained from individual BAC libraries. We use this extensive dataset to highlight sharply contrasted patterns of molecular evolution of each of the two genes controlling self-incompatibility themselves (the lock and the key) as well as of the genomic region surrounding them. We find strong collinearity of the flanking regions on each side of the S-locus together with high levels of sequence similarity. In contrast, the S-locus region itself shows spectacularly deep gene genealogies, high variability in size and gene organization, as well as complete absence of sequence similarity in intergenic sequences and striking accumulation of transposable elements. Of particular interest, we demonstrate that dominant and recessive S-haplotypes experience sharply contrasted patterns of molecular evolution. Indeed, dominant haplotypes exhibit larger size and a much higher density of transposable elements, being matched only by that in the centromere. Overall, these properties highlight that the S-locus presents many striking similarities with other regions involved in the determination of mating-types, such as sex chromosomes in animals or in plants, or the mating-type locus in fungi and green algae.

II - AUTHOR SUMMARY

Self-incompatibility is a common genetic system preventing selfing through recognition and rejection of self pollen in hermaphroditic flowering plants. In the Brassicaceae family, this system is controlled by a single genomic region, called the S-locus, where many distinct specificities segregate in natural populations. In this study, we obtained genomic sequences comprising the S-locus in two closely related Brassicaceae species, *Arabidopsis lyrata* and *A. halleri* and we analyzed their diversity and patterns of molecular evolution. We report compelling evidence that the S-locus presents many similar properties with other genomic regions involved in the determination of mating-types in mammals, insects, plants or fungi. In particular, in spite of their diversity, these genomic regions all show absence of similarity in intergenic sequences, large depth of genealogies, highly divergent organization and accumulation of transposable elements. Moreover, some of these features were found to vary according to dominance of the S-locus specificities, suggesting that dominance/recessivity interactions are key drivers of the evolution of this genomic region.

III - INTRODUCTION

Sexual reproduction entails the combination of genetic material from different individuals to produce offspring. Yet in many species mating is not entirely random, being only possible between individuals with either distinct sexes or distinct mating types (BILLIARD *et al.* 2011). Sexes or mating types are typically determined by very distinctive genomic tracts known as sex chromosomes in animals (BACHTROG 2005; ROSS *et al.* 2005) and plants (MARAIS *et al.* 2008; YU *et al.* 2007), sex-determining loci in honeybees (HASSELMANN and BEYE 2006), mating-type loci in green algae (FERRIS *et al.* 2010; GOODENOUGH *et al.* 1995) and fungi (HOOD *et al.* 2004; LEE *et al.* 1999; MENKIS *et al.* 2008; WHITTLE and JOHANNESSON 2011) or self-incompatibility (SI) loci in plants (BILLIARD *et al.* 2011). In spite of the wide diversity of organisms and types of molecular and genetic systems involved, these genomic regions typically share several common features. In particular, the genes that directly determine the sexes or the mating types are often tightly linked, sometimes with a large genomic region containing many genes, in which recombination is suppressed. Such regions can include most of a chromosome (*e.g.* the male-determining region of mammalian Y chromosomes). Recombination suppression in these genomic regions is typically accompanied by a variety of degeneration signatures (BACHTROG 2005; HOOD *et al.* 2004; RICE 1996; SKALETSKY *et al.* 2005) such as low efficiency of natural selection, low gene density and accumulation of repeated DNA such as transposable elements (TEs).

At present, a comprehensive understanding of the forces driving evolution of these genomic regions is still missing (BACHTROG *et al.* 2011). In particular, two sets of issues remain unanswered. First, the process by which recombination is suppressed and the shape of the transition between recombining and non-recombining regions is not known. In sex chromosomes of mammals and those of the plant *Silene latifolia*, the level of X-Y divergence increases with increasing distance from the boundary with the recombining (pseudo-autosomal) region. Recombination suppression is therefore thought to have occurred in successive and discrete steps (BERGERO and CHARLESWORTH 2009; BERGERO *et al.* 2008; CHARLESWORTH *et al.* 2005; LAHN and PAGE 1999; LAWSON HANDLEY *et al.* 2004; ROSS *et al.* 2005; SKALETSKY *et al.* 2005), possibly involving large chromosomal inversions. Second, the factors determining the size of the non-recombining region remain poorly understood. In mammals, the size of the Y chromosome is 37% that of the X (ROSS *et al.* 2005; SKALETSKY *et al.* 2005), while in *Silene latifolia* it is 150% that of the X (MARAIS *et al.* 2008).

Homomorphic self-incompatibility (SI) is a highly relevant genetic system to address these issues. SI functions to prevent self-fertilization in hermaphroditic plants (DE NETTANCOURT 2001). While relatively widespread (being present in at least 94 flowering plant families (IGIC and LANDE 2008)), homomorphic SI has been described at the molecular level in only a handful of taxa (reviewed in (REA and NASRALLAH 2008; TAKAYAMA and ISOGAI 2005)). The genetics of SI involves a single genomic region or a small number of regions. All of the few incompatibility loci that have been characterized at the molecular level contain at least two genes, one expressed in pistils and the other in anthers for

sporophytic SI; in gametophytic SI systems, the pollen-S gene is expressed in pollen and there are sometimes multiple genes (KUBO *et al.* 2010). These genes encode proteins that physically interact in a haplotype-specific manner, ultimately allowing normal cross-pollen germination and/or growth when proteins are produced by haplotypes carrying different specificities, but preventing it when pollen and pistils express cognate specificities, in particular avoiding self-fertilization.

Evolutionary properties of the genes controlling SI have been studied in several taxa, including the Brassicaceae, Solanaceae and Papaveraceae species (FRANKLIN-TONG and FRANKLIN 2003; HISCOCK and MCINNIS 2003). In accordance with negative frequency-dependent selection theory (WRIGHT 1939), these genes show remarkable evolutionary features. First, the S-locus typically has very high haplotype diversity, with up to >100 distinct specificities in natural populations within species (see (CASTRIC and VEKEMANS 2004) for a review). Second, because they are maintained within species for extended periods of time, these haplotypes show high nucleotide divergence among specificities within species (CASTRIC *et al.* 2010) and trans-specific polymorphism between closely related species (DWYER *et al.* 1991). Third, to maintain specific recognition, the pollen and pistil genes are expected to be in strong linkage disequilibrium and hence to constitute co-adapted haplotypic combinations (SATO *et al.* 2002). Indeed, recombination between the two component genes would disrupt specific recognition, leading to self-compatible haplotypes (CASSELMAN *et al.* 2000; KAWABE *et al.* 2006). Several studies in different SI systems confirmed that recombination among haplotypes in the S-locus is highly infrequent (CASSELMAN *et al.* 2000; CASTRIC *et al.* 2010; CHARLESWORTH and AWADALLA 1998; KAWABE *et al.* 2006; VIEIRA *et al.* 2003), and consequently that pollen and pistil genes are expected to follow the same evolutionary history. Fourth, in species whose SI system is sporophytic (DE NETTANCOURT 2001), complex dominance relationships have been described among S-haplotypes controlling both pollen and pistil phenotypes (KUSABA *et al.* 2002). Sporophytic SI has been described at the molecular level in a single family, the Brassicaceae. In both *Brassica* and *Arabidopsis*, the dominance relationships among haplotypes are partly related to their phylogenetic distance, with roughly four different classes in *A. lyrata*, corresponding to four phylogenetic groups (PRIGODA *et al.* 2005) and two dominance classes in *Brassica* corresponding to two phylogenetic groups (HATAKEYAMA *et al.* 1998; UYENOYAMA 1995). In line with theoretical expectations (BILLIARD *et al.* 2007; SCHIERUP *et al.* 1997), dominant and recessive S-haplotypes appear to experience contrasted evolutionary dynamics (CASTRIC *et al.* 2010). In particular, recessive haplotypes occur at higher frequency and may form homozygote combinations between distinct gene copies of a given S-allele, which may allow for the possibility of recombination in the absence of large sequence divergence (CASTRIC *et al.* 2010).

Because of linkage to the targets of negative frequency-dependent selection, the surrounding genomic region is also expected to show deeper coalescence than the genomic background, and parting high sequence divergence among haplotypes (SCHIERUP *et al.* 2001). The physical extent of this genomic region is potentially large, in inverse proportion to the extent of local recombination restriction within

the S-locus. Analysis of the S-locus in different species belonging to different SI systems confirmed that this genomic region is indeed highly heteromorphic in terms of sequence similarity among haplotypes (ENTANI *et al.* 2003; SHIBA *et al.* 2003; TANG *et al.* 2007; TOMITA *et al.* 2004; WHEELER *et al.* 2003). However detailed analyses of the patterns of molecular evolution in the S-locus region are lacking because full sequences of the region are available in just a handful of haplotypes and in a few taxa belonging to different SI systems. In the best documented SI system, that of the Brassicaceae, twelve S-haplotypes were sequenced in the cultivated species of the *Brassica* genus (CUI *et al.* 1999; FUJIMOTO *et al.* 2006; FUKAI *et al.* 2003; KIMURA *et al.* 2002; SHIBA *et al.* 2003; SUZUKI *et al.* 1999; TAKUNO *et al.* 2007). However, many of these sequences lack the flanking regions, hence preventing comparative analysis. In addition, three haplotypes of the S-locus were sequenced in *A. thaliana*, one of which is a recombinant haplotype between two of the three main haplogroups currently segregating in the species (SHERMAN-BROYLES *et al.* 2007; TANG *et al.* 2007; THE *ARABIDOPSIS* GENOME INITIATIVE 2000). However, although the breakdown of SI is arguably recent in *A. thaliana* (BECHSGAARD *et al.* 2006), the three available sequences encode for non-functional haplotypes and may have decayed substantially, especially in light of the rapid genomic changes that occurred since the split with *A. lyrata* (HU *et al.* 2011). Only five haplotypes from natural populations have been sequenced in Brassicaceae with functional SI, all from *Arabidopsis lyrata* (BOGGS *et al.* 2009a; GUO *et al.* 2011; KUSABA *et al.* 2001). Additionally, two haplotypes carrying non-functional specificities were also reported and sequenced in this species (GUO *et al.* 2011).

Here, we obtained full sequences for a sample of 11 S-haplotypes from natural populations of *A. halleri* and *A. lyrata,* comprehensively distributed across the four phylogenetic classes described in these species. We first used these data to determine accurately the boundaries of the non-recombining S-locus region and evaluated its extent, by studying the breakdown of sequence similarity and changes in inter-haplotype phylogenetic patterns at the interface between the flanking regions and the S-locus. We then investigated patterns of variation among haplotypes in the genomic distance between *SCR* and *SRK*, in their relative orientation, and in the occurrence of additional ORFs or pseudogenes. We compared in particular the complement of transposable elements across haplotypes and asked whether the contrasted evolutionary processes acting on dominant and recessive haplotypes had left different molecular signatures. Finally, we took advantage of the complete haplotypic combinations of the two component genes *SCR* and *SRK* in *A. lyrata* and *A. halleri* to investigate their pattern of co-divergence in natural populations.

IV - RESULTS

The genomic sequences of seven *A. halleri* and four *A. lyrata* S-locus haplotypes were obtained through sequencing of 14 bacterial artificial chromosome (BAC) clones extracted from 10 individual genomic

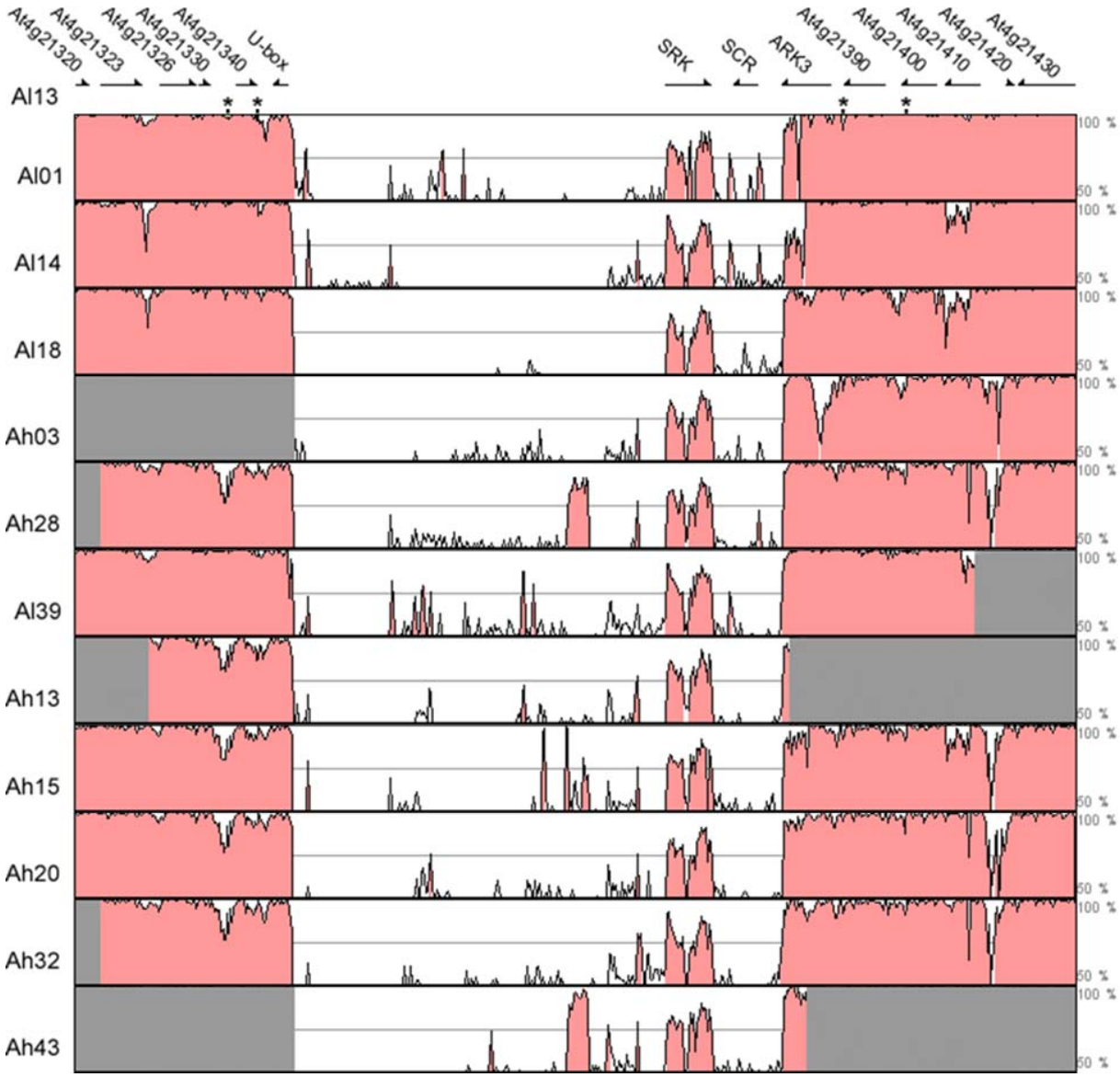


Figure 1. Sequence conservation in the S-locus region between *Al13* (the reference *A. lyrata* genome) and each of the other haplotypes. Portions of sequences not available for some haplotypes were colored in gray. For clarity, transposable elements outside of the S-locus in *Al13* were extracted from the sequence, and their locations are indicated by an asterisk.

libraries. Libraries were screened with probes from the two genes immediately flanking the S-locus region (*U-box* and *ARK3*). Positive clones were checked using BAC-end sequencing and further validated by PCR targeted on *SRK* sequences using haplotype-specific primers (LLAURENS *et al.* 2009a). Full BAC sequences were then obtained using 454 pyrosequencing technology. Because of the large sequence divergence among haplotypes, individual sequencing reads were assembled *de novo*, resulting in two to nine large contigs for each clone, with an average clone size of 98 kb and mean coverage of 57X. Attempts to increase coverage did not eliminate the gaps, suggesting that they may contain repetitive sequences. We thus used long-range PCR to validate the proposed assemblies, focusing on gaps interrupting *SCR* or *SRK* sequences (*AhSRK15, AlSRK01, AlSCR39* and *AhSCR03*). All these PCR amplifications succeeded, and contigs carrying different exons of *SCR* or *SRK* were thus confirmed to be consecutive. Detailed characteristics of the BAC clone sequences are reported in Table S1.

A - RECOMBINATION SUPPRESSION AND THE BOUNDARIES OF THE S-LOCUS

To determine the precise location of the boundaries of the S-locus region, we compared sequences from twelve S-locus haplotypes (additionally including the reference haplotype *Al13* from the *A. lyrata* full genome sequence (HU *et al.* 2011)) using the VISTA software (MAYOR *et al.* 2000), looking for a transition in the levels of sequence similarity among haplotypes. As shown in Figure 1 and Figure S1, the percentage of sequence conservation between haplotypes is fairly high in flanking regions on both sides of the S-locus, but plummets very sharply between about 300 bp upstream of the starting codon of the *U-box* on one side and near the stop codon of *ARK3* on the other side. Hence, we define the S-locus as this region of very low similarity lying between these two breakpoints. Synteny is remarkably well conserved outside the S-locus region, except for the presence or absence of some transposable elements in intergenic regions (which were removed from the reference sequence in Figure 1 for clarity). High sequence similarity among haplotypes and high collinearity of flanking genes in the region outside of the S-locus suggest that recombination among haplotypes does occur outside the region delimited by these breakpoints. Additional evidence comes from the observation that elevated diversity, as expected for neutral sites linked to sites under balancing selection (SCHIERUP *et al.* 2001), is only apparent for the two immediately flanking genes (the *U-box* and *ARK3*), while levels of synonymous nucleotide diversity are comparable with that of the genomic background (≈ 2%, (ROSS-IBARRA *et al.* 2008; RUGGIERO *et al.* 2007)) for all genes located further away on the chromosome (Figure S2). In contrast, within the S-locus, sequence similarity is almost completely lacking, the only notable exceptions being the seven exons of *SRK* and some transposable elements of the same family. Interestingly, a pseudogenized partial duplicate of the *ARK3* gene (from the end of the first exon to the end of the gene) is found within the S-locus in three different haplotypes: *Al01, Ah15* and *Ah43*. These partial duplicates of *ARK3* within the S-locus region could be responsible for the observation by Hagenblad *et al.* (2006) of the occurrence of a pseudogenized paralog of *ARK3* in some haplotypes,

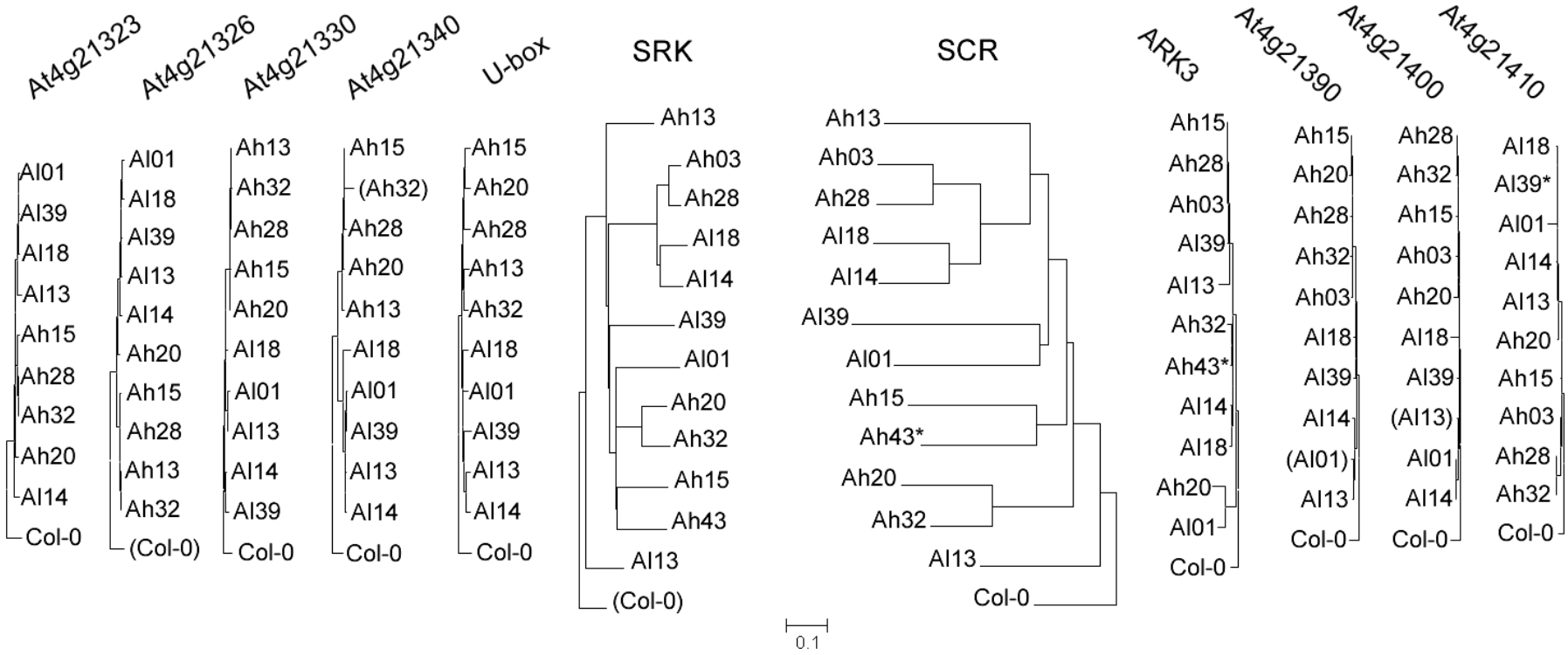


Figure 2. Gene phylogenies in and around the S-locus region. Phylogenies were obtained by the Minimum Evolution method, and are based on coding sequences, with the *A. thaliana* reference sequence (Col-0) as an outgroup. Asterisks indicate partial sequences, and brackets non functional sequences. The inversion in the *SCR* coding sequence of Col-0 was de-inverted (*i.e.* restored to its original functional configuration in *A. halleri*) according to Tsuchimatsu *et al.* (2010).

including one carrying allele *Al01* at *SRK*. A similar partial duplicate sequence of *ARK3* was found in the S-locus region of the recombinant *C24* haplotype of *A. thaliana*, and it was hypothesized that this motif acted as the recombination breakpoint between the two common haplotypes *A* and *C* (SHERMAN-BROYLES *et al.* 2007). Interestingly, the duplicated *ARK3* sequences in *Al01, Ah15* and *Ah43* are more similar to *ARK3* gene copies present in haplotypes other than their own (Figure S3). Assuming that this second copy initially originated through gene duplication from the same chromosome, this observation implies that inter-haplotype recombination does occur at the genomic position of this gene, and hence supports our conclusion that *ARK3* indeed lies outside the non-recombining region. Moreover, while the partial duplicates of *ARK3* in *Ah15* and *Ah43* are closely related, that of *Al01* is not phylogenetically close, suggesting at least two independent duplication events.

B - THE S-LOCUS HAS LOW GENE DENSITY AND SHOWS IMPORTANT STRUCTURAL REARRANGEMENTS

Annotation of the S-locus region revealed only the two incompatibility genes, *SCR* and *SRK*, plus TEs (see below). A single copy of *SCR* and of *SRK* was found in each haplotype, whereas a previous study (KUSABA *et al.* 2001) described two copies of *SCR* in one haplotype from *A. lyrata* (*Al20*). Multiple gene copies are therefore the exception rather than the rule in the S-locus of *Arabidopsis*. Sequencing of the 206.7 Mb *A. lyrata* genome predicted 32,670 genes (HU *et al.* 2011), *i.e.* approximately 0.16 genes per kb. With only two genes in about 60 kb, the S-locus appears to have very low gene density (*ca.* 4.8 times lower than the genomic background). Striking differences in the timescales of gene genealogies for the S-locus genes *SCR* and *SRK* as compared to the flanking genes were observed (Figure 2), with much deeper genealogies for *SCR* and *SRK*, as expected for genes under strong frequency-dependent selection (VEKEMANS and SLATKIN 1994). Moreover, the gene genealogies of *SCR* and *SRK* (Figure 2) were found to be more congruent than expected by chance (I_{cong} = 1.53; P-value = 0.0014 (DE VIENNE *et al.* 2007)). Specifically, the phylogenetic classes defined based on *SRK* sequences (PRIGODA *et al.* 2005) (class I: *Al01*; class II: *Ah03*, *Ah28*, *Al18* and *Al14*; class III: *Al13*; class IV: all other haplotypes), are perfectly conserved in the *SCR* tree.

In contrast, the phylogenetic relationships among haplotypes were strikingly different for the flanking genes (Figure S4), as reported for haplotypes of the *U-box* and the *ARK3* genes in *A. thaliana* (TSUCHIMATSU *et al.* 2010). Indeed, in our dataset gene genealogies of the flanking genes tend to cluster according to species overall, rather than to S-locus phylogenetic classes. This further supports the conclusion that the non-recombining region is confined to the S-locus and is determined by the two breakpoints identified based on sequence similarity.

The S-locus region is variable in size across haplotypes, spanning from 31 kb (haplotype *Al14*) to 110 kb (haplotype *Ah15*) with an average size of 62 kb. Given that sequences from the S-locus of haplotypes *Ah03, Ah13* and *Ah43* are incomplete, these estimates are lower bounds. Also, several

Table 2. Description of the different haplotypes.

Haplotype	Phylogenetic class	Size of the S-locus	*SCR* - *SRK* distance
Al01	I	42 614	2 906
Al14	II	30 909	8 671[a]
Al18	II	65 495	12 227
Ah03	II	34 512	742
Ah28	II	87 805	25 748
Al13	III	37 013	1 752
Al39	IV	55 787	6 601
Ah13	IV	73 401	17 028
Ah15	IV	109 864	618
Ah20	IV	56 764	3 636[a]
Ah32	IV	52 987	1 974
Ah43	IV	93 791	4 147

[a] Because of the uncertainty on the orientation of some contigs, the indicated distance is the minimum distance between *SCR* and *SRK*.

libraries that we constructed could not be exploited because no single clone showed both flanking genes used for screening, suggesting that the S-locus haplotypes they contain may have been larger than the average 100 kb typical of the BAC clones in our libraries. With an average size of 74 kb, haplotypes from *SRK* phylogenetic class IV are generally larger than haplotypes from classes I to III, showing an average size of 50 kb (Table 2). Figure 3 summarizes the gene organization within the S-locus and includes data from Kusaba *et al.* (2001), Boggs *et al.* (2009a) and Guo *et al.* (2011). Globally, we found that gene organization within the S-locus is highly variable with regard to both the relative orientation of *SCR* and *SRK* (tail-to-tail, head-to-head or in the same direction) and the distance separating them (from less than 1 kb to about 26 kb; Table 2). These patterns also vary among haplotypes within each of the *SRK* phylogenetic classes, with the exception of class II haplotypes showing mostly *SCR* and *SRK* oriented tail-to-tail and a location of *SRK* consistently very close to the flanking gene *ARK3* in head-to-head orientation. Strikingly, these class II haplotypes were already reported to show common features that distinguish them from other phylogenetic classes (CHARLESWORTH *et al.* 2003; PRIGODA *et al.* 2005). We found here that the strong sequence similarity previously noted in the kinase domain of these haplotypes (CHARLESWORTH *et al.* 2003) is extended to the whole intergenic region (about 900 bp in length) between *SRK* and *ARK3* (Figure S5), in contrast to comparisons with other classes of haplotypes or between classes (Figure S1). Interestingly, this same intergenic region is also conserved between class II haplotypes and haplotypes *Ah15* and *Ah43*, two of the three haplotypes carrying a pseudogeneized duplicated copy of *ARK3*. This observation strongly suggests that the duplication involved a recombination event between these haplotypes and a class II haplotype. Interestingly, while (GUO *et al.* 2011) suggested that haplotypes *Al38* and *Al50* lack the second exon of the *SCR* gene, we were able to detect the second exon upon closer examination applying the same approach than in our own data, suggesting that these haplotypes are indeed functional.

C - INVASION BY TRANSPOSABLE ELEMENTS AND THE EFFECT OF DOMINANCE

Transposable elements annotation with the CENSOR (KOHANY *et al.* 2006) and PLOTREP (TOTH *et al.* 2006) programs revealed a strong density and diversified complements of TEs in the S-locus, with a representation of most families known in the *A. thaliana* genome (detailed annotation and a complete list of TEs for each haplotype are shown in Figure S6 and Table S2). In order to determine whether these observations are uncommon in the genomic background, we also used CENSOR (KOHANY *et al.* 2006) to estimate TE density along the *A. lyrata* genome divided in non-overlapping windows of 100 kb. Variation of TE density along chromosome 7 confirmed that the TE density of the S-locus sharply departs from its chromosomal background, being matched only by the centromeric region (Figure 4, and Figure S7 for the other chromosomes). This difference is not due to an invasion by a single class of TEs, since the quantitative difference in density was observed for most TE families (Figure 5).

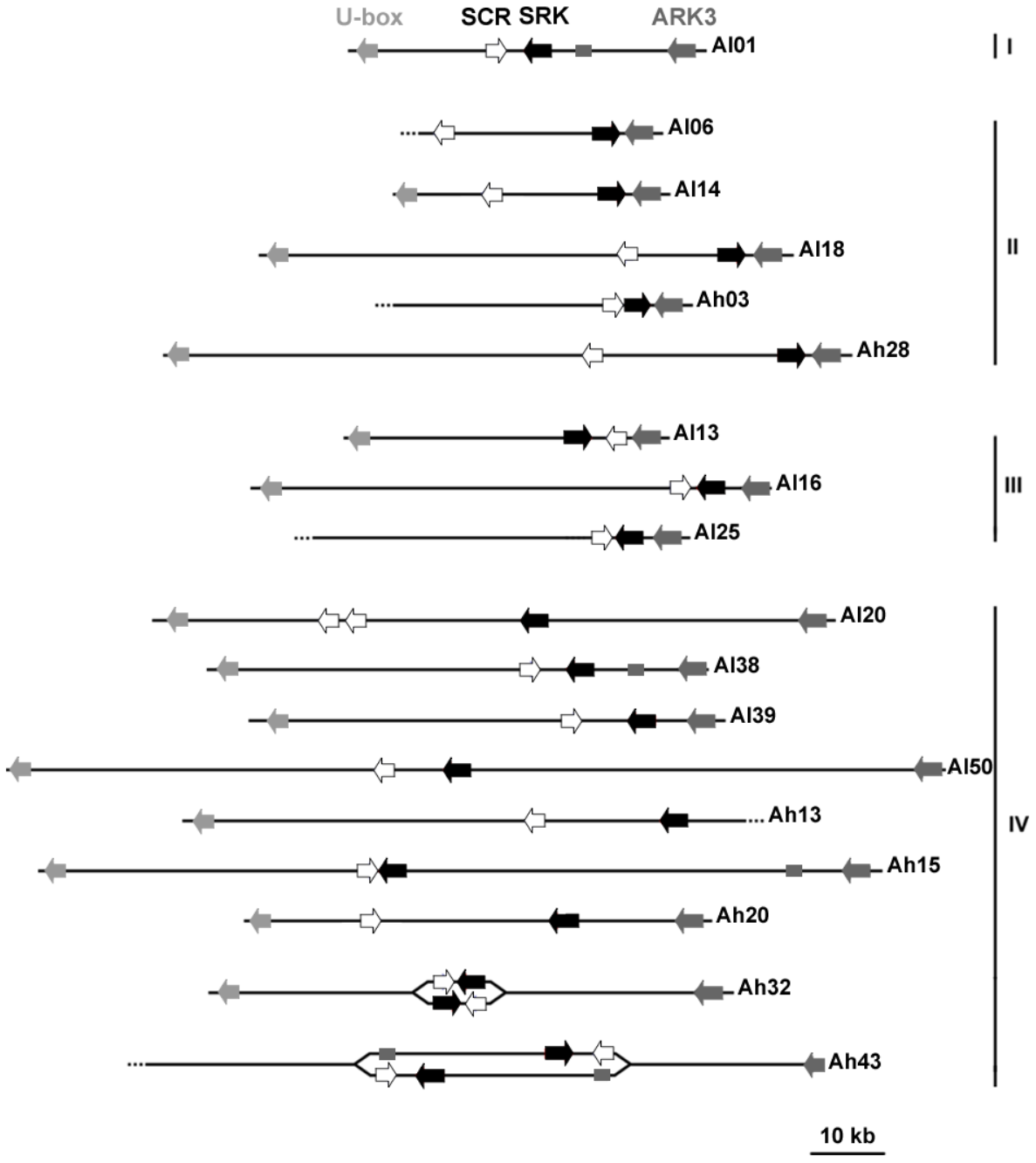


Figure 3. Structural variation within the S-locus. The direction of *SCR*, *SRK* and the two flanking genes is shown taking into account their approximate distances. Both possibilities are depicted when the orientation of genes remains unknown due to unoriented contigs. The presence of a pseudo-*ARK3* sequence is represented by a dark gray rectangle. Organization of haplotypes *Al20*, *Al06, Al25*, *Al16, Al38* and *Al50* are based on Kusaba *et al.* (2001), Boggs *et al.* (2009a) and Guo *et al.* (2011).

While most haplotypes have higher TE density than the genomic background, there is striking variability in TE density among haplotypes. Indeed, TE density depends on *SRK* phylogenetic classes, which are themselves associated with dominance with higher density in the more dominant haplotypes (Figures 6A and 6B). Since levels of dominance are in turn expected to correlate with S-haplotype frequency in natural populations (SAMPSON 1974; SCHIERUP *et al.* 1997), we plotted TE density against haplotype frequency, as estimated from S-locus genotype surveys in *A. lyrata* (SCHIERUP *et al.* 2008) and *A. halleri* (P. Goubet *et al.* unpublished data). We find that variation in TE density is even better captured by haplotype frequencies, with rare haplotypes being more enriched in TEs than more frequent haplotypes (Figure 6C).

V - DISCUSSION

Our results confirm that the S-locus in *A. halleri* and *A. lyrata* differs significantly from its genomic background in several respects: gene density is particularly low, gene genealogies are much deeper as compared to the flanking genes, gene order and orientation vary extensively, sequence similarity among haplotypes in intergenic sequences is completely lacking and the density of transposable elements is particularly elevated, being matched only by that in the centromere. Most of these properties are shared with many genetic systems controlling patterns of mating such as sex chromosomes, sex-determining loci or mating-type loci.

A - SIZE OF GENOMIC REGIONS INVOLVED IN MATING-TYPE DETERMINATION

The S-locus and its flanking regions had strikingly contrasted patterns of sequence conservation among haplotypes. This result is in line with two previous investigations comparing three and five haplotypes in *A. thaliana* (TANG *et al.* 2007) and *A. lyrata* (GUO *et al.* 2011) respectively. Based on a more extensive collection of S-haplotypes, we could precisely map the breakdown of synteny to two narrow regions very close to the 5' or 3' ends of the coding regions of the flanking genes *U-box* or *ARK3*, respectively. Using this objective criterion to define the S-locus itself, we find that the S-locus has an average size of 62 kb, ranging from 31 to 110 kb among haplotypes, much larger than the average distance of 7 kb between the two S-locus genes, *SCR* and *SRK*, ranging from 1 kb to 26 kb. An orthologous sporophytic SI system occurs in the genus *Brassica*, although the S-locus is located in a different genomic region than in *Arabidopsis*. Based on the available sequences of four *B. rapa* haplotypes (FUKAI *et al.* 2003; TAKUNO *et al.* 2007) and using a similar criterion to define the S-locus we determined that the size of the S-locus was somewhat lower than in *Arabidopsis*, ranging from 28 to more than 60 kb. In contrast, the distance separating *SCR* and *SRK* was less variable, ranging from 2 to 11 kb. In *Brassica*, however, the S-locus generally comprises a third gene, *SLG* which is a paralog of *SRK* lacking the kinase domain, and the overall region comprising these three genes ranged from 23 to 43 kb. In *Ipomoea trifida* (Convolvulaceae), which also exhibits sporophytic SI but of a different

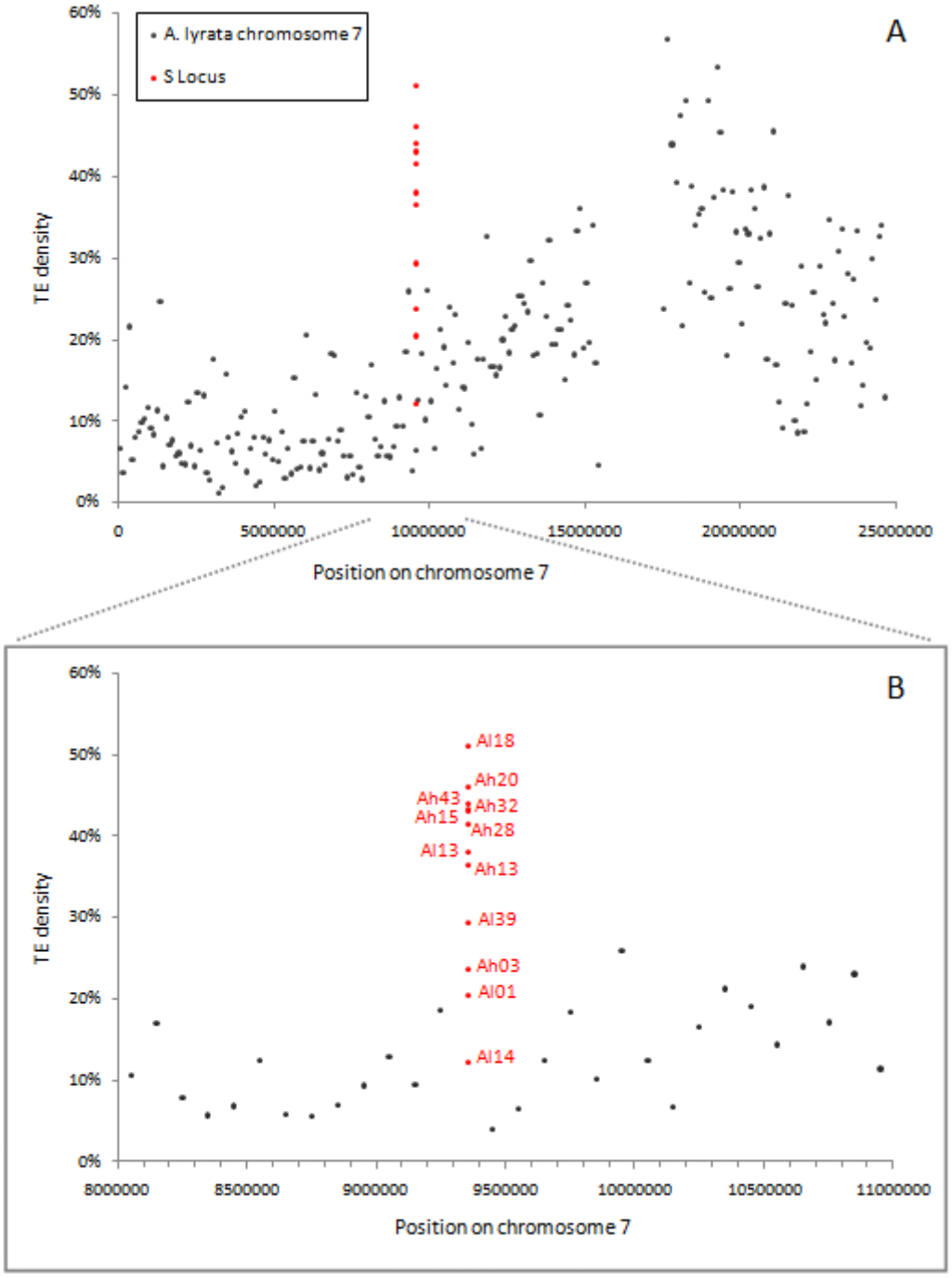


Figure 4. A. TE density along *A. lyrata* chromosome 7, and comparison with the S-locus data. Transposable elements contents were calculated using CENSOR (KOHANY *et al.* 2006) for non overlapping windows of 100 kb.

B. Zoom on the 3 Mbp region around the S-locus. The dashed line represents a 95% confidence interval on the TE densities of this 3 Mbp genomic region.

molecular nature, the S-haplotype-specific divergent regions between the only two sequenced haplotypes (*S1* and *S10*) span 50 and 34kb, respectively (RAHMAN *et al.* 2007). In the gametophytic SI system of *Prunus dulcis* and *P. mume* (Rosaceae), the S-locus was estimated as being a divergent genomic region of about 70 kb (USHIJIMA *et al.* 2001) and 15 to 27 kb (ENTANI *et al.* 2003), respectively. In *Antirrhinum hispanicum* (Plantaginaceae), the distance between the two component genes of haplotype S_2 is 9 kb (ZHOU *et al.* 2003). However, a major difference between the S-locus of the Brassicaceae and that of the Plantaginaceae and Solanaceae is that in the latter the pollen phenotype can be encoded by different members of the gene family to which the male determinant belongs, so that the S-locus comprises more than two genes (KUBO *et al.* 2010), making this comparison tricky. Overall, in spite of the large diversity of species and molecular mechanisms involved in the different SI systems, the size of S-loci seems to be fairly constant across taxa, ranging from 27 to about 110 kb, with *Arabidopsis* species apparently in the upper part of the range.

Beyond the comparison with S-loci of other plants, the size of the S-locus can also be compared with that of the mating-type loci in fungi or green algae. In the basidiomycete *Cryptococcus neoformans*, sex determination is controlled by a locus including genes encoding a pheromone and its receptor. Haplotypes of this mating-type locus, α and a, represent a genomic region of approximately 105 to 130 kb (LENGELER *et al.* 2002), hence slightly larger than the S-locus in *A. halleri* and *A. lyrata*. In another basidiomycete, *Ustilago hordei,* the mating-type locus consists of a single region comprising two complexes, a and b, between which recombination is suppressed. The distance between these two complexes was estimated to be 500 kb and 413 kb in the *MAT-1* and *MAT-2* strains, respectively (LEE *et al.* 1999). In the ascomycete *Neurospora tetrasperma*, the non-recombining region comprising the mating-type locus covers 78.4 % of the chromosome length, *i.e.* 6.9 Mbp (MENKIS *et al.* 2008). In green algae, the mating-type locus of the unicellular *Chlamydomonas reinhardtii* consists of a highly rearranged 200-kb region (FERRIS *et al.* 2002) while that of the multicellular *Volvox carteri* is about 500 % larger and contains many ORFs. Interestingly, *C. reinhardtii* is an isogamous species with two morphologically undistinguishable mating-types (GOODENOUGH *et al.* 1995) while *V. carteri* shows morphological differentiation between the mating-types, suggesting the general conclusion that genomic regions involved in mating-type systems that are not associated with morphological differences between mates may span smaller genomic regions. In other words, the accumulation of genes with a role in expression of the morphological differences between mating-types (FERRIS *et al.* 2010) may contribute to some extent to the variation in size of the mating-type locus, in addition to transposable elements and non coding DNA accumulating in these regions. Because in homomorphic SI, the mating-types are not associated with morphological differences, the S-loci may retain an apparently smaller size.

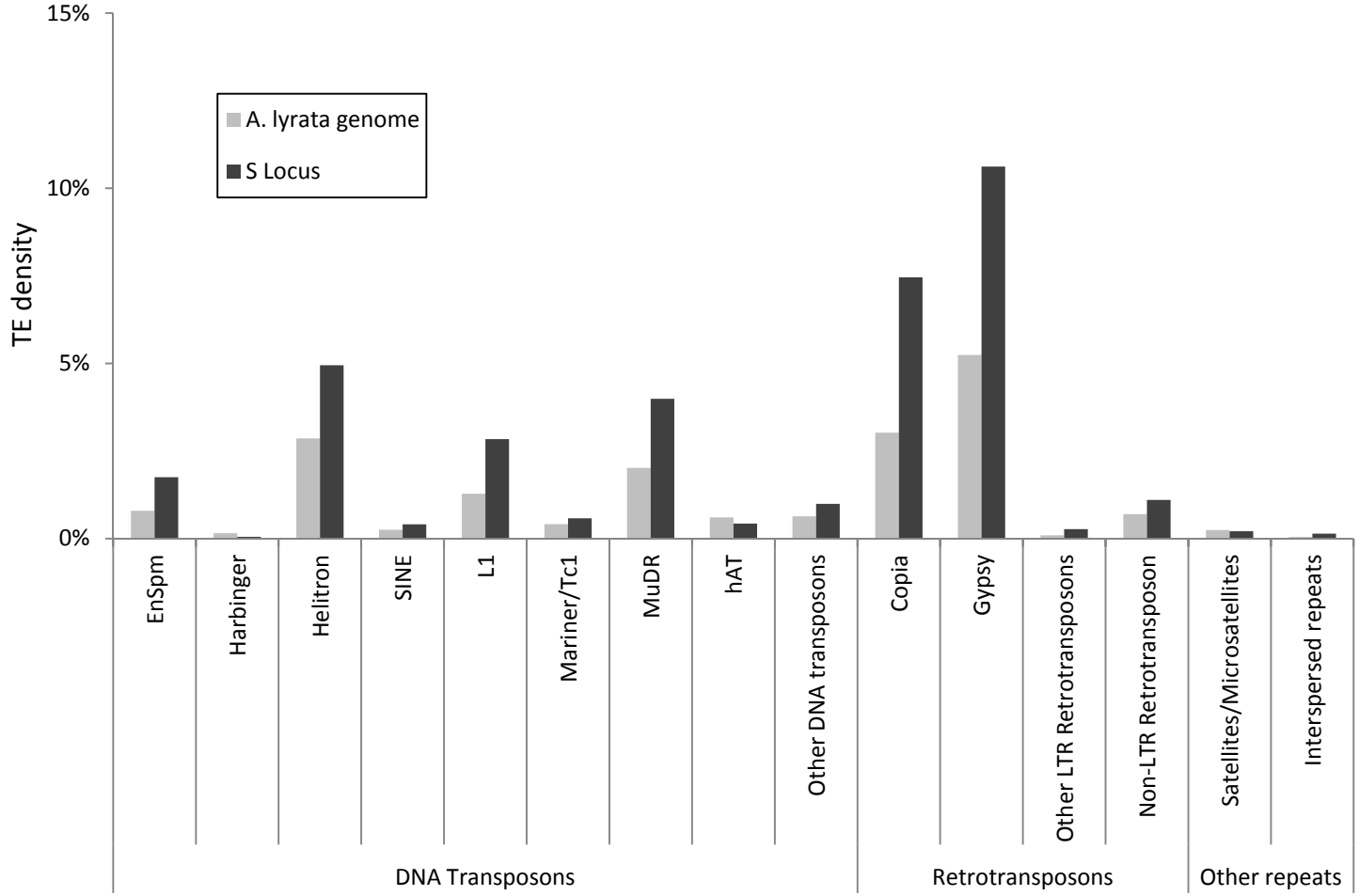


Figure 5. Comparative density in different families of transposable elements for the entire genome of *A. lyrata*, and the S-locus of *A. lyrata* and *A. halleri*. Transposable elements classification refers to Kapitonov and Jurka (2008).

B - STRUCTURAL REARRANGEMENTS, YET SHARED EVOLUTIONARY HISTORY BETWEEN *SCR* AND *SRK*

Only six sequences of *SCR* were already described in *Arabidopsis* because of the difficulty of finding conserved regions to perform PCR amplifications (BOGGS *et al.* 2009a; BOGGS *et al.* 2009b; GUO *et al.* 2011; KUSABA *et al.* 2001; TSUCHIMATSU *et al.* 2010). Our important sequencing effort of the S-locus region resulted in the successful identification of full *SCR* sequences in ten new S-haplotypes in *A. halleri* and *A. lyrata* and only the second exon of *SCR* in one haplotype (haplotype *Ah43*, for which we could not obtain the full S-locus sequence), along with their cognate *SRK* partner. These results do not support the hypothesis of existence of non-functional haplotypes carrying only partial SCR sequences, as proposed by Guo *et al.* (2011), as we were able to localize the missing coding sequence for their two putative non-functional haplotypes when applying the ALN (GOTOH 1982) software fed with all known *SCR* sequences. Congruence of *SCR* and *SRK* phylogenies reflects the coevolution necessary to maintain the specific SCR-SRK protein-protein recognition, and clearly indicates that recombination between the two SI genes has been precluded. Comparison of phylogenies between *SCR* and the S domain of *SRK* was already investigated by Sato *et al.* (2002) for twelve haplotypes in *Brassica oleraceae*. They found that the hypothesis of an identical topology for the two trees was not rejected. Edh *et al.* (2009a) also compared *SCR* and *SRK* phylogenies in *Brassica rapa, Brassica oleraceae* and *Brassica cretica* class II haplotypes, but congruence between topologies could not be clearly demonstrated, perhaps as a consequence of the concerted evolution of the *SLG* and *SRK* genes within haplotypes, or of the more recent history of diversification within the class II lineage. In contrast, in the ascomycete *Neurospora* (HALL *et al.* 2010), the non-self recognition system is controlled by two tightly linked genes, *het-c* and *pin-c*. In agreement with our results in the S-locus, congruence was found between topologies of the phylogenies of these two genes, but not with those of the flanking genes. When more *SCR/SRK* sequences become available, it will be interesting to study in more details the co-evolutionary process.

Based on the study of nine haplotypes in *A. thaliana*, *A. lyrata* and *Capsella rubella*, Guo *et al.* (2011) proposed that head-to-head orientation of *SCR* and *SRK,* with *SCR* at *U-box* side and *SRK* at *ARK3* side, was the ancestral orientation in the *Arabidopsis/Capsella* lineage. However, the lack of conserved orientation pattern in our results based on a much larger number of haplotypes suggests that, in spite of the shared evolutionary history of *SCR* and *SRK,* the S-locus has experienced a history of frequent inversions and genomic rearrangements. At this stage, we argue that ancestral orientation cannot be deduced. Strong structural variation among haplotypes seems to be a common feature of S-loci (FOBIS-LOISY *et al.* 2004) and genomic rearrangements, particularly inversions, are known to be frequent in low recombination regions such as in sex chromosomes of mammals (LEMAITRE *et al.* 2009; ROSS *et al.* 2005; SKALETSKY *et al.* 2005) and plants (BERGERO *et al.* 2008) or in the mating-type locus of green algae (FERRIS *et al.* 2002). Evidence of gradual suppression of recombination was found in sex chromosomes, with formation of evolutionary strata (BERGERO and CHARLESWORTH 2009; BERGERO *et al.* 2008; CHARLESWORTH *et al.* 2005; LAHN and PAGE 1999; LAWSON HANDLEY *et al.* 2004; ROSS *et al.* 2005;

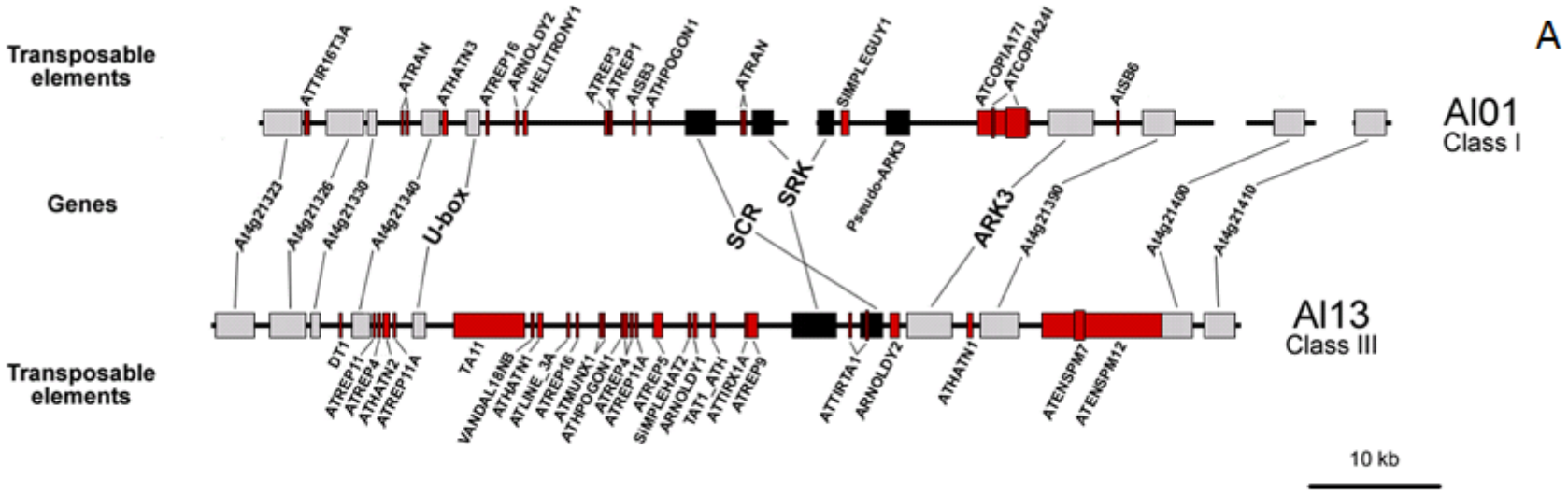


Figure 6. A. Comparative annotation of genes and transposable elements for a recessive haplotype, *Al01*, and a dominant one, *Al13*. The S-locus genes are represented by black rectangles, and other genes by light gray ones. Transposable elements, annotated using CENSOR (KOHANY *et al.* 2006) and PLOTREP (TOTH *et al.* 2006), are indicated in red.

SKALETSKY *et al.* 2005). These strata, composed of genes which stop recombining and therefore start diverging at the same time, are thought to have been caused by large inversions in the non-recombining sex chromosome (LAHN and PAGE 1999). Like in sex chromosomes, inversions in the S-locus could have contributed to the reduction in recombination among haplotypes. However, no discrete strata in levels of divergence among haplotypes could be identified, as the proportion of sequence similarity changes abruptly to zero within the S-locus region.

C - TRANSPOSABLE ELEMENTS ACCUMULATION IN SEX-DETERMINING REGIONS

Our results show that transposable elements are a major component of the S-locus region, as previously noted in other taxa (FUJIMOTO *et al.* 2006; TOMITA *et al.* 2004; WHEELER *et al.* 2003). They were firstly analyzed on a wide scale and their density was found to be higher in most haplotypes than in the genomic background. Such accumulation has already been observed in other genomic regions involved in mating-type and gender determination, and is not exclusive to the S-locus. Bachtrog (2005) investigated four regions of the neo-sex chromosomes, containing homologous gene pairs, in *Drosophila miranda*. In each case, the neo-Y showed several transposable elements insertions that were absent from the neo-X. Similarly, Marais *et al.* (2008) analyzed genetic degeneration of the Y chromosome in *Silene latifolia*, by examining seven sex-linked genes. Comparison of Y-linked genes and their X-linked homologs provided evidence that some of the Y-linked genes showed higher intron sizes, due to the accumulation of transposable elements. In the mating-type locus of the basidiomycete *Ustilago hordei*, sequencing of one of the two haplotypes, *MAT-1*, revealed that this genomic region was particularly rich in both retroelements and repetitive DNA compared to *U. maydis*, in which the a and b complexes are unlinked (BAKKEREN *et al.* 2006). Similarly, the chromosome carrying the mating-type locus in the fungus *Microbotryum violaceum* was found to be enriched in transposable elements as compared to autosomal chromosomes (HOOD *et al.* 2004). In *A. thaliana*, Wright *et al.* (2003) compared the transposable elements accumulation in chromosome arms and in low-recombining regions surrounding the centromeres, *i.e.* centromeres, pericentromeric regions and heterochromatic knobs. These regions of reduced recombination were shown to exhibit greater TE copy numbers than chromosome arms, particularly for Gypsy retrotransposons and EnSpm transposons. Interestingly, our results showed that precisely these two TE families present densities twice higher in the S-locus than in the overall genome of *A. lyrata*, suggesting that the increased TE density noticed in the S-locus is effectively linked to the restricted recombination.

D - TE ACCUMULATION: DRIVEN BY RECOMBINATION SUPPRESSION AND MUTATIONAL HAZARD?

Strikingly, we found that not all haplotypes present the same TE coverage, with dominant S-haplotypes (*SRK* phylogenetic classes III and IV) having higher TE density than those belonging to recessive classes (I and II). Two hypotheses can be proposed to explain this pattern. Firstly, this heterogeneity could be related to variation among haplotypes in the frequency of recombination in the

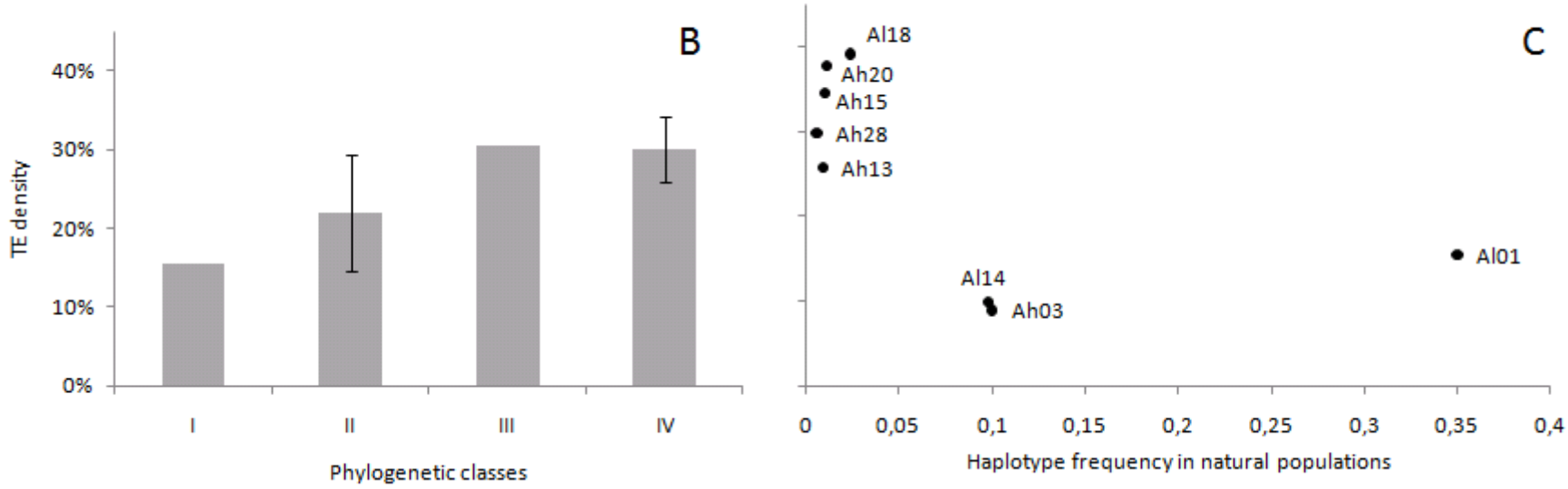


Figure 6. B. Mean TE density in the different phylogenetic classes. TE density corresponds to the percentage of a sequence which is covered by transposable elements. Standard deviation is indicated for classes II and IV, in which multiple haplotypes could be analysed.

C. TE density of the different haplotypes according to their frequency in natural populations. Haplotype frequencies are based on *SRK* fragments. *A. lyrata* data concern 12 icelandic populations (SCHIERUP *et al.* 2008) and A. *halleri* data concern 39 european populations (unpublished data).

S-locus region. Indeed, a previous work revealed signatures of intragenic recombination in *SRK* in S-haplotypes belonging to recessive classes I and II only (CASTRIC *et al.* 2010). It was suggested that such recombination events occurred in individuals carrying two copies of the same functional S-haplotype, a feature that is more probable for recessive haplotypes, which are predicted to have high frequencies in natural populations (SCHIERUP *et al.* 1997). Our observation that the density of TEs is inversely related to haplotype population frequency also suggests that recombination, occurring only for the most common recessive haplotypes, plays an active role in preventing TEs accumulation in the S-locus. However, haplotype frequency also determines the effective population size of S-locus gene copies (VEKEMANS and SLATKIN 1994), and S-haplotypes are known to reach different frequencies according to their dominance level. For example, in *A. lyrata*, the frequency of the most recessive haplotype was 12.75 times higher than that of the most dominant haplotypes in Icelandic natural populations (SCHIERUP *et al.* 2008). This suggests the non-exclusive explanation that genetic drift, stronger in low-frequency dominant haplotypes, determines the degree of TE accumulation, in agreement with the mutational-hazard model of Lynch and Conery (LYNCH and CONERY 2003). We argue that such differences in recombination and/or in level of genetic drift, and hence differences in accumulation of TEs, may be an important source of variation of the size of the S-locus among haplotypes. Sex-chromosomes in mammals also feature such differences in opportunities for recombination and in effective population size (HELLBORG and ELLEGREN 2004). Our study shows that recessive S-haplotypes tend to behave as the X chromosome and dominant haplotypes as the Y chromosome.

VI - METHODS

A - CONSTRUCTION OF BAC LIBRARIES

High Molecular Weight (HMW) DNA was prepared from young leaves of six *A. halleri* and four *A. lyrata* haplotypes. For each extraction, approximately 20 grams of frozen leaf tissue was ground to powder in liquid nitrogen with a mortar and pestle used to prepare megabase-size DNA embedded in agarose plugs. HMW DNA of the various genotypes was prepared as described by Peterson *et al.* (2000) and modified as described in (GONTHIER *et al.* 2010). Embedded HMW DNA was partially digested with *Hind*III (New England Biolabs, Ipswich, Massachusetts), subjected to two size selection steps by pulsed- field electrophoresis, using a BioRad CHEF Mapper system (Bio-Rad Laboratories, Hercules, California), and ligated to pIndigoBAC-5 *Hind*III-Cloning Ready vector (Epicentre Biotecnologies, Madison, Wisconsin). Pulsed-field migration programs, electrophoresis buffer, and ligation desalting conditions were performed according to (CHALHOUB *et al.* 2004).

To evaluate the average insert size of each library, BAC DNA was isolated from about 384 randomly selected clones in each library, restriction enzyme digested with the rare cutter NotI, and analyzed by

Pulsed-Field Gel Electrophoresis (PFGE). All fragments generated by NotI digestion contained the 7.5 kb vector band and various insert fragments. Analysis of the insert sizes from the various BAC libraries showed a mean insert size comprised between 80 kb and 175 kb. Since the haploid genome of *A. lyrata* and *A. halleri* is estimated around 230 Mb and 250 Mb respectively, we picked the number of BAC clones required to obtain a library coverage of 5 genome equivalents.

B - BAC LIBRARIES SCREENING

High-density colony filters were prepared from all the nine BAC libraries constructed using a robotic workstation QPix2 XT (Genetix). BAC clones were spotted in double using a 5x5 or 6x6 pattern onto 22x22 cm Immobilon-Ny+ filters (Millipore Corporate, Billerica, Massachusetts). On each filter, 27 648 to 41 472 unique clones were spotted in duplicate, and clones were grown at 37°C for 17 h. Filters were then processed as follows: (1) denaturation on Whatman paper soaked with a solution of 0.5M NaOH and 1.5M NaCl for 4 min at room temperature, and for 10 min at 100°C, (2) neutralization on Whatman paper soaked with 1M TrisHCl pH 7.4, and 1.5M NaCl for 10 min, incubation in a solution of 0.25 mg/mL proteinase K (Sigma Aldrich, St. Louis, Missouri) for 45 min at 37°C, baking for 45 min at 80°C, and (3) fixation by UV on a Biolink 254 nm crosslinker (Thermo Fischer Scientific, Waltham, Massachusetts) with an energy of 120,000 µJoules. Radiolabelling of probes and hybridization of the filters were performed as described in (GONTHIER *et al.* 2010). Hybridized filters were imaged with a Storm 860 PhosphorImager (GE Healthcare, Little Chalfont, UK), and analyses were performed using the HDFR software (Incogen, Williamsburg, Virginia). Positive BAC clones detected by hybridization were validated individually by PCR amplification using the primer pairs used for probes synthesis (Table S3), and visualisation of PCR products after agarose gel electrophoresis.

C - SEQUENCING

A total of thirteen BACs were sequenced at Genoscope. Two other clones (containing haplotypes *Al39* and *Ah43*) were sequenced at CNRGV and a last one (containing the haplotype *Al14*) was sequenced by the society Beckman Coulter Genomics SA. All clones were sequenced using a 454 multiplexing technology on Titanium sequencer version (www.roche.com). Assemblies were performed by Newbler (www.roche.com) and only contigs representing the extremities of the BACs were organized at this step.

D - SEQUENCE FINISHING

BAC sequences were obtained in two to nine contigs. Suggestion of orientation was provided with assembly for some sequences, but in most cases, only the first and last contigs were oriented. The relative order and orientation of other contigs were therefore unknown. When exons of *SCR* or *SRK* were in two different contigs (*i.e.* haplotypes *Al01* and *Ah15* for *SRK*, *Ah03* and *Al39* for *SCR*), primers

were defined with Primer3 (ROZEN and SKALETSKY 2000) on both contigs and PCR were performed in order to confirm the contiguity of the two contigs.

E - SEQUENCE ANNOTATION

Annotation of BAC sequences was performed using two programs of gene structure prediction with *Arabidopsis* parameters, FGENESH (SALAMOV and SOLOVYEV 2000) and GENSCAN (BURGE and KARLIN 1997). FGENESH has the advantage of being more accurate in detecting *Arabidopsis* genes but GENSCAN is more sensitive. Detected ORFs were blasted using BLASTX (GISH and STATES 1993) and the obtained proteins were then aligned on BAC sequences with SPALN (GOTOH 2008) and FGENESH+ (SALAMOV and SOLOVYEV 2000) softwares. Because of its high nucleotide diversity, *SCR* was rarely detected by these two programs. Known *SCR* proteins were thus aligned on BAC sequences using the semiglobal alignment procedure implemented on ALN (GOTOH 1982), which is more sensitive than SPALN and FGENESH+. Results of this analysis were then examined by eye in order to find the *SCR* gene and the typical cystein residues of its protein. Transposable elements were annotated with CENSOR (KOHANY *et al.* 2006) using the *A. thaliana* repetitive elements [v16.02] database of Repbase Update (JURKA *et al.* 2005). Results were then filtered and defragmented with PLOTREP (TOTH *et al.* 2006), using a minimum coverage of merged fragments of 10 %.

F - COMPARISON OF SEQUENCES AND PHYLOGENETIC ANALYSIS

The full BAC sequences were aligned and compared using the "glocal" alignment procedure (BRUDNO *et al.* 2003) implemented in VISTA (MAYOR *et al.* 2000). This kind of alignment is able to detect rearrangements and inversions in sequences, and is particularly appropriate for divergent regions like the S-locus. Protein sequences of genes were aligned with CLUSTALW (THOMPSON *et al.* 1994). Alignments were then manually adjusted and phylogenetic trees were constructed with a Minimum Evolution (ME) analysis using MEGA version 5 (TAMURA *et al.* 2011).

G - ANALYSIS OF THE TRANSPOSABLE ELEMENTS CONTENT

A PERL script was developed to compare TE density between the twelve S haplotypes and the *A. lyrata* genome. CENSOR (KOHANY *et al.* 2006) was used in local on BAC sequences, excluding the S-locus flanking regions, and on non-overlapping windows of 100 kb along the eight chromosomal sequences of the *A. lyrata* genome version Araly1 (http://genome.jgi-psf.org/Araly1/Araly1.download.html (HU *et al.* 2011)). Densities were thus calculated for each transposable elements family in the *A. lyrata* genome and in the S-locus, according to the classification in Kapitonov and Jurka (2008).

VII - ACKNOWLEDGEMENTS

We thank Deborah Charlesworth, Gabriel Marais and Tatiana Giraud for discussion and helpful comments on the manuscript. We are also grateful to Eric Schmitt and Angélique Bourceaux for taking excellent care of plants in the greenhouse. Yalong Guo kindly shared sequences from haplotypes *Al16*, *Al38* and *Al50*. Jesper Bechsgaard and Mikkel H. Schierup provided the *A. lyrata* plant material used in this study. Maude Pupin, Hélène Touzet, Camille Roux and Clémentine Vitte provided computational advice on the use of software for data analysis. This project was funded by Genoscope project AP2006/07-projet#13. P. G. was supported by a CNRS doctoral grant.

CHAPITRE II - RUPTURE DE L'AUTO-INCOMPATIBILITE CHEZ *ARABIDOPSIS THALIANA*

Contrairement à ses deux espèces sœurs *A. halleri* et *A. lyrata*, *A. thaliana* est une espèce auto-compatible, ce qui implique qu'elle ait perdu son système d'auto-incompatibilité au cours de son histoire. Cette différence de système de reproduction se retrouve notamment au niveau du nombre d'haplotypes au locus S. En effet, alors que trente à quarante haplogroupes sont connus chez *A. lyrata* et *A. halleri* (CASTRIC *et al.* 2008), seules trois anciennes spécificités ont été identifiées chez *A. thaliana* (SHIMIZU *et al.* 2004) : l'haplogroupe *A*, l'haplogroupe *B* et l'haplogroupe *C*.

Trois séquences génomiques comprenant la région du locus S chez *A. thaliana* ont à ce jour été publiées dans la littérature :

- Col-0, un haplotype *A* provenant d'une accession allemande dont le génome a été le premier génome de plante entièrement séquencé (THE *ARABIDOPSIS* GENOME INITIATIVE 2000).

- Cvi-0, un haplotype *B* provenant d'une accession du Cap Vert (TANG *et al.* 2007).

- C24, provenant d'une accession portugaise et proposé comme étant un recombinant entre deux haplotypes appartenant aux haplogroupes *A* et *C* (SHERMAN-BROYLES *et al.* 2007).

L'accession marocaine Ita-0, dont la région génomique comprenant le locus S fait partie de nos données séquencées, possède un haplotype *C*. Ainsi, l'obtention d'un haplotype du seul haplogroupe principal à n'avoir pas encore été séquencé constitue un enjeu majeur dans la comparaison des différentes spécificités d'*A. thaliana*, ainsi que dans la confirmation de l'évènement de recombinaison ayant abouti à l'haplotype C24. De plus, les trois haplogroupes principaux d'*A. thaliana* étant maintenant disponibles, une analyse de leur distribution en population naturelle a également pu être intégrée à cette partie, permettant d'étudier les patrons de dégénérescence d'un locus d'auto-incompatibilité après la rupture du système.

Ce chapitre fait l'objet d'un article en préparation : « Analysis of a reference sequence of haplogroup C of the Arabidopsis thaliana self-incompatibility locus and implications for the evolution of self-compatibility ».

TITLE: ANALYSIS OF A REFERENCE SEQUENCE OF HAPLOGROUP *C* OF THE *ARABIDOPSIS THALIANA* SELF-INCOMPATIBILITY LOCUS AND IMPLICATIONS FOR THE EVOLUTION OF SELF-COMPATIBILITY

Contributions:

Acquisition and assembly of reads from 107 accessions of the 1001 genomes project: Magnus Nordborg[1] and Quan Long[1]

Perl script: Sophie Gallina[2]

Data analysis, interpretation and manuscript preparation: Pauline Goubet[2], Xavier Vekemans[2] and Vincent Castric[2]

Affiliations:

[1] Gregor Mendel Institute, 1030 Vienna, Austria

[2] Laboratoire de Génétique et Evolution des Populations Végétales, CNRS FRE 3268, Bâtiment SN2, Université Lille Nord de France, Cité scientifique, F-59655 Villeneuve d'Ascq Cedex, France

I - ABSTRACT

Self-incompatibility (SI) is a common genetic system preventing selfing in hermaphroditic plants through recognition and rejection of self pollen. This genetic system is controlled by a genomic region, called the S-locus, at which a large number of haplotypes are typically maintained by negative frequency-dependent selection. Paradoxically, while selfing is commonly considered as an evolutionary dead-end, the evolution from outcrossing to selfing is one of the most prevalent transitions in the flowering plants. *Arabidopsis thaliana* is an autogamous species, which has lost SI as a result of several independent disrupting mutations. In contrast with its close relatives with functional SI *A. halleri* and *A. lyrata*, only three divergent S-haplogroups were identified in *A. thaliana*. Full sequences from only two of these haplogroups and from one recombinant haplotype were obtained previously and analyzed. Here, we obtained a sequence from the third haplogroup through full sequencing of a BAC clone from an individual library. Examination of this sequence revealed apparently intact coding sequences of the two interacting self-incompatibility genes, *SCR* and *SRK*, suggesting that the breakdown of SI in this haplogroup was due to mutation in non-coding regions of these genes or at another locus. We then took advantage of this complete set of S-haplogroup sequences to assemble the whole S-locus genomic region of a set of 107 north-European accessions by mapping on these templates the raw reads from the 1001 *A. thaliana* genomes project. We compared the structure of the S-haplotypes across accessions, in particular focusing on the identification of recombination events. We found that recombinants were strikingly frequent in natural populations (40 %) and proposed a scenario for the degeneration of the S-locus in *A. thaliana*.

II - INTRODUCTION

Outcrossing is a common mating system consisting in the breeding of unrelated individuals, and promoted by various mechanisms such as dichogamy or heteromorphic and homomorphic self-incompatibility. The evolution and maintenance of outcrossing has prompted a number of studies aimed at characterizing its genetic benefits, including increased genetic diversity and limitation of inbreeding depression (CHARLESWORTH 2006). Yet, transitions from outcrossing to selfing are among the most prevalent in flowering plants (BARRETT 2002), having occurred independently in hundreds of lineages (GOODWILLIE 1999; SCHOEN *et al.* 1997). Two sets of factors can cause the benefits of selfing to outweight those of outcrossing. First, selfing can be favored when pollinators are scarce, in small populations (GOODWILLIE *et al.* 2005) or during the initial steps of colonization, providing reproductive assurance (BAKER 1967) under these conditions. Second, selfing also provides an "automatic" transmission advantage, since autogamous individuals transmit gametes through both ovule and pollen when selfing, but through either pollen or ovules only when outcrossing (FISHER 1941).

Self-incompatibility (SI) is a genetic system thought to occur in about 40% of Angiosperm species (IGIC and LANDE 2008), which functions to limit inbreeding depression by preventing self-fertilization in hermaphrodites (DE NETTANCOURT 2001). In most documented cases, SI is controlled by a single genomic region (the S-locus) at which a large number of divergent haplotypes are maintained by negative frequency-dependent selection (WRIGHT 1939) and comprising two tightly linked genes, one expressed in pollen and one in pistils (REA and NASRALLAH 2008; TAKAYAMA and ISOGAI 2005). In the Brassicaceae, the pollen gene *SCR* produces a ligand protein and the pistil gene *SRK* produces its receptor (STEIN *et al.* 1991). Physical interaction between these two proteins at the stigmatic surface is responsible for the pollen-pistil recognition step necessary to trigger the incompatibility reaction in the presence of self-pollen.

Beside theoretical approaches, the number of empirical systems allowing to investigate the proximal causes and precise scenario of the transition from outcrossing to selfing has remained very limited. *A. thaliana* is a highly selfing species in a family (Brassicaceae) where SI is thought to be ancestral. It has diverged from the self-incompatible *A. lyrata* and *A. halleri* about 3-5.8 million years ago (CLAUSS and KOCH 2006). Multiple lines of evidence suggests that the breakdown of SI in *A. thaliana* was recent (BECHSGAARD *et al.* 2006; CHARLESWORTH and VEKEMANS 2005; TANG *et al.* 2007), making this biological system especially relevant to study the molecular nature of the mutations having caused the transition to selfing. *A. thaliana* seems to have lost SI as a result of several independent disrupting mutations (BOGGS *et al.* 2009b; SHIMIZU *et al.* 2008). Strikingly, while *A. lyrata* and *A. halleri* present at least 38 and 30 divergent S-haplogroupes, respectively (CASTRIC *et al.* 2008), only three S-haplogroups (*A*, *B* and *C*) have been identified in *A. thaliana* (SHIMIZU *et al.* 2004), all of which have closely related orthologs in either *A. halleri* or *A. lyrata* (BECHSGAARD *et al.* 2006). Whereas haplogroup *B* is only found in offshore African islands, haplogroups *A* and *C* are largely distributed throughout Eurasia, North Africa and North America (SHERMAN-BROYLES *et al.* 2007). Three full sequences of the genomic region comprising the inactivated S-locus in *A. thaliana* have previously been obtained: one from the reference *A. thaliana* Col-0 genome belonging to haplogroup *A* (THE *ARABIDOPSIS* GENOME INITIATIVE 2000), one from the Cvi-0 accession belonging to haplogroup *B* (TANG *et al.* 2007), and the last one from the Portuguese C24 accession thought to belong to haplogroup C. Careful examination of this sequence suggested that it actually results from recombination between haplotypes belonging to haplogroups *A* and *C* (SHERMAN-BROYLES *et al.* 2007), although a reference sequence for haplogroup *C* has not yet been obtained. All three sequences showed obvious signs of functional inactivation of the component genes of SI. In Col-0, ΨSRK_A presents a premature stop codon and ΨSCR_1 seems to be truncated as a result of an inversion comprising a part of its second exon (2009b). Tsuchimatsu *et al.* (2010) further showed that restoring the inverted fragment of ΨSCR_1 to its original position (as inferred from its functional ortholog in *A. halleri*) was sufficient to rescue SI in several accessions with haplotype *A*, hence demonstrating that those accessions carried no other inactivating mutation apart from this causal mutation. In Cvi-0, while the SCR_B coding sequence seems to be functional, ΨSRK_B also presents a

premature stop codon (SHIMIZU *et al.* 2008). C24 completely lacks any *SCR* sequence and was proposed to be an *A*-*C* recombinant because of the presence of rearranged *SRK* sequences from both haplogroups *A* and *C* and the presence of a chimeric version of an S-locus flanking gene, *ARK3*, which would have been the recombination breakpoint between the two parental haplotypes (SHERMAN-BROYLES *et al.* 2007). Moreover, functional studies showed that complementation of several accessions (including C24 and Cvi-0) with functional *SCR* and *SRK* sequences from *A. lyrata* and *Capsella grandiflora* restored SI (BOGGS *et al.* 2009a; BOGGS *et al.* 2009b; NASRALLAH *et al.* 2004), demonstrating that the breakdown of SI in those accessions was only due to the inactivation of the S-locus. In other accessions (including Col-0), however, the same experiment did not restore SI, suggesting that these accessions also carry mutations in other genes necessary for SI (LIU *et al.* 2007).

Sherman-Broyles *et al.* (2007) and Shimizu *et al.* (2008) used the available genomic sequences of the S-locus region in Col-0, Cvi-0 and C24 accessions to design probes targeted to specific components of each haplogroup in order to describe the molecular diversity of this genomic region in a set of *A. thaliana* accessions. Based on the distribution of haplotypes in natural populations, Sherman-Broyles *et al.* (2007) proposed a scenario for how and when the loss of SI occurred in *A. thaliana*. According to this scenario, transition to selfing would have occurred while the species recolonized Europe from Asian and Mediterranean glacial refuges during the Pleistocene, *i.e.* about 30 000 years ago. Haplogroups *A* and *C* would have originated in Asian and Mediterranean glacial refuges respectively, while haplogroup *B* would have been maintained in the Cap Verde Islands, *i.e.* in its current distribution area. Independent mutations causing the breakdown of SI would have occurred for each haploptype, including the inactivating inversion of haplogroup *A* (BOGGS *et al.* 2009b; TSUCHIMATSU *et al.* 2010), either in the refuges themselves or during postglacial recolonization. Recombination events between haplogroups from the *A* and *C* groups would be associated with the suture zones that were created when lineages from both refugia came into contact. Nevertheless, this scenario is based on probes, targeted on small genomic fragments, and not on the entire S-locus. Moreover, no sequence of haplogroup *C* is yet available and while SRK_C seems to be intact (SHIMIZU *et al.* 2008), no PCR amplification succeeded for SCR_C so that its occurrence is still unknown (SHIMIZU *et al.* 2008).

Here, we obtained the genomic sequence of a haplotype belonging to haplogroup *C*, the third and last major S-locus haplogroup in *A. thaliana* which had not been sequenced yet. We analyzed its sequence and searched for mutations possibly damaging the function of SI genes in order to understand the breakdown of self-incompatibility in this haplogroup. This sequence then allowed us to take advantage of a first set of available genomic data from 107 north-European accessions from the 1001 genomes project to assemble the S-locus region in each of these accessions and evaluate the pattern of degeneration of the S-locus in *A. thaliana*. In particular, we studied the distribution of the identified haplotypes in natural populations of Northern Europe, focusing more specifically on the detection of different recombination events between haplogroups *A* and *C*.

III - METHODS

A - CONSTRUCTION OF THE BAC LIBRARY

High Molecular Weight (HMW) DNA was prepared from young leaves of the Ita-0 accession. Approximately 20 grams of frozen leaf tissue was ground to powder in liquid nitrogen with a mortar and pestle used to prepare megabase-size DNA embedded in agarose plugs. HMW DNA was prepared as described by Peterson *et al.* (2000) and modified as described in (GONTHIER *et al.* 2010). Embedded HMW DNA was partially digested with *HindIII* (New England Biolabs, Ipswich, Massachusetts), subjected to two size selection steps by pulsed-field electrophoresis, using a BioRad CHEF Mapper system (Bio-Rad Laboratories, Hercules, California), and ligated to pIndigoBAC-5 *HindIII*-Cloning Ready vector (Epicentre Biotechnologies, Madison, Wisconsin). Pulsed-field migration programs, electrophoresis buffer, and ligation desalting conditions were performed according to (CHALHOUB *et al.* 2004).

B - SCREENING OF THE BAC LIBRARY

High-density colony filters were prepared from the BAC library constructed using a robotic workstation QPix2 XT (Genetix). BAC clones corresponding to 36 microplates of 384 wells were spotted in duplicate onto 22x22 cm Immobilon-Ny+ filters (Millipore Corporate, Billerica, Massachusetts) and grown at 37°C for 17 h. Filters were then processed as follows: (1) denaturation on Whatman paper soaked with a solution of 0.5M NaOH and 1.5M NaCl for 4 min at room temperature, and for 10 min at 100°C, (2) neutralization on Whatman paper soaked with 1M TrisHCl pH 7.4, and 1.5M NaCl for 10 min, incubation in a solution of 0.25 mg/mL proteinase K (Sigma Aldrich, St. Louis, Missouri) for 45 min at 37°C, baking for 45 min at 80°C, and (3) fixation by UV on a Biolink 254 nm crosslinker (Thermo Fischer Scientific, Waltham, Massachusetts) with an energy of 120,000 µJoules. Radiolabelling of probes and hybridization of the filters were performed as described in (GONTHIER *et al.* 2010). Hybridized filters were imaged with a Storm 860 PhosphorImager (GE Healthcare, Little Chalfont, UK), and analyses were performed using the HDFR software (Incogen, Williamsburg, Virginia). Positive BAC clones detected by hybridization were validated by PCR amplification using the primer pairs used for probes synthesis (see the Table S3 of Chapter I), and visualisation of PCR products after agarose gel electrophoresis.

C - SEQUENCING AND FINISHING

The BAC clone containing the haplotype C was sequenced at Genoscope using a 454 multiplexing technology on Titanium sequencer version (www.roche.com). Because the high sequence divergence at the S-locus prevented simple alignment on the genomic reference, a *de novo* assembly was performed by Newbler (www.roche.com) and the sequence was obtained in several contigs. Only contigs representing the extremities of the BAC were initially oriented. To order and orientate the rest

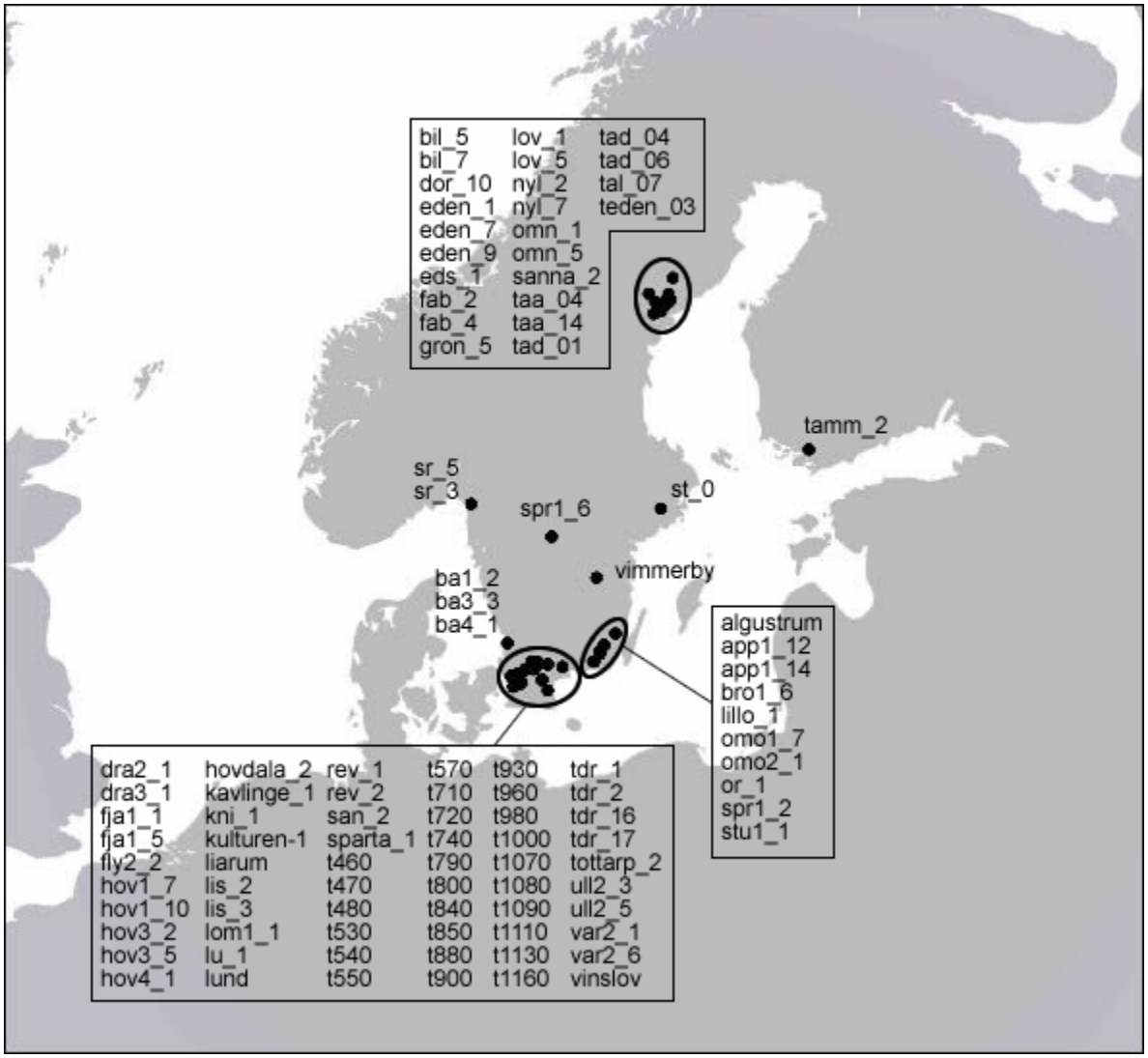


Figure 7. Origin of analyzed accessions. 26 accessions come from northeast Sweden, 78 from southern Sweden and one from southern Finland. The bur_0 accession is not indicated but comes from Ireland.

of the contigs, PCR were performed. Long-range rather than classical PCR had to be employed because the presence of repeated sequences (including transposable elements) near the extremities of most contigs imposed constraints on the design of the primers. Primers were therefore designed with Primer3 (ROZEN and SKALETSKY 2000) on each contig in regions without transposable elements or repeated sequences.

D - SEQUENCE ANNOTATION

Annotation of BAC sequences was performed using FGENESH (SALAMOV and SOLOVYEV 2000) and GENSCAN (BURGE and KARLIN 1997) with *Arabidopsis* parameters. Detected ORFs were blasted using BLASTX (GISH and STATES 1993) and the obtained proteins were then aligned on BAC sequences with SPALN (GOTOH 2008) and FGENESH+ (SALAMOV and SOLOVYEV 2000) softwares. The S-locus genes being extremely divergent, known sequences of *SRK* and *SCR* were aligned on the BAC sequence with SPALN (GOTOH 2008) and ALN (GOTOH 1982). The amino-acid sequence of *SCR* and *SRK* was then examined for obvious inactivating mutations, such as the presence of either large deletions or inversions, stop codons or the absence of a typical set of eight cysteine residues in SCR. Transposable elements were annotated with CENSOR (KOHANY *et al.* 2006) using the *A. thaliana* repetitive elements [v16.02] database of Repbase Update (JURKA *et al.* 2005). Results were then filtered and defragmented with PLOTREP (TOTH *et al.* 2006), using a minimum coverage of merged fragments of 10 %. Alignment of SCR protein sequences was performed using the Muscle software (EDGAR 2004).

E - CONFIRMATION OF THE ORIGIN OF C24 BY RECOMBINATION

Using the "glocal" alignment procedure (BRUDNO *et al.* 2003) implemented in VISTA (MAYOR *et al.* 2000), the full BAC sequence of Ita-0 was compared with Col-0 and C24 sequences, in order to check whether the C24 S-locus haplotype originated through recombination between haplogroups *A* and *C*, as previously suggested (SHERMAN-BROYLES *et al.* 2007). This "glocal" alignment procedure is indeed appropriate for studying a highly divergent region like the S-locus, because it can detect inversions and other rearrangements. On a smaller scale, the two parts of the chimeric *ΔARK3* sequence of C24 were aligned using CLUSTALW (THOMPSON *et al.* 1994) with corresponding parts of *ARK3* sequences from Col-0, Cvi-0, C24 and Ita-0. Phylogenetic trees were constructed with a Minimum Evolution (ME) analysis using MEGA version 5 (TAMURA *et al.* 2011).

F - ANALYSIS OF THE S-LOCUS REGION IN ACCESSIONS FROM THE 1001 GENOMES PROJECT

The 1001 genomes project (SCHNEEBERGER *et al.* 2011; WEIGEL and MOTT 2009) aims to sequence the whole genome of 1001 accessions of *A. thaliana*, and produces important data for the analysis of polymorphism in this reference plant. Here, we used the paired end reads from whole genome sequencing of 107 accessions from Northern Europe (Figure 7) as part of the 1001 genomes project. In this project (1001genomes.org), each accession was sequenced using Illumina sequencing with paired-

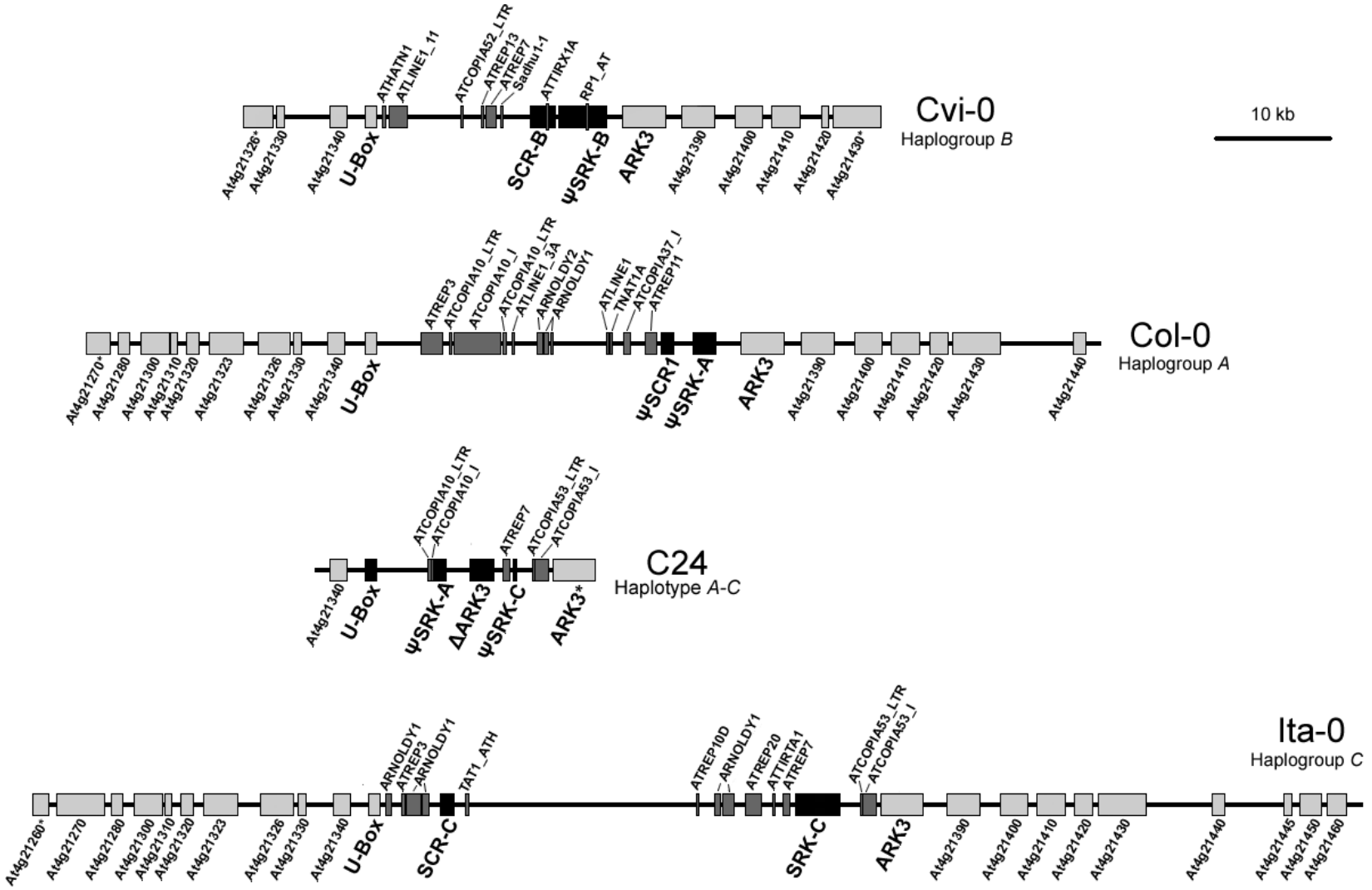


Figure 8. Annotation of the S-locus genes and transposable elements for accessions Cvi-0 (haplogroup *B*), Col-0 (haplogroup *A*), C24 (haplotype *A*-*C*) and Ita-0 (haplogroup *C*) of *A. thaliana*. The S-locus genes are represented by black rectangles, with delimitation of their exons. Genes of the flanking regions are depicted in light gray. Transposable elements are shown in dark gray. Asterisks indicate sequences of genes which are not complete because the BAC clone ends in them. Data for haplogroups *A, B* and haplotype *A*-*C* are from (THE *ARABIDOPSIS* GENOME INITIATIVE 2000; SHERMAN-BROYLES *et al.* 2007; TANG *et al.* 2007), respectively.

end reads of 76 bp. According to a survey of genomic DNA sequence polymorphism (NORDBORG *et al.* 2005), the accessions we analyzed corresponded to two different clusters (one for accessions from Northern Sweden and Finland, comprising 26 accessions, the second for accessions from Southern Sweden and from Ireland, comprising 81 accessions) which could be related to different post-glaciation colonization routes from Russia and from Europe. Because of the large sequence divergence across haplotypes, the S-locus can typically not be assembled using standard procedures (*i.e.* mapping on a single genomic reference, Col-0). Hence, for each of these accessions, four new alignments of the S-locus were performed using the Burrows-Wheeler Aligner (LI and DURBIN 2009), taking in turn each known sequence (Col-0 for haplogroup *A*, Cvi-0 for haplogroup *B*, Ita-0 for haplogroup *C* and the recombinant C24) as a reference.

We firstly analyzed 40 of the 107 accessions and searched manually for fragments of sequences matching the references. Based on this initial set of accessions, we defined several groups of haplotypes, according to which fragments of the reference sequences they contained. Note that this method can detect regions that are either conserved or deleted as compared to the reference, but not insertion events. A *perl* script was then developed to automatically detect the S-haplotype groups. Based on conserved fragments detected in the first analysis, the script estimated the conservation of these fragments by calculating the proportion of Ns in the fasta nucleotidic sequences resulting from the assemblies. A fragment was considered to be conserved if the reconstructed sequence comprised less than 30 % Ns and not conserved if it comprised more than 50 % Ns. Between these two thresholds, a manual examination was required to determine whether the fragment was conserved. The analyzed accession was then assigned to one of the S-haplotype groups.

Based on shared fragments with reference haplotypes, the analysis did not allow us to orientate the detected fragments compared to each other. To solve this incertitude, we used the BAM files from the assembly data, and extracted the reads at the extremities of each fragment. Then, we searched for reads with one mate on a fragment, and one mate on another fragment. Two fragments were therefore considered to be consecutive if they presented several overlapping reads at their extremities.

Detected types of accessions were then compared to each other, and combined to propose an evolutionary scenario for their generation, involving recombination and deletion events. Polymorphism of the putative ancestral haplotypes of our scenario was analyzed using DnaSP version 5.00.07 (ROZAS *et al.* 2003) and compared with available data from functional relatives (CASTRIC *et al.* 2010) in order to evaluate the influence of the transition to self-compatibility on the genetic diversity.

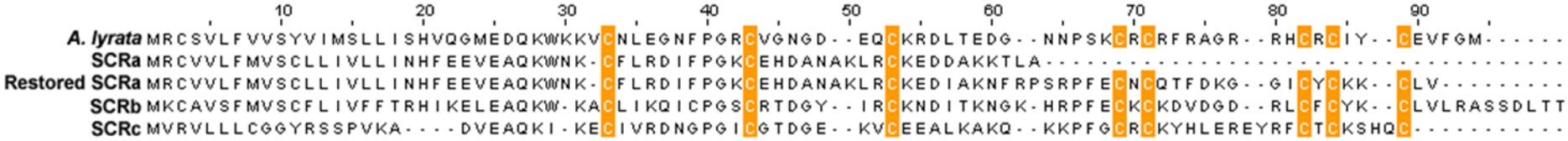


Figure 9. Alignment of SCR protein sequences from *A. thaliana* and *A. lyrata*. ΨSCR$_A$ (Col-0), restored SCR$_A$ (TSUCHIMATSU *et al.* 2010), SCR$_B$ (Cvi-0), SCR$_C$ (Ita-0) and reference *A. lyrata* SCR (AlSCR13) were aligned using Muscle (EDGAR 2004).

IV - RESULTS

A - ANALYSIS OF THE REFERENCE SEQUENCE OF HAPLOGROUP *C*

The sequence of the S-locus genomic region containing haplogroup *C* was obtained from the Ita-0 accession through full sequencing of a BAC (Bacterial Artificial Chromosome) clone from an individual library. Coverage was 22 X and the assembly provided a sequence of 116,959 bp divided into five contigs, ranging from 8,755 to 47,646 bp. The order and orientation of these contigs were successfully confirmed by amplifications in long-range PCR (see Table S4 and Figure S8).

The S-locus region of haplogroup *C* (defined as the divergent region between the two flanking genes, *U-box* and *ARK3*, see Chapter I) was found to be 43,763 bp long. Gene prediction and annotation revealed the presence of the two SI genes, *SCR* and *SRK* as well as several transposable elements (Figure 8). Strikingly, in contrast with other available S-locus sequences in *A. thaliana* (SHERMAN-BROYLES *et al.* 2007; TANG *et al.* 2007; THE *ARABIDOPSIS* GENOME INITIATIVE 2000), no apparent disrupting mutation was found in either *SCR* or *SRK*. Indeed, no stop codon, deletion or inversion was detected in their coding sequence. As in previously described functional haplotypes of *Brassica* or *Arabidopsis* (BOGGS *et al.* 2009a; CUI *et al.* 1999; FUJIMOTO *et al.* 2006; FUKAI *et al.* 2003; KIMURA *et al.* 2002; KUSABA *et al.* 2001; SHIBA *et al.* 2003; SUZUKI *et al.* 1999; TAKUNO *et al.* 2007), the SCR protein presented all eight cysteine residues (Figure 9) that are predicted to form disulfure bridges and be essential for proper SCR folding.

B - CONFIRMATION OF THE RECOMBINATIONAL ORIGIN OF C24

Based on the sequence analysis of Col-0 and C24, a previous study (SHERMAN-BROYLES *et al.* 2007) suggested that C24 had been produced by a recombination event between haplotypes belonging to haplogroups *A*, like Col-0, and *C*. We confirmed this scenario by comparing the Ita-0 sequence with the sequence part of C24 which is not homologous to haplogroup *A*, *i.e.* between exon 5 of *ΔARK3* and *ARK3*. Throughout this portion, C24 and Ita-0 show high sequence conservation (Figure 10) and share most transposable elements as well as partial *SRK* sequences (Figure 9). Further, a more detailed analysis of the chimeric *ΔARK3* sequence in C24 shows that exons 1 to 5 of *ΔARK3* are more closely related to haplogroup *C* than to other haplogroups, while exon 7 is more closely related to haplogroup *A* than to other haplogroups (Figure 11). Altogether, these results strongly support the hypothesis that haplotypes belonging to haplogroups *C* and *A* recombined, using *ΔARK3* as recombination breakpoint, to produce haplotype C24, as proposed by Sherman-Broyles *et al.* (2007). Interestingly though, neither haplogroups *A* nor *C* was found to have a *ΔARK3*, although this has been detected in several functional haplotypes from *A. lyrata* and *A. halleri* (Chapter I).

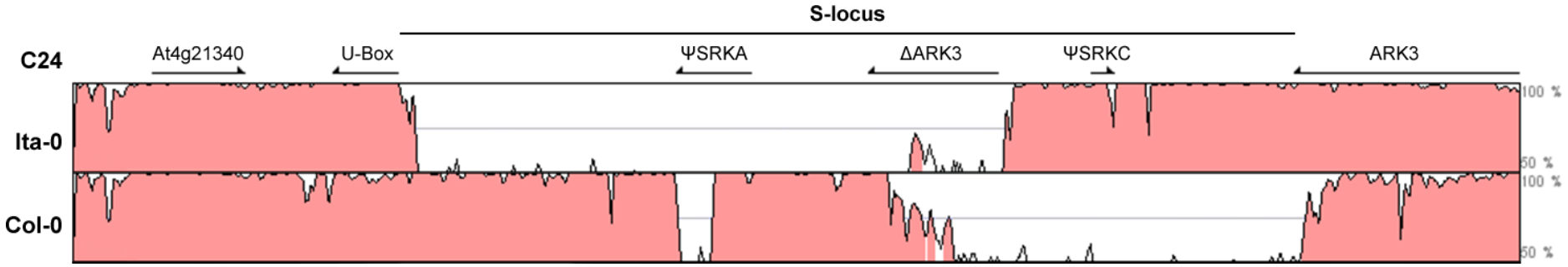


Figure 10. Sequence conservation in the S-locus region between the recombinant C24, and the two haplogroups from which it would have been produced, calculated with the VISTA software (MAYOR *et al.* 2000). Annotation of C24 genes and pseudogenes is indicated above the VISTA diagram.

C - DETECTION OF S-HAPLOTYPES IN EUROPEAN ACCESSIONS

1 - Description of detected haplotypes

Based on full-genome short-read sequencing data, we assembled the S-locus region of 107 accessions from Northern Europe, taking the four known sequences (Col-0, Cvi-0, Ita-0 and C24) as references. These assemblies were obtained with a mean coverage of 18 X. It can be noted that no couple of sequences were found to be exactly identical, even if they were sampled in the same location. A first analysis resulted in the identification of eleven genomic fragments within the S-locus region that belong to either haplogroup *A* or *C* (but not *B*, as expected for north-European accessions (TANG *et al.* 2007)) and are either present or absent in each of the 107 different accessions (Figure 12). Combining this information resulted in ten different groups of haplotypes. Yet, as we explain in the supplemental results, the evolutionary scenario necessary to explain the formation of all these groups from haplogroups *A* and *C* requires a highly unlikely succession of events such as a double recombination followed by at least two homoplasic deletions of the exact same fragment located in Ita-0 between positions 66,042 and 78,204. Further analysis of the reads with a single mate mapped at the extremities of this fragment suggested that it actually originated in Ita-0 by insertion from another genomic location on chromosome 2 that would have been specifically inserted into the S-locus. Hence, considering this fragment as a secondary insertion into Ita-0 and removing it from our analysis, we detected six distinct groups of haplotypes, one belonging to haplogroup *A*, two belonging to haplogroup *C* (*C2* and *C3*), and the last three corresponding to *AC* recombinants (Figure 12). The first group, comprising 53% of the accessions, showed good matches to Col-0 but not to Ita-0 or Cvi-0, and was thus equivalent to haplogroup *A*. We checked that all haplotypes of this group had a pseudogenized SCR_A, as suggested by Tsuchimatsu *et al.* (2010). The second and third groups, comprising respectively 5.6 and 0.9 % of the accessions, showed good matches to Ita-0, but not to Col-0 or Cvi-0, suggesting that they belong to haplogroup *C*. However, none of the haplotypes presented the complete sequence of Ita-0, so they were classified as haplotypes *C2* and *C3* from haplogroup *C*. As compared to Ita-0, *C2* haplotypes lack the 12 kb-fragment we chose to ignore in the analysis, and *C3* haplotypes lack a 31 kb sequence fragment between the *U-box* flanking gene and the Atrep20 transposon. Both SRK_C and SCR_C genes are present in *C2*, but only SRK_C is present in *C3*. All other accessions (40 %) present some fragments with good matches to Col-0 and some others to Ita-0. These haplotypes clearly constitute recombinant haplotypes between haplogroups *A* and *C*, with three different haplotypes named *R1* to *R3*. All recombinants present the fragment from haplogroup *C* between the *ARK3* flanking gene and the Atrep7 transposon, containing a partial sequence of SRK_C, but they differ according to the portion of sequence from haplogroup *A* they contain. The complete Col-0 sequence is conserved in haplotype *R1*, including ΨSRK_A and ΨSCR_A. Haplotypes *R2* and *R3* lack fragments of 18 and 23 kb, respectively, with respect to the Col-0 sequence and they only exhibit partial sequences of ΨSRK_A. According to the conserved fragments, the previously sequenced recombinant haplotype, C24 (SHERMAN-BROYLES *et al.* 2007), would belong to type *R3*. Frequencies of

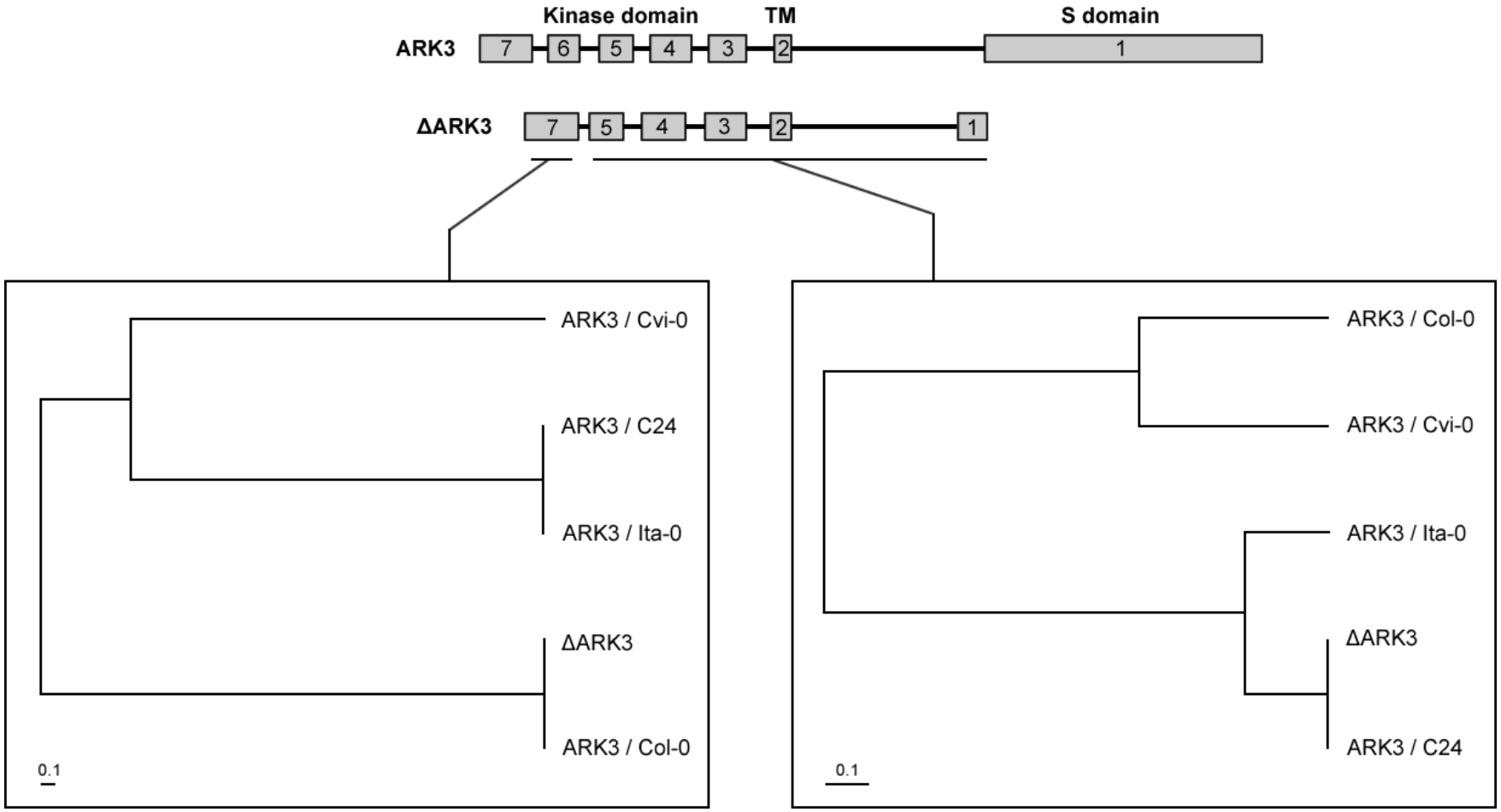


Figure 11. A. Schematized structure of an *ARK3* gene and of *ΔARK3*. Exons are indicated by gray rectangles. TM: Transmembrane domain.
B. Phylogeny of the exon 7 of *ΔARK3* from C24 and *ARK3* from C24, Col-0, Cvi-0 and Ita-0.
C. Phylogeny of the exons 1 to 5 of *ΔARK3* from C24 and *ARK3* from C24, Col-0, Cvi-0 and Ita-0. Both phylogenies were constructed using a Minimum Evolution analysis (TAMURA *et al.* 2011).

the different detected types are indicated on Figure 12, and a complete list of accessions is available in table S5. It can be noted that, if haplogroup *A* and *AC* recombinants are detected across the whole surveyed distribution, none of the accessions from Northern Sweden and Finland were found to carry a haplotype *C2* or *C3*. This observation could result from the small sample of accessions in this Northern region (26 accessions) and the low frequency of haplotypes *C2* and *C3* (7 %). But it could also reflect the fact that haplogroup *C* may have colonized Europe from the Mediterranean refuge (SHERMAN-BROYLES *et al.* 2007). Accessions from Northern Sweden and Finland having potentially colonized Europe from Russia (NORDBORG *et al.* 2005), haplogroups *C* could be expected to be less frequent.

The raw reads were used to orientate the different detected fragments and reconstitute a linear sequence from the set of component fragments shown on Figure 12. By this analysis, consecutive fragments were organized relative to one another through the detection of reads whose two mates mapped to two different fragments. In particular, when considering the assembly taking the C24 sequence as a reference, overlapping reads were found in all recombinant haplotypes between haplogroup *A* and the *ΔARK3* sequence, and between *ΔARK3* and haplogroup *C*, respectively. Thus, all recombinant haplotypes were shown to present a *ΔARK3* sequence (Figure 13).

2 - Evolutionary scenario

Shared fragments between the different groups of haplotypes and the presence of a *ΔARK3* sequence in all recombinant haplotypes suggest that the latter did not appear independently. A simple scenario for their generation can therefore be proposed (Figure 13). According to this scenario, the *C2* haplotype would be ancestral in haplogroup *C*, and would have independently produced haplotypes *C1* (Ita-0) and *C3* by an insertion and a deletion event, respectively. Haplotype *R1* contains the entire sequence of haplogroup *A*, a portion of haplogroup *C* as well as a chimeric duplicated copy of *ARK3* (*ΔARK3*), and would have been generated through recombination between *C2* and A haplotypes, accompanied by several small deletions in the SRK_C sequence. A large deletion from haplotype *R1* would have then produced the *R2* haplotype, followed by heterologous recombination between the two LTRs of the Atcopia10 retrotransposon that would have produced the *R3* haplotype (Figure 13).

We computed the level of nucleotide polymorphism for comparisons among haplotype groups. Nucleotide diversity of the two putative ancestral haplotypes, *A* and *C2*, was found to be sensibly similar with $\Pi = 0.00889$ and $\Pi = 0.00638$, respectively. In these haplotypes, nucleotide diversity was also calculated separately for the S domain of *SRK*. We found $\Pi_S = 0$ and $\Pi_A = 0.00170$ for haplogroup *A*, and $\Pi_S = 0$ and $\Pi_A = 0.00154$ for *C2* haplotypes. By comparison, the synonymous nucleotide diversity within functional haplogroups from *A. halleri* was found to vary from 0 to 0.00581, with on average $\Pi_S = 0.00441$ (CASTRIC *et al.* 2010). This reduction of diversity in *A. thaliana* as compared to *A. halleri* could firstly be explained by a limited geographical origin of our data. But it could also reflect

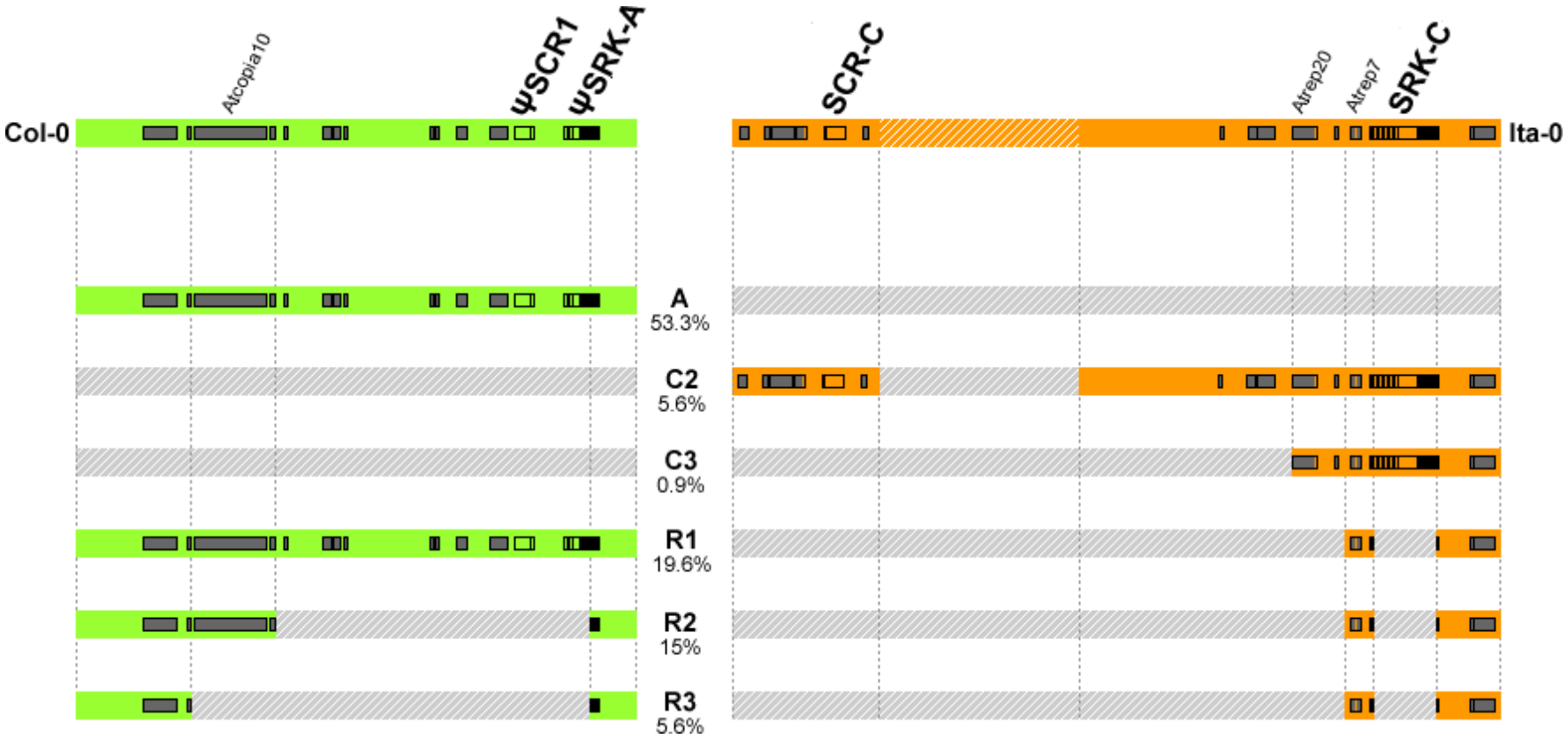


Figure 12. Annotation of the reference haplotypes Col-0 (green) and Ita-0 (orange), and fragments from these haplotypes detected in the six types of accessions. Note that the relative position of the fragments is not respected in this representation. The fragment of Ita-0 which is ignored in the analysis is hatched in white. Frequencies of the detected types in the 107 accessions are indicated below their name.

an hitch-hiking effect due to a rapid increase in frequency of inactivated S-haplotypes in response to selection for self-compatibility in *A. thaliana.*

The presence of *ΔARK3* in all *AC* recombinant haplotypes might reflect the occurrence of a duplicated *ARK3* sequence in the ancestral haplotype of haplogroup *C*. To investigate this issue, we searched for raw reads from *C2* and *C3* haplotypes with one mate on the *ΔARK3*, in the assembly taking C24 as reference, and one mate in the part of haplogroup *C* which is not found in the recombinant in the assembly taking Ita-0 as reference. No such reads were identified, suggesting that none of the *C2* and *C3* haplotypes we analyzed have the duplicated copy of *ARK3*.

V - DISCUSSION

Most studies aiming at characterizing the diversity of the S-locus region in *A. thaliana* used PCR or northern blot approaches, targeting only specific genomic fragments (up to several hundred bp) of the S-locus. Here, we used a combination of data from a new BAC library sequence and from whole genome sequencing of 107 accessions. This constitutes the most important dataset ever analyzed and allowed us to study the diversity of the S-locus on a region of average size 32 kb. This change of scale provides new perspectives to investigate the evolution of self-compatibility in *A. thaliana*.

A - ABSENCE OF APPARENT DISRUPTING MUTATION IN BOTH SCR_C AND SRK_C

Our analysis of the first haplogroup *C* reference sequence did not allow us to point out the cause of the SI breakdown in this specificity, since the two self-incompatibility genes, *SCR* and *SRK*, are both present in the S-locus and seem to have intact coding sequences. In contrast, previous analyses of haplogroup *C* accessions in *A. thaliana* could not detect the presence of the *SCR* gene and suggested that breakdown of incompatibility could have been related to a deletion of this component gene (SHIMIZU *et al.* 2008). We suggest that previous attempts to detect the SCR_C gene failed because of the high sequence divergence among *SCR* sequences, as well as the large fragment of non-coding sequence separating *SCR* from *SRK* (approximately 30 kb). In other haplogroups, potentially functional sequences of the self-incompatibility genes were previously mentioned for either *SRK* (for haplogroup *A* (TSUCHIMATSU *et al.* 2010)), or *SCR* (for haplogroup *B* (TANG *et al.* 2007)). However, no accession carrying both potentially functional *SCR* and *SRK* had already been reported in *A. thaliana*. We cannot exclude the possibility of a disrupting mutation less obvious than a stop codon or a truncated coding sequence, as for instance in the promoter region, but for SRK_C a previous study showed that this gene is actively expressed (SHIMIZU *et al.* 2004). However, if SCR_C and SRK_C were confirmed to be functional, this would imply that the causal disrupting mutation in this accession did not hit the self-incompatibility determinants themselves, but other genes necessary for the functioning of the system.

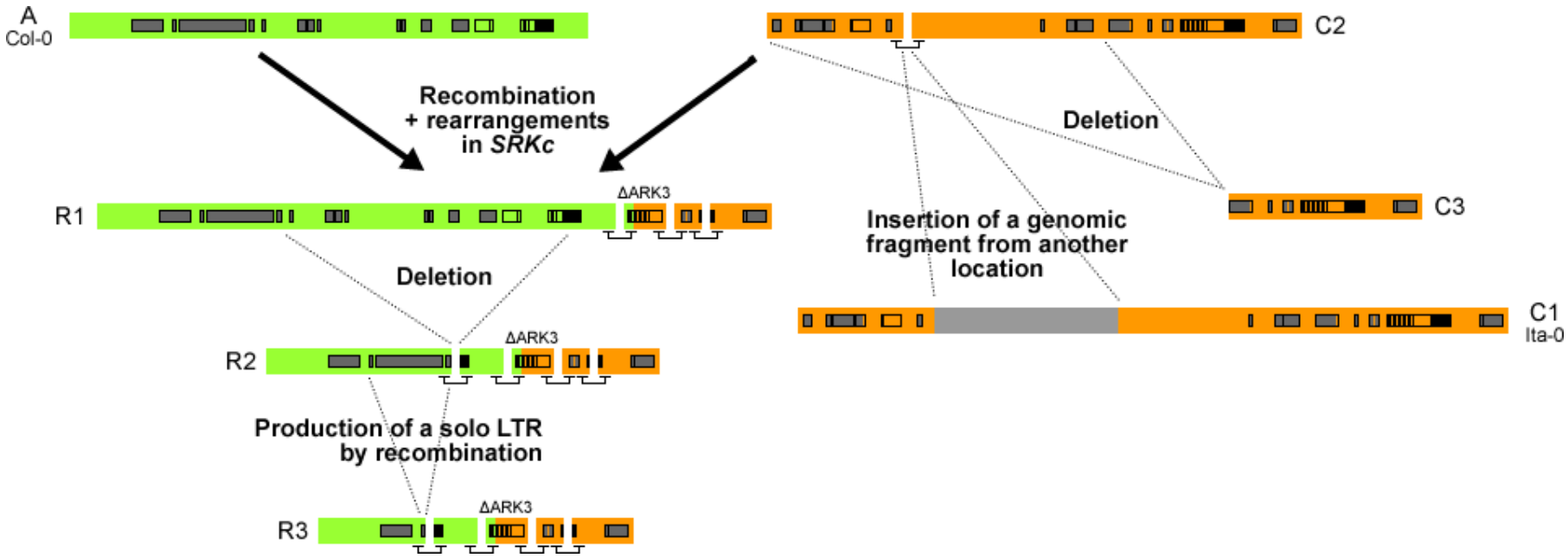


Figure 13. **Scenario for the generation of the different detected haplotypes.** Fragments, which were confirmed to be consecutive by the identification of mates mapping to different fragments, are indicated by brackets.

B - RECOMBINANT HAPLOTYPES ARE MORE FREQUENT THAN PREVIOUSLY DESCRIBED

Investigation of 107 north-European accessions of *A. thaliana* revealed that recombinants between haplogroups *A* and *C* were much more frequent (40 %) than expected based on previous studies. Based on amplification and gel blot analysis of genomic DNA, Sherman-Broyles *et al*. (2007) surveyed 70 accessions distributed across the worldwide distribution area of the species. They estimated that 23 % of these accessions were recombinants and similar to C24. Interestingly, they analyzed 10 accessions of Northern Europe that were present in our data. They considered all these accessions as belonging to haplogroup *A*, whereas our genomic approach detected 5 haplotypes *A*, 4 haplotypes *R1* and one haplotype *R2* in the same sample, *i.e.* 50 % of recombinant haplotypes. Similarly, Shimizu *et al.* (2008) genotyped the S-locus of 297 accessions using primers defined to assay the presence or absence of eight regions of haplogroup *A*, and of the S domain of ΨSRK_C and ΨSRK_B. They concluded that 93.9 % of the accessions belonged to haplogroup *A*, with different deletion types. However, it should be noted that their analysis would not have permitted to detect recombinants like C24, because this type of accessions only shows partial sequences of ΨSRK_C. Indeed, they also analyzed three accessions of Northern Europe that were present in our data (lu-1, st_0 and tamm_2) and assigned them to haplogroup *A,* while we detected that one of them (lu-1) shared also fragments with haplogroup *C* and was a *R1* recombinant haplotype. These results strongly suggest that reports based on targeted PCR or northern blot of short fragments greatly underestimate the importance of recombination among haplogroups, because they only target specific regions. In contrast, genomic methods consider the totality of the region and are more appropriate to detect recombinants.

C - IMPLICATION OF *ΔARK3* IN THE ANCESTRAL RECOMBINATION EVENT

In the S-locus, recombination is expected to be restricted, in order to maintain linkage between *SCR* and *SRK*, and to avoid the breakdown of the system. Due to the loss of self-incompatibility, these selective pressures are supposed to be relaxed in *A. thaliana*. Nevertheless, the high sequence divergence among haplotypes is expected to prevent recombination between them. Thus, a recombination event between haplogroups *A* and *C* would require an element of high similarity between the two sequences. The presence of *ΔARK3* in all types of recombinants suggests that this truncated duplication of *ARK3* constitutes the initial recombination breakpoint, as proposed by Sherman-Broyles *et al.* (2007) based on the analysis of the C24 accession. It can be noted that, even if the precise mechanism of the duplication is not known, such duplications have already been reported in functional haplotypes of *A. halleri* and *A. lyrata* (see Chapter I and also Hagenblad *et al.* (2006) and Guo *et al*. (2011)). By examining the conformation of recombinants, a scenario for the ancestral recombination event can be hypothesized, involving the partial duplication of *ARK3*. As depicted in Figure 14, this partial duplication is thought to have occurred in a haplotype belonging to haplogroup *C*. It would then have allowed a heterologous recombination event with the *ARK3* gene of haplogroup *A*, generating a recombinant haplotype with fragments of both haplogroups *A* and *C*, separated by

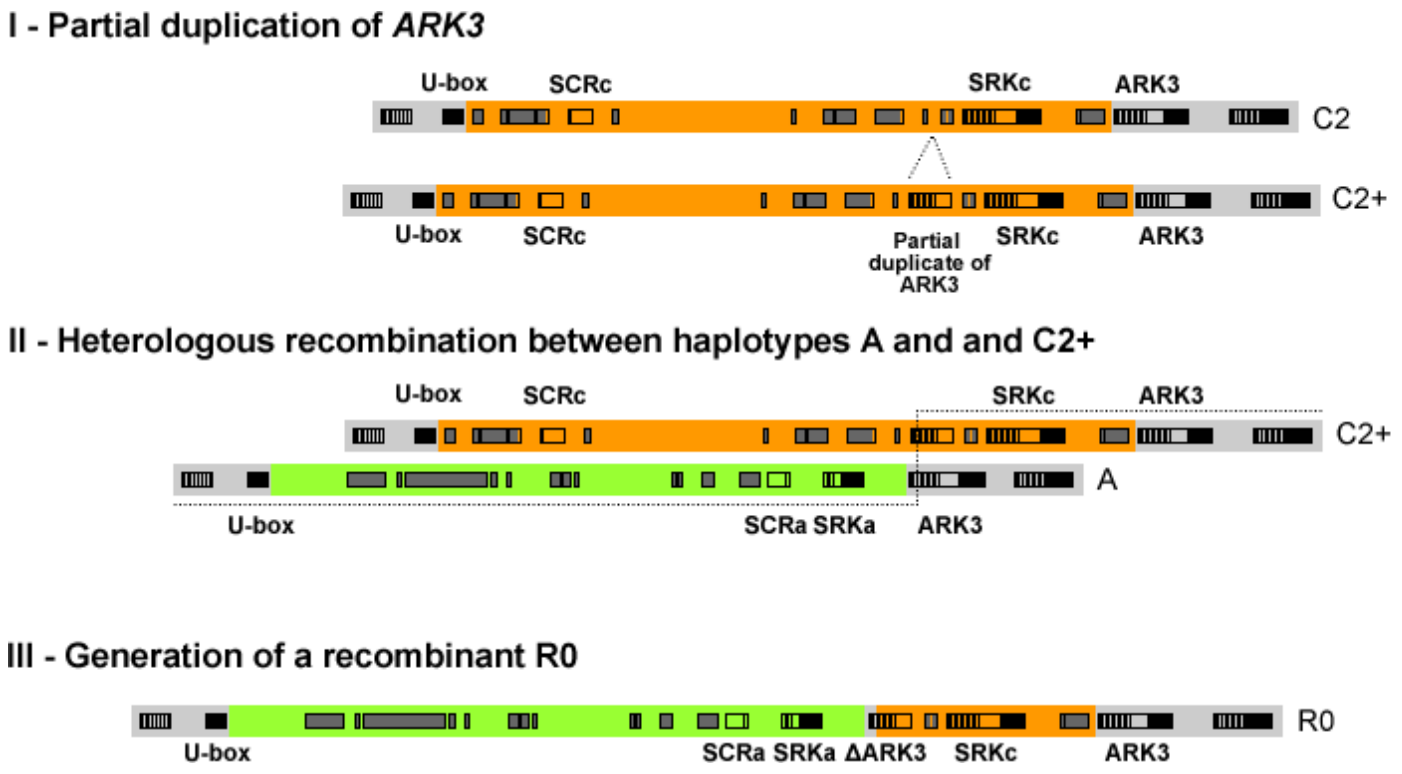


Figure 14. Generation of a recombinant haplotype from haplotypes *A* and *C2*, through a partial duplication of *ARK3*. Flanking regions of the S-locus are depicted in light gray. It has to be noted that types *C2+* and *R0* are hypothetical and were not detected in analyzed accessions.

ΔARK3. However, no haplotype *C2* carrying a partial duplication of *ARK3* in the S-locus (called type *C2+* in the scenario) was detected in our data, so this hypothesis could not be confirmed. Since we had only six accessions of haplotype *C2* in the data we analyzed, investigation of further accessions might help to resolve this issue. The deletion in the sequence of SRK_C in recombinants is another point of incertitude for which analysis of supplementary accessions could be helpful. Indeed, none of the recombinant accessions (*i.e.* 43 accessions) was shown to exhibit a complete SRK_C (such potential accessions were called haplotypes *R0* in the scenario) and none of the ancestral haplotype *C2* presented this deletion. It might suggest that this rearrangement occurred at the same time as the recombination event or that recombinants carrying functional *SRK* sequences were selected against.

D - THE EVOLUTION OF SELFING IN *A. THALIANA*

Some implications for the evolution of selfing in *A. thaliana* can be highlighted by our results. First, the observation of a full coding sequence of *SCR* in haplogroup *C* is one additional argument in favor of a recent loss of self-incompatibility in the *A. thaliana* lineage (BECHSGAARD *et al.* 2006; CHARLESWORTH and VEKEMANS 2005; TANG *et al.* 2007). Indeed, the fact that SRK_C has an intact coding sequence and is expressed in pistils (SHIMIZU *et al.* 2004) implies that the transition to selfing is not ancient enough to have allowed the accumulation of disrupting mutations in both *SCR* and *SRK* in this haplogroup. Second, we found an unexpected high frequency of *AC* recombinants in natural populations. Consequently, the absence of accessions from haplogroup *C* with apparent disrupting mutation suggests that the recombination may have played a role in the evolution of selfing. Recombinants could thus have increased in frequency due to the benefits of compatibility but it would imply that at least one of the two ancestral haplogroups was still functional when the initial recombination occurred. Because all recombinant haplotypes *R1* carried the non-functional ΨSCR_A sequence, this suggests that recombination occurred in an *AC* heterozygote with a functional *C* haplotype and a non-functional *A* haplotype. This parental individual should have been self-incompatible as haplogroup *C* belongs to a phylogenetic group of haplotypes that is dominant, in *A. lytata*, over that of haplogroupe *A* (BECHSGAARD *et al.* 2006). This hypothesis is supported by the low frequency of haplogroup *C*, which could reflect a breakdown of self-incompatibility in a second step.

VI - PERSPECTIVES AND FUTURE DIRECTIONS

The availability of reference sequences from two of the three S-haplogroups, and accumulation of data from the 1001 genomes project open new perspectives in the study of the breakdown of self-incompatibility in the recent history of *A. thaliana*. First, we produced a reference sequence for the third haplogroup (*C*) and showed that SCR_C and SRK_C were potentially functional in accession Ita-0. Hence, we could not identify the cause of SI breakdown in haplogroup *C* accessions. A functional analysis of the self-incompatibility genes in this haplogroup would therefore be needed in order to

point out the initial disrupting mutation. Second, the three main haplogroups in *A. thaliana* are the remnant of the ancestral polymorphism of the S-locus. Thus, comparison of these haplogroups with their functional orthologs in the closely related self-incompatible species *A. lyrata* and *A. halleri* could provide new insights in the patterns of degeneration of an inactivated S-locus. A reference sequence for the ortholog of haplogroup *B* in *A. lyrata*, *Al16*, was recently made available (GUO *et al.* 2011) but ortholog sequences of haplogroups *A* (*Ah04* in *A. halleri* and *Al37* in *A. lyrata*) and *C* (*Al36* in *A. lyrata*) are still lacking. Third, we analyzed a large number of accessions in Northern Europe and found evidence for a single recombination event between haplogroups *A* and *C*, as well as several deletions. Analysis of supplementary accessions from a larger set of locations could lead to detection of other recombination events, potentially due to other *ARK3* duplications. Moreover, examination of supplementary accessions could also provide evidence of the intermediate haplotypes (*C2+* and *R0*) that we have proposed in our scenario.

CHAPITRE III - COEVOLUTION DES PROTEINES POLLEN ET PISTIL AU LOCUS D'AUTO-INCOMPATIBILITE CHEZ *ARABIDOPSIS*

Contributions :

Utilisation du programme Comap sur les données : Julien Dutheil[1,2]

Analyse des résultats, interprétation et rédaction : Pauline Goubet[3], Xavier Vekemans[3] et Vincent Castric[3]

Affiliations :

[1] Max Planck Institute for Terrestrial Microbiology, Department of Organismic Interactions, Karl-von-Frisch-Straße 10, D-35043 Marburg, Germany

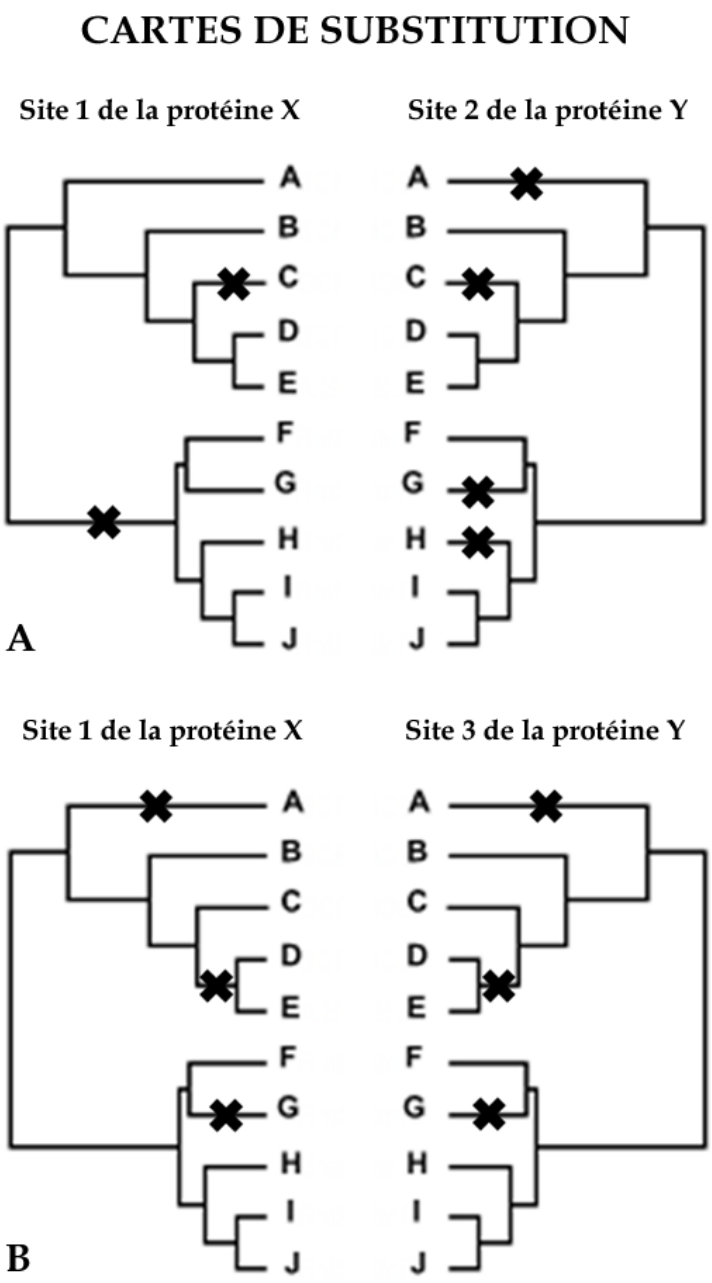
[2] Institut des Sciences de l'Evolution de Montpellier, CNRS UMR 5554, Bâtiment 22, Université des Sciences et Techniques du Languedoc, Place E. Bataillon, F-34095 Montpellier Cedex 5, France

[3] Laboratoire de Génétique et Evolution des Populations Végétales, CNRS FRE 3268, Bâtiment SN2, Université Lille Nord de France, Cité scientifique, F-59655 Villeneuve d'Ascq Cedex, France

I - INTRODUCTION

Les études en biologie évolutive ont souvent été considérées à l'échelle d'une espèce, d'un organisme ou d'un gène alors que ces entités sont impliquées dans des réseaux d'interaction complexes, dont il est maintenant largement admis qu'ils doivent être pris en compte. En biologie, la coévolution peut être définie comme étant l'évolution des fréquences d'un ensemble de mutations dans une de ces entités, en réponse à l'évolution des fréquences de mutations dans une autre entité qui lui est liée. De tels changements évolutifs ont été largement étudiés entre espèces différentes, par exemple à travers les interactions entre les hôtes et les parasites (HAFNER and NADLER 1988), entre les prédateurs et leurs proies (BARRIO *et al.* 2010), entre les plantes et leurs pollinisateurs (JOHNSON and ANDERSON 2010), ou entre symbiontes (MOYA *et al.* 2008). La coévolution peut également être étudiée au niveau moléculaire. En effet, différentes positions au sein d'une macromolécule peuvent être soumises à des contraintes structurelles, si elles se retrouvent à proximité au niveau de sa structure tridimensionnelle, ou à des contraintes fonctionnelles, si elles sont impliquées dans un même site actif. Différentes approches méthodologiques ont été développées pour mettre en évidence de la coévolution entre résidus d'acides aminés, dont la méthode des cartes de substitution, consistant à tester la corrélation de l'occurrence des mutations sur les branches des phylogénies de plusieurs sites (DUTHEIL and GALTIER 2007; DUTHEIL *et al.* 2005). Par cette approche, des groupes de sites en situation de coévolution ont par exemple été mis en évidence dans le cas des myoglobines (DUTHEIL and GALTIER 2007; NEHER 1994), des protéines impliquées dans le transport et le stockage de l'oxygène chez les vertébrés. A une échelle plus large, les positions d'une molécule peuvent être impliquées dans une interaction physique avec d'autres molécules, comme par exemple dans les systèmes génétiques de type récepteur-ligand (aussi appelés systèmes clé-serrure).

Par leur reconnaissance de type clé-serrure lorsque le grain de pollen est déposé sur le pistil, les protéines SCR et SRK constituent un modèle pertinent pour une analyse de coévolution. En effet on observe au sein des espèces auto-incompatibles chez les Brassicaceae une très grande diversité de spécificités correspondant à des paires SCR/SRK montrant une évolution conjointe (voir Sato *et al.* (2002) et le Chapitre I). En particulier, des travaux théoriques suggèrent une dynamique de diversification impliquant des mutations au niveau des séquences pollen, suivies de mutations compensatrices dans les séquences pistil associées (GERVAIS *et al.* 2011; UYENOYAMA *et al.* 2001). Cependant, dans le genre *Arabidopsis*, un nombre très restreint de séquences du gène *SCR* était jusqu'à ce jour disponible dans la littérature (BOGGS *et al.* 2009a; BOGGS *et al.* 2009b; KUSABA *et al.* 2001; TSUCHIMATSU *et al.* 2010) en raison de son extrême divergence. Grâce au séquençage de la région complète du locus d'auto-incompatibilité pour onze haplotypes différents, l'approche génomique employée ici a l'intérêt de permettre d'accéder aux séquences du gène pollen. Dans ce contexte, l'acquisition d'un nombre conséquent de couples de gènes pollen *SCR* et pistil *SRK* nous permet d'effectuer une analyse préliminaire de la coévolution nécessaire au maintien de la reconnaissance de



Encadré 9. Exemple de cartes de substitutions illustrant deux sites indépendants (A) et deux sites qui coévoluent (B).

Cet exemple suppose une analyse de la coévolution entre une protéine X et une protéine Y pour lesquelles dix séquences sont disponibles. Les cartes établies attribuent dans ce cas précis le même poids à toutes les mutations, représentées par des croix. Dans le premier cas, les mutations ne se situent pas sur les mêmes branches de l'arbre et les deux sites sont donc jugés indépendants. A l'inverse, dans le second cas, les deux sites montrent des substitutions sur les mêmes branches de l'arbre. Il y a donc coévolution.

Les données des cartes de substitution sont utilisées par le programme Comap (DUTHEIL *et al.* 2005; DUTHEIL & GALTIER 2007) sous forme de vecteurs de substitution.

leurs protéines et à l'évolution de nouvelles spécificités. Pour cela, le programme Comap (DUTHEIL and GALTIER 2007; DUTHEIL *et al.* 2005) a été utilisé sur les seize couples de protéines SCR-SRK disponibles chez *A. halleri* et *A. lyrata*.

II - MATERIEL ET METHODES

A - STRUCTURE ET FONCTION DES PROTEINES SCR ET SRK

D'un point de vue fonctionnel, la protéine SRK est composée de trois domaines : le domaine extracellulaire (domaine S) permet à la protéine du pollen de se fixer sur celle du pistil ; le domaine transmembranaire permet l'ancrage de la protéine SRK à la surface du pistil et le domaine intracellulaire (domaine kinase) est impliqué dans la transmission du signal d'auto-incompatibilité. Le domaine S contient lui-même trois sous-domaines fonctionnels distincts : un sous-domaine B-lectin, un sous-domaine glycoprotéique (SLG) et un sous-domaine PAN-APPLE (NAITHANI *et al.* 2007). De plus, des régions hypervariables ont été définies au sein du domaine extracellulaire dans le genre *Brassica* (HVR1, HVR2, HVR3 et CVR (NISHIO and KUSABA 2000)) et comprendraient des sites particulièrement impliqués dans la fixation du ligand SCR sur son récepteur SRK (KEMP and DOUGHTY 2007).

La protéine SCR se distingue quant à elle par un grand nombre de cystéines relativement à sa petite taille. Elle comporte en effet dans la plupart des cas huit résidus de cystéine parfaitement conservés au sein de son deuxième exon et qui forment deux à deux des ponts disulfures influençant la structure tridimensionnelle de la protéine (CHOOKAJORN *et al.* 2004). Une certaine variation de ce nombre de cystéines est cependant possible puisque des séquences protéiques comportant sept ou neuf résidus ont été reportés chez *Brassica oleracea* (SATO *et al.* 2002). Dans nos données relatives à *A. halleri* et *A. lyrata*, seule la séquence AlSCR39 se différencie des autres en ne présentant que six résidus conservés.

B - ORIGINE DES SEQUENCES SCR ET SRK

Dans le cadre de cette analyse, seize couples de protéines SCR-SRK d'*A. halleri* et *A. lyrata*, provenant des données de cette thèse et de la littérature, ont été utilisés (voir le Tableau 3 pour le détail). Ces données comprennent un couple de classe I, cinq couples de classe II, quatre couples de classe III et six couples de classe IV. L'exon 1 de *AhSCR43* n'ayant pas été détecté dans la séquence de l'haplotype *Ah43* (voir le chapitre I), le couple AhSCR43-AhSRK43 n'a pas été intégré à l'analyse.

C - RECONSTRUCTIONS PHYLOGENETIQUES POUR SCR ET SRK

La méthode de détection de la coévolution (DUTHEIL and GALTIER 2007; DUTHEIL *et al.* 2005) se base sur des corrélations phylogénétiques et son application nécessite que les phylogénies des protéines

Tableau 3. Provenance des couples de protéines SCR-SRK analysés dans ce chapitre.

Haplotype	Classe	SCR	SRK
Al01	I	Données de la thèse	Données de la thèse
Al06	II	ACU29640 (BOGGS *et al.* 2009a)	ACU29642 (BOGGS *et al.* 2009a)
Al14	II	Données de la thèse	Données de la thèse
Al18	II	Données de la thèse	Données de la thèse
Ah03	II	Données de la thèse	Données de la thèse
Ah28	II	Données de la thèse	Données de la thèse
Al13	III	BAB40984 (KUSABA *et al.* 2001)	BAB40986 (KUSABA *et al.* 2001)
Al25	III	ACU29641 (BOGGS *et al.* 2009a)	ACU29643 (BOGGS *et al.* 2009a)
Al37	III	ACN63521 (BOGGS *et al.* 2009b)	ABF71379 (BECHSGAARD *et al.* 2006)
Ah04	III	ADG01814 (TSUCHIMATSU *et al.* 2010)	ABF71368 (BECHSGAARD *et al.* 2006)
Al20	IV	BAB40985 (KUSABA *et al.* 2001)	BAB40987 (KUSABA *et al.* 2001)
Al39	IV	Données de la thèse	Données de la thèse
Ah13	IV	Données de la thèse	Données de la thèse
Ah15	IV	Données de la thèse	Données de la thèse
Ah20	IV	Données de la thèse	Données de la thèse
Ah32	IV	Données de la thèse	Données de la thèse

[a] La fin de l'exon 7 n'est pas disponible pour AhSRK04 et AlSRK37.

SCR et SRK présentent des topologies congruentes. Pour vérifier cela, les séquences protéiques de SCR et SRK ont été alignées en utilisant le programme Muscle (EDGAR 2004) implémenté sur Jalview (WATERHOUSE *et al.* 2009). Après ajustement manuel, les phylogénies ont été construites par la méthode de « Minimum Evolution » implémentée sur MEGA (TAMURA *et al.* 2011). La congruence de leur topologie a ensuite été testée par la méthode de De Vienne *et al.* (2007), basée sur une mesure appelée MAST (Maximum Agreement SubTrees). Cette méthode détermine pour un grand nombre de couples de phylogénies générées au hasard le nombre minimum de branches qui doivent être supprimées pour que les deux arbres soient identiques. La distribution de ces mesures est ensuite comparée à la valeur observée, et il peut ainsi être déterminé si un couple d'arbres est plus congruent que ce que l'on attendrait par hasard.

D - ANALYSE DE COEVOLUTION

Sur la base des alignements des deux protéines et de leur phylogénie, le programme Comap estime pour chaque site ne comportant pas de gaps le nombre de substitutions associées à chaque branche de l'arbre, sous forme d'une carte de substitution. Le degré de coévolution de chaque groupe de sites est calculé sur la base de la comparaison des cartes de substitution des sites concernés (voir l'encadré 9). La significativité des groupes est alors évaluée par une méthode de bootstraps paramétriques s'appuyant sur un jeu de 100 000 simulations présentant les mêmes paramètres que les données analysées. Une correction pour tests multiples est également appliquée dans chaque cas.

Deux analyses différentes peuvent être effectuées, et ont été lancées sur nos données : l'approche par paires (DUTHEIL *et al.* 2005) et l'approche par clusters (DUTHEIL and GALTIER 2007). L'approche par paires permet de détecter des couples de sites en situation de coévolution entre deux protéines distinctes. Elle a donc été utilisée entre les protéines du pollen et du pistil. L'approche par clusters a quant à elle l'avantage de permettre la détection de groupes de sites, et non plus seulement de couples, qui coévoluent au sein d'une même séquence. Cette analyse a été réalisée sur un alignement concaténé des deux protéines étudiées, et permet de détecter des groupes de sites en coévolution entre SCR et SRK, mais aussi au sein de SCR ou de SRK. Par souci d'optimisation du temps de cette seconde analyse, il a été choisi de ne pas y intégrer le domaine transmembranaire et le domaine kinase de SRK, qui ne sont *a priori* pas impliqués dans la reconnaissance du pollen et du pistil.

Ces analyses tiennent compte des propriétés biochimiques des acides aminés, à savoir le volume, la charge, la polarité et la distance chimique Grantham, qui est une mesure synthétique combinant le volume, la polarité et la composition atomique (GRANTHAM 1974). Différentes cartes de substitutions sont donc établies et comparées selon qu'elles attribuent soit le même poids à toutes les substitutions, soit des poids différents selon les propriétés biochimiques des acides aminés mis en cause. Elles permettent de mettre en évidence deux types de sites en situation de coévolution, c'est-à-dire qui présentent des co-substitutions : les évènements de substitution tendent à se trouver sur les mêmes

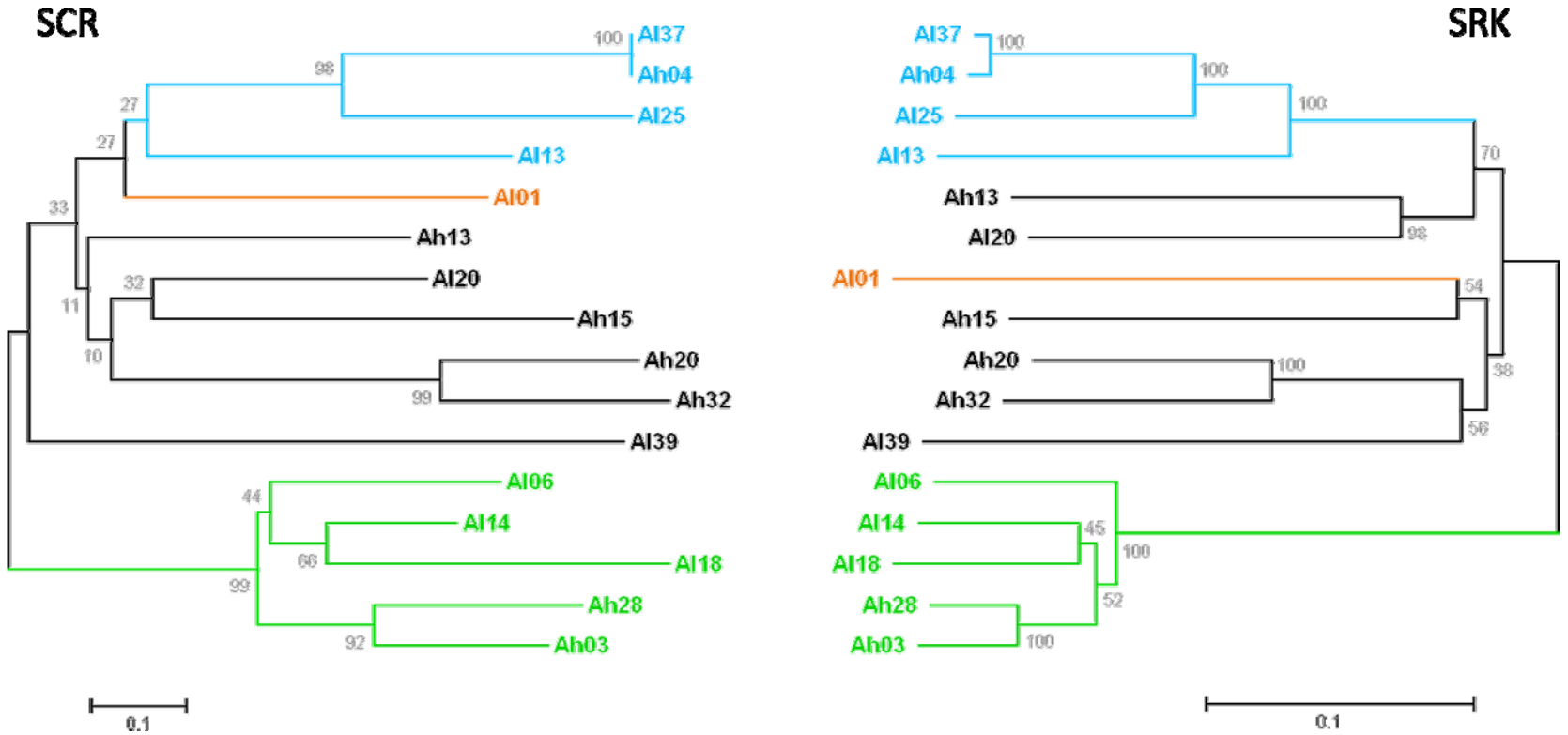


Figure 15. Phylogénie des protéines SCR et SRK analysées. Les séquences protéiques ont été alignées à l'aide du programme Muscle (EDGAR 2004), et l'arbre construit en utilisant la méthode « Minimum Evolution » implémentée sur MEGA (TAMURA *et al.* 2011). Les quatre classes de dominance sont indiquées par des couleurs différentes : la classe I en orange, la classe II en vert, la classe III en bleu et la classe IV en noir.

branches de l'arbre phylogénétique. Tout d'abord, des sites peuvent être détectés comme coévoluant sous corrélation s'ils présentent des co-substitutions simples ou associées à des corrélations des propriétés biochimiques (charge, polarité, volume et distance Grantham). Par ailleurs, la coévolution par compensation peut également être détectée en fonction de ces mêmes propriétés biochimiques. Dans le cas du volume par exemple, il s'agit de co-substitutions mettant en cause des groupes de sites qui conservent un volume global fixe alors que les volumes individuels de chacun de ces sites varient. Ce type de coévolution est principalement détecté dans le cas de contraintes structurales ou de sélection négative. Dans notre cas d'étude, on peut supposer que la compensation est principalement liée au maintien de la reconnaissance entre les deux protéines d'une spécificité, alors que la corrélation est davantage liée à l'évolution de nouvelles spécificités.

E - REPARTITION DES SITES DETECTES

Afin de tester si les sites détectés par l'analyse réalisée se répartissaient préférentiellement dans les régions hypervariables définies dans le domaine S de SRK chez *Brassica* (NISHIO and KUSABA 2000), l'approche par sites candidats de Comap a été utilisée. Cette analyse permet de désigner des couples de sites et de mesurer leur degré de coévolution. Nous avons donc sélectionné un nombre arbitraire de 200 couples de positions SCR-SRK à partir des alignements des deux protéines. La moitié de ces couples implique des positions situées dans les régions hypervariables de SRK (NISHIO and KUSABA 2000) et l'autre moitié implique des positions situées dans le reste du domaine S. Les statistiques de coévolution de chacun de ces couples ont été calculées et un test de Student a permis de déterminer si l'un ou l'autre des deux groupes présentait un degré de coévolution significativement plus important.

III - RESULTATS

A - CONGRUENCE DES PHYLOGENIES

Dans un premier temps, les phylogénies respectives des protéines SCR et SRK ont été comparées (Figure 15). Si la congruence n'est pas parfaite entre les deux topologies, on remarque néanmoins que les quatre classes de dominance, qui avaient été définies d'après la phylogénie du gène *SRK* (PRIGODA *et al.* 2005), sont conservées chez SCR comme précédemment montré dans le chapitre I. De plus, ces phylogénies sont jugées plus congruentes qu'attendu par hasard : I_{cong} = 1.72, P-value = 0.0001 (DE VIENNE *et al.* 2007). Cette congruence justifie donc l'utilisation de l'approche par carte de substitutions dans le cadre de la détection de sites sous coévolution. La séquence de SCR étant relativement petite (environ 80 acides aminés) par rapport à celle de SRK (environ 850 acides aminés), il a été choisi d'utiliser la phylogénie de SRK dans les analyses décrites ci-dessous.

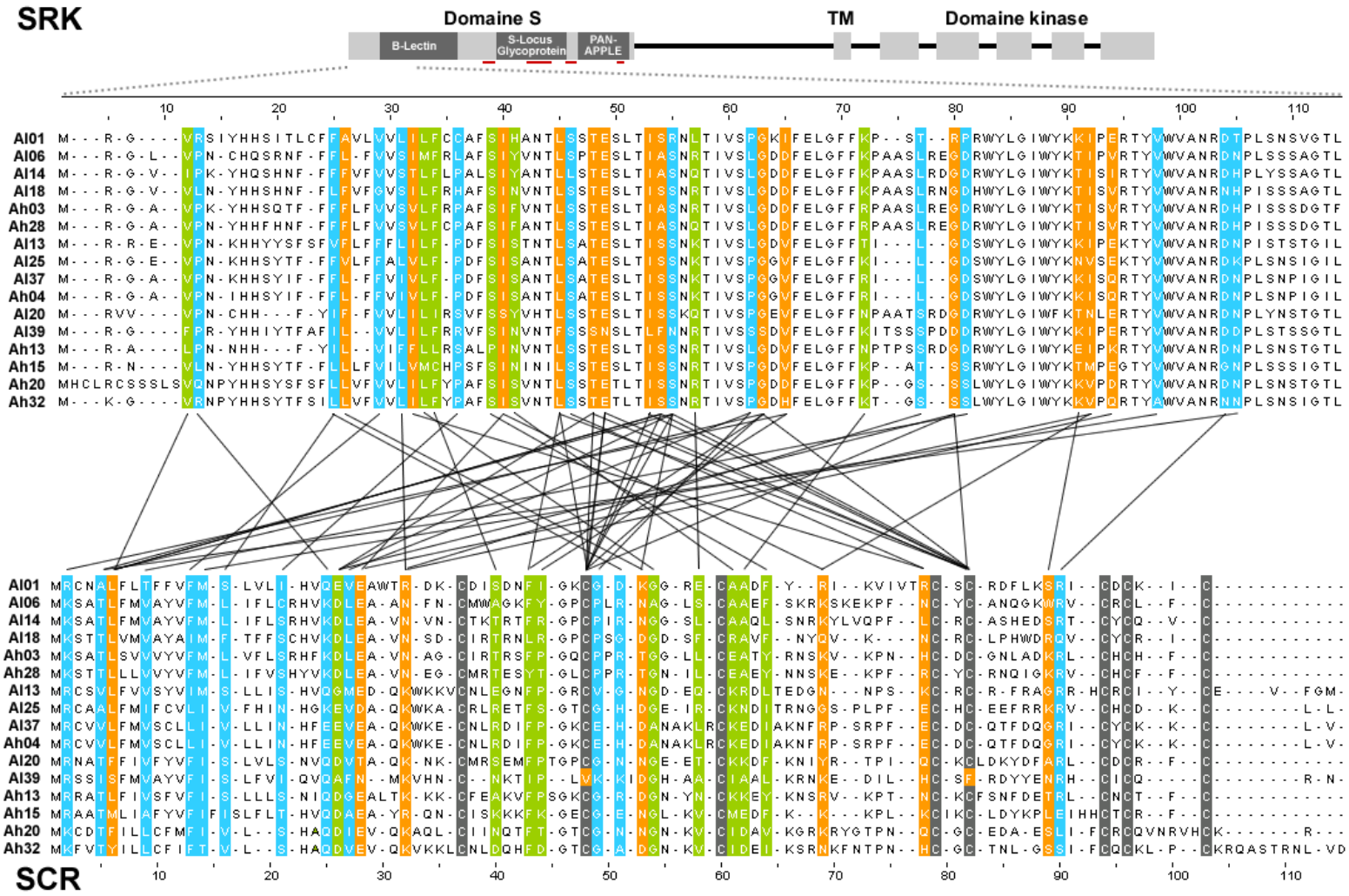


Figure 16.A. Couples de positions montrant un signal de coévolution entre les protéines SCR et SRK, dans le cadre de l'analyse par paires (DUTHEIL *et al.* 2005). Cette représentation prend en compte la totalité de la protéine SCR, mais seulement les positions 1 à 115 de la protéine SRK. Une représentation schématique de *SRK*, comprenant les différentes régions fonctionnelles du domaine S en gris foncé (MARCHLER-BAUER *et al.* 2011) et les régions hypervariables chez *Brassica* (HVR1, HVR2, HVR3 et CVR respectivement (NISHIO & KUSABA 2000)) en rouge, est incluse en haut de la figure afin de fournir plus de précision quant à la localisation de la région étudiée. En vert sont représentées les positions détectées par compensation, en bleu les positions détectées par corrélation, et en orange les positions détectées dans les deux cas. Par ailleurs, les huit résidus de cystéine caractéristiques de la plupart des protéines SCR sont indiqués en gris. TM : Domaine transmembranaire.

B - ANALYSE PAR PAIRES

Utilisée entre les protéines SCR et SRK, l'analyse par paires a permis de détecter un grand nombre de couples de sites montrant un signal de coévolution : 399 couples sous corrélation et 1 361 sites sous compensation (P-value < 0.05). Cependant, dans chacun de ces cas, aucun couple ne reste significatif après correction pour tests multiples. En choisissant un seuil arbitraire de P-value = 0.01, 64 et 35 couples de sites sont respectivement détectés sous corrélation et sous compensation (voir l'annexe 1 pour le détail). De manière intéressante, tous les sites détectés sur la protéine SRK avec une P-value inférieure à 0.01 se répartissent en seulement deux ensembles clairement distincts dans le domaine S, le premier entre les positions 12 et 105 de l'alignement (Figure 16.A) et le second entre les positions 200 et 279 (Figure 16.B). Le premier ensemble de sites comprend le début de la région fonctionnelle B-lectin et le second comprend le début de la région fonctionnelle SLG (S-Locus Glycoprotein). La région hypervariable HVR1 et une partie de la région HVR2, définies chez *Brassica* (NISHIO and KUSABA 2000), sont également comprises dans ce second ensemble. Aucun site potentiellement sous coévolution n'est détecté dans le reste du domaine S de la protéine SRK, y compris dans les régions hypervariables HVR3 et CVR (NISHIO and KUSABA 2000), ou dans les domaines transmembranaire et kinase. Au niveau de la protéine SCR, les sites détectés se répartissent sur l'ensemble de la séquence. Cependant, il peut être remarqué que les sites comprenant les deux résidus de cystéine qui ne sont pas conservés chez AlSCR39 (positions 48 et 82 sur l'alignement) sont détectés comme coévoluant avec de nombreux autres sites, que ce soit par corrélation ou par compensation. De manière plus générale, 30 sites sont détectés à la fois sous corrélation et sous compensation. 69 sites sont détectés uniquement sous corrélation et 54 sites uniquement sous compensation.

C - ANALYSE PAR CLUSTERS ENTRE SCR ET SRK

L'analyse par clusters a été utilisée sur la totalité de la protéine SCR et sur le domaine S de la protéine SRK. Entre SCR et SRK, 9 et 16 groupes pouvant comprendre jusque 8 sites ont respectivement été détectés par corrélation et par compensation (Figures 17.A et 17.B ; voir également l'annexe 1 pour le détail). Contrairement à l'analyse par paires, ces groupes se répartissent dans la totalité du domaine S de SRK et ne sont pas restreints à des zones particulières. On retrouve ainsi des sites potentiellement sous coévolution dans les régions fonctionnelles B-lectin, SLG ou PAN-APPLE ainsi que dans les régions hypervariables HVR1, HVR2 et CVR (NISHIO and KUSABA 2000). Au niveau de SCR, il peut être remarqué que les positions 48 et 82, caractérisées par la perte de deux résidus de cystéine chez AlSCR39, montrent un signal de coévolution par corrélation avec un groupe de quatre sites sur SRK. D'un point de vue général, 7 sites sont à la fois détectés sous corrélation et sous compensation. 21 sites le sont uniquement sous corrélation et 34 sites uniquement sous compensation.

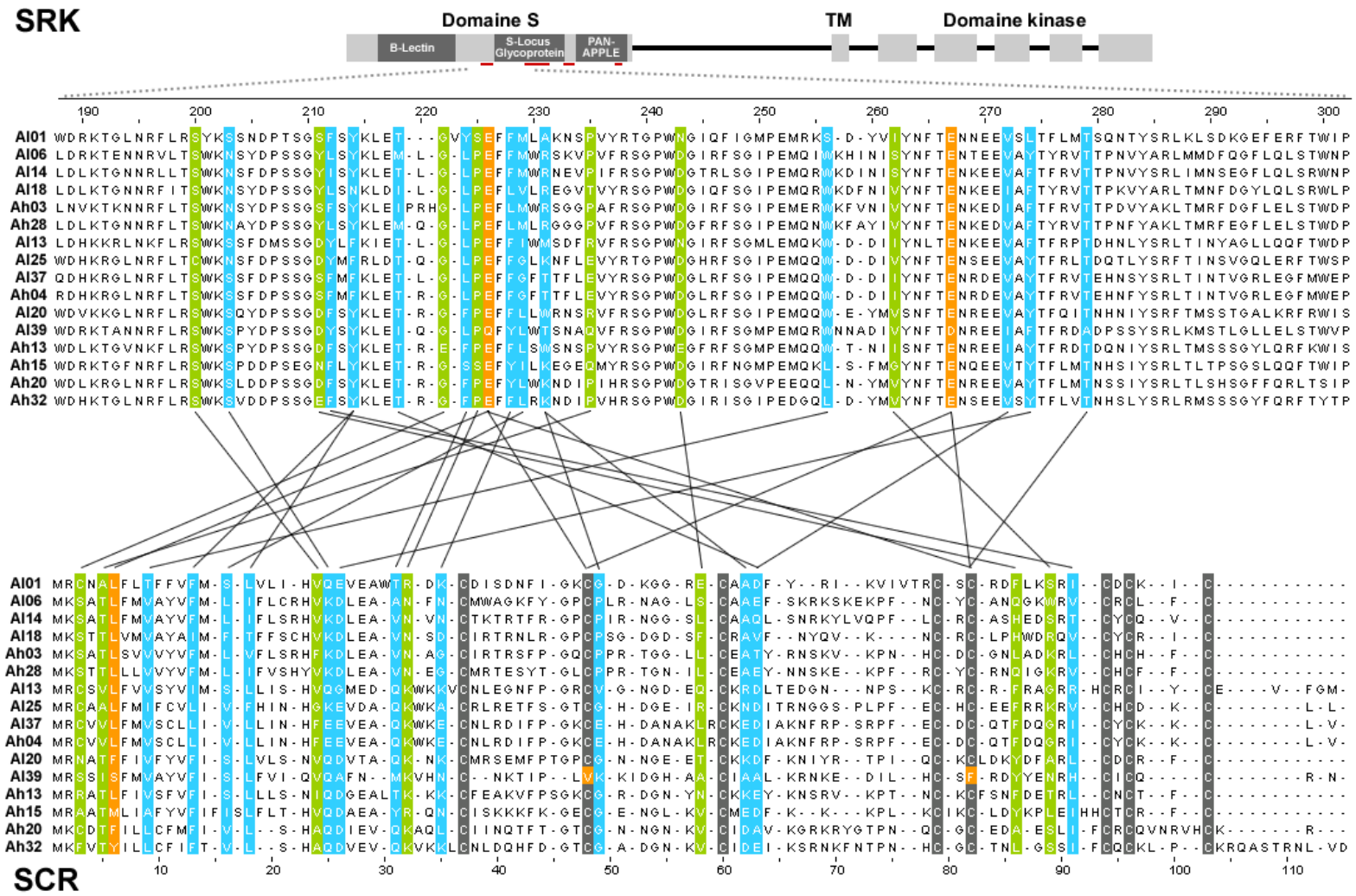


Figure 16.B. Couples de positions montrant un signal de coévolution entre les protéines SCR et SRK, dans le cadre de l'analyse par paires (DUTHEIL *et al.* 2005). Cette représentation prend en compte la totalité de la protéine SCR, mais seulement les positions 188 à 302 de la protéine SRK. Une représentation schématique de *SRK*, comprenant les différentes régions fonctionnelles du domaine S en gris foncé (MARCHLER-BAUER *et al.* 2011) et les régions hypervariables chez *Brassica* (HVR1, HVR2, HVR3 et CVR respectivement (NISHIO & KUSABA 2000)) en rouge, est incluse en haut de la figure afin de fournir plus de précision quant à la localisation de la région étudiée. En vert sont représentées les positions détectées par compensation, en bleu les positions détectées par corrélation, et en orange les positions détectées dans les deux cas. Par ailleurs, les huit résidus de cystéine caractéristiques de la plupart des protéines SCR sont indiqués en gris. TM : Domaine transmembranaire.

D - ANALYSE PAR CLUSTERS AU SEIN DE SRK

L'analyse par clusters a également permis de détecter des groupes de sites n'impliquant que des groupes de positions sur SRK (voir l'annexe 1 pour la liste complète des groupes détectés). Ainsi, 11 et 25 groupes de sites ont respectivement été détectés sous corrélation et sous compensation à l'intérieur du domaine S de SRK (Figure supplémentaire 11). Ces groupes, constitués de 2 à 10 sites, semblent se répartir de manière relativement homogène dans le domaine S de SRK et ne se retrouvent pas préférentiellement dans l'une de ses régions fonctionnelles ou dans les régions hypervariables. A l'inverse, aucun groupe n'impliquant que des sites sur SCR n'a été mis en évidence par cette analyse.

E - COMPARAISON DE L'ANALYSE PAR PAIRES ET DE L'ANALYSE PAR CLUSTERS

De manière globale, les analyses par paires et par clusters semblent apporter des résultats différents dans la détection de sites sous coévolution entre les protéines SCR et SRK (Figure 18) et la question de la cohérence entre ces deux types d'approches peut être posée. Au niveau de la protéine SRK, les sites détectés dans le cadre de l'analyse par paires sont très localisés alors qu'ils sont dispersés dans la totalité du domaine S dans le cadre de l'analyse par clusters. Au total, 55 sites sont détectés dans le domaine S par l'analyse par paires, et 43 sites par l'analyse par clusters lorsque l'on ignore les groupes détectés uniquement au sein de la protéine du pistil. 11 sites sont communs entre les deux analyses. La probabilité de tirer ainsi au moins 11 sites communs par deux tirages aléatoires de 55 et 43 sites parmi les 416 sites sans gaps de l'alignement étant faible (loi hypergéométrique, probabilité = 0.0157), on peut en déduire que les deux analyses sont cohérentes entre elles. Au niveau de SCR, 38 et 17 sites sont respectivement mis en évidence par la méthode par paires et la méthode par clusters, avec 13 de ces sites communs entre les deux analyses. De la même manière que pour SRK, la probabilité de tirer au moins 13 sites en commun dans deux tirages indépendants de 38 et 17 sites parmi 64 sites est très faible (loi hypergéométrique, probabilité = 0.00721) et les deux analyses sont par conséquent plus cohérentes entre elles qu'attendu par hasard.

Dans les différents types d'analyses, les sites détectés sous coévolution ne semblent pas se répartir préférentiellement dans les régions hypervariables connues chez *Brassica* (NISHIO and KUSABA 2000). Cette tendance a été testée à l'aide de deux groupes de couples de positions définis selon qu'ils impliquaient ou non des positions dans les régions hypervariables. Le degré de coévolution de ces deux groupes n'a pas été jugé statistiquement différent (Test de Student, p-value = 0.383). Les régions hypervariables définies chez *Brassica* ne semblent par ailleurs pas particulièrement correspondre à des pics de diversité chez *Arabidopsis* (Figure 18).

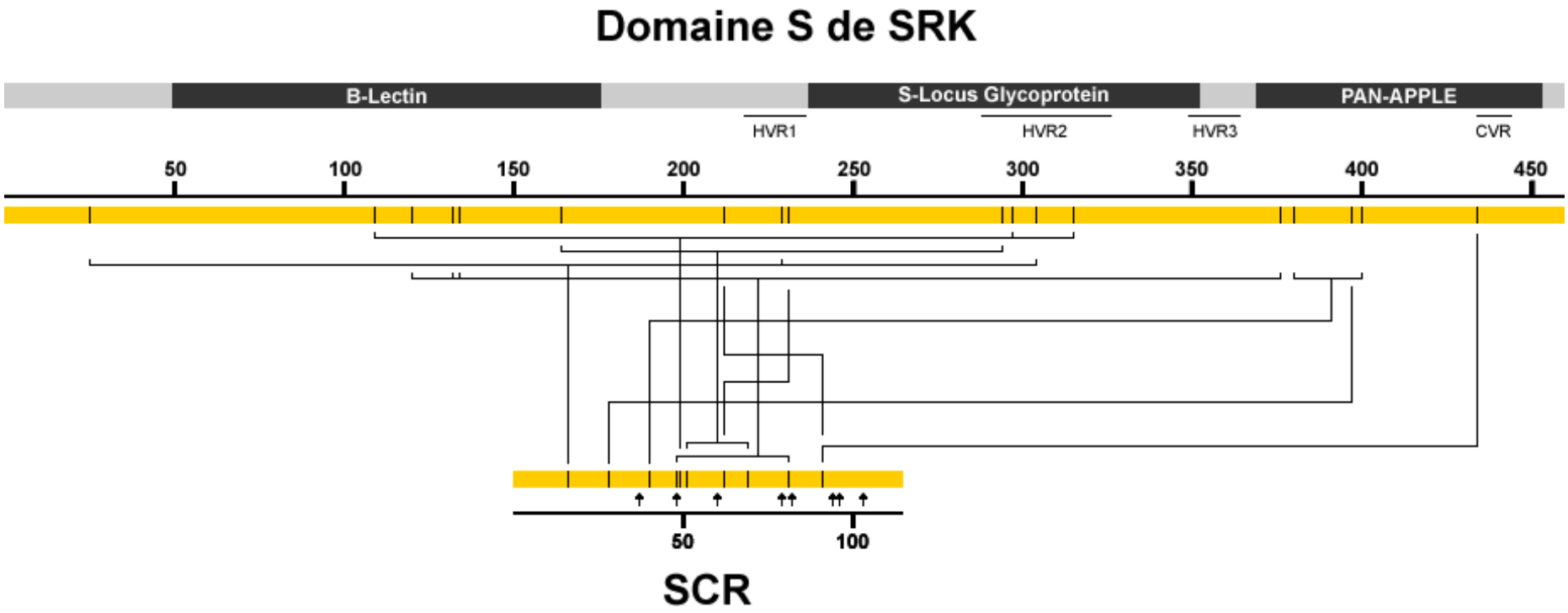


Figure 17.A. Groupes de positions montrant un signal de coévolution par corrélation entre le domaine S de la protéine SRK et la protéine SCR dans le cadre de l'analyse par clusters (DUTHEIL & GALTIER 2007). Une représentation schématique du domaine S de SRK est incluse en haut de la figure, indiquant les différentes régions fonctionnelles (MARCHLER-BAUER *et al.* 2011) et les régions hypervariables connues chez *Brassica* (NISHIO & KUSABA 2000). Les résidus de cystéine caractéristiques de la protéine SCR sont quant à eux indiqués par des flèches. Les positions indiquées se réfèrent à l'alignement des séquences protéiques, et non aux positions absolues sur les séquences.

IV - DISCUSSION

La reconnaissance de type clé-serrure des protéines du pollen et du pistil impose un processus de coévolution. Si les patrons d'évolution du locus d'auto-incompatibilité ont dans la plupart des cas été étudiés sur la base d'un seul des deux gènes impliqués, le gène *SRK* en l'occurrence, l'approche génomique permet ici d'avoir accès au second partenaire de l'interaction et d'étudier directement ce processus.

A - CORRELATION ET COMPENSATION

La coévolution est un rapport réciproque entre deux ou plusieurs entités biologiques, des protéines dans notre cas présent, qui varient simultanément en fonction les unes des autres. Cette relation évolutive est le reflet des contraintes que subissent ces protéines qui interagissent. Dans le cadre du système d'auto-incompatibilité et de la coévolution des protéines du pistil et du pollen, deux processus peuvent être mis en cause. Afin qu'il ne soit pas rompu, le système d'auto-incompatibilité nécessite que les protéines du pistil et du pollen maintiennent leur reconnaissance. La coévolution peut donc tout d'abord être liée à cette contrainte spécifique. On peut imaginer que c'est notamment le cas des sites soumis à compensation : un site mute pour compenser les changements intervenus à une ou plusieurs autres positions qui lui sont liées et ainsi éviter la rupture de l'auto-incompatibilité au sein de la spécificité concernée. Par ailleurs, la coévolution peut également être liée à l'évolution de nouvelles spécificités. On peut alors imaginer que c'est davantage le cas des sites soumis à corrélation : les deux protéines d'une spécificité donnée évoluent conjointement en s'éloignant de leur état ancestral. La nouvelle spécificité engendrée bénéficiera alors de l'avantage du rare conféré par la sélection fréquence-dépendante et sera fortement sélectionnée.

B - COMPARAISON DES SITES DETECTES DANS LE GENRE *ARABIDOPSIS* ET LE GENRE *BRASSICA*

Les genres *Arabidopsis* et *Brassica* partageant le même système d'auto-incompatibilité sporophytique, on peut se demander s'ils tendent à montrer les mêmes sites en situation de coévolution. Pour répondre à cette question, nos résultats peuvent être comparés avec une analyse par clusters réalisée au sein de SRK chez plusieurs espèces du genre *Brassica* (DUTHEIL and GALTIER 2007). Cette analyse a permis de mettre en évidence 14 groupes impliquant un total de 30 sites sous corrélation, et 23 groupes impliquant 40 sites sous compensation. Si les séquences utilisées de *Brassica* sont alignées avec les données présentées ici, les sites détectés dans les deux cas peuvent être confrontés. Sur les 30 sites mis en évidence sous corrélation, trois (les sites 63, 194 et 397) sont retrouvés dans nos données. De la même manière, sur les 40 sites détectés sous compensation, cinq (les sites 31, 63, 93, 129 et 366) sont retrouvés dans nos données. La probabilité de détecter au hasard au moins autant de sites en commun au travers de deux analyses successives étant élevée (loi hypergéométrique, probabilité = 0.452 dans le cas de la corrélation et probabilité = 0.651 dans le cas de la compensation),

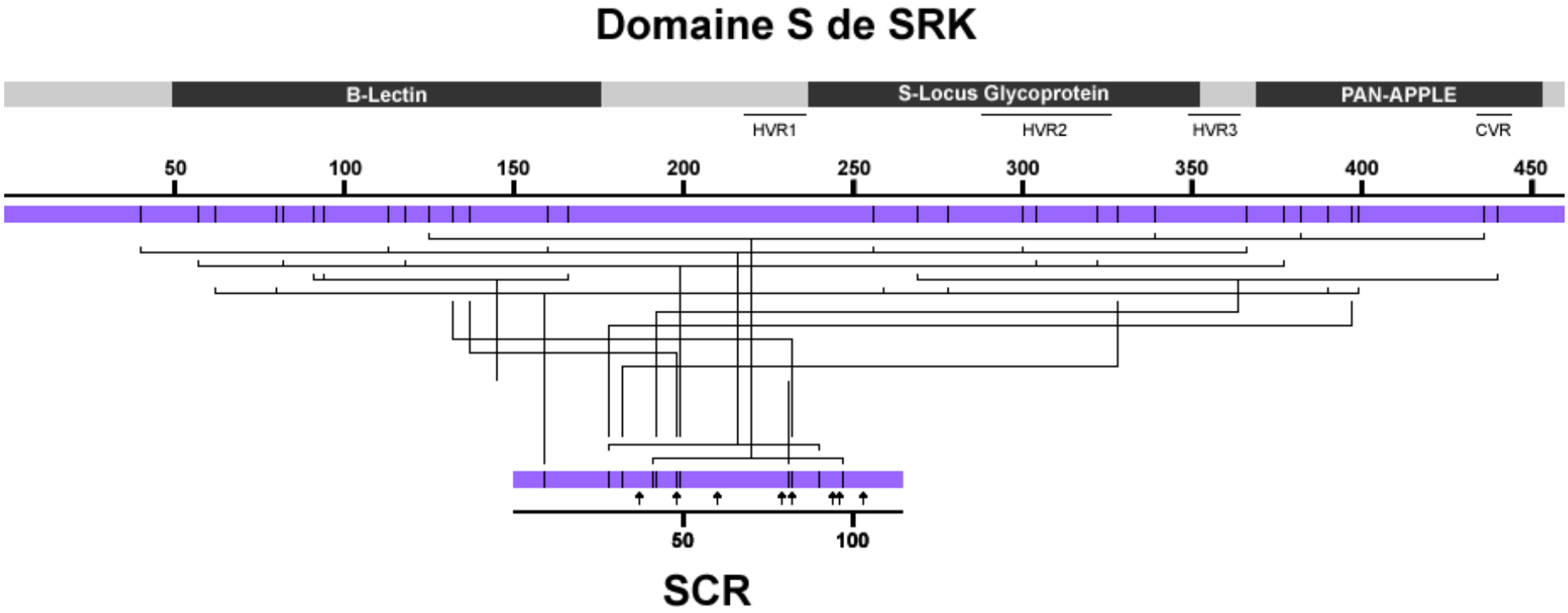


Figure 17.B. Groupes de positions montrant un signal de coévolution par compensation entre le domaine S de la protéine SRK et la protéine SCR dans le cadre de l'analyse par clusters (DUTHEIL & GALTIER 2007). Une représentation schématique du domaine S de SRK est incluse en haut de la figure, indiquant les différentes régions fonctionnelles (MARCHLER-BAUER *et al.* 2011) et les régions hypervariables connues chez *Brassica* (NISHIO & KUSABA 2000). Les résidus de cystéine caractéristiques de la protéine SCR sont quant à eux indiqués par des flèches. Les positions indiquées se réfèrent à l'alignement des séquences protéiques, et non aux positions absolues sur les séquences.

les deux analyses ne sont pas plus cohérentes entre elles que ce qu'il serait attendu par hasard. Deux hypothèses peuvent être avancées pour expliquer ces différences. Tout d'abord, avec 16 séquences étudiées, notre jeu de données est moins important que chez *Brassica* (53 séquences (DUTHEIL and GALTIER 2007)). On pourrait donc s'attendre à ne détecter qu'un sous-ensemble du signal mis en évidence chez *Brassica*. D'autre part, on sait que le genre *Brassica* a connu un fort goulot d'étranglement, à l'issue duquel n'ont été conservés que deux allèles d'auto-incompatibilité (CASTRIC and VEKEMANS 2007; EDH *et al.* 2009a). Les différents allèles présents aujourd'hui chez *Brassica* étant probablement issus d'un processus de re-diversification à partir de ces deux seuls allèles, on peut imaginer que les sites ayant tendance à coévoluer, dans le but de maintenir la reconnaissance ou d'évoluer vers une nouvelle spécificité, ne sont pas les mêmes que dans le genre *Arabidopsis* qui n'a pas subi cette re-diversification. Sous cette hypothèse, les sites impliqués dans l'interaction pollen-pistil seraient distincts dans les différentes lignées alléliques, en accord avec l'observation que les sites sous sélection positive chez SRK ne sont pas identiques entre les différentes espèces (CASTRIC and VEKEMANS 2007). Cette hypothèse explique en outre que les sites détectés dans nos analyses ne se regroupent pas préférentiellement dans les régions hypervariables définies dans le genre *Brassica* (NISHIO and KUSABA 2000) et est supportée par l'analyse de la diversité le long du domaine S de SRK (Figure 18).

C - COEVOLUTION LIEE A L'ABSENCE D'UN PONT DISULFURE

La cystéine est un acide aminé caractérisé par la présence d'un groupement sulfhydryle qui lui permet de former des ponts disulfures. Dans la majeure partie des cas, la protéine SCR compte huit résidus de cystéine conservés (BOGGS *et al.* 2009a; BOGGS *et al.* 2009b; KUSABA *et al.* 2001; TSUCHIMATSU *et al.* 2010). Dans la structure tridimensionnelle de la protéine, ces résidus de cystéine se regroupent par deux, chaque couple étant lié par un pont disulfure (CHOOKAJORN *et al.* 2004). La protéine AlSCR39 se distingue des autres par le fait qu'elle ne présente pas le deuxième et le cinquième résidu de cystéine (voir l'annexe 1.A pour l'alignement des protéines SCR). Ces deux résidus étant impliqués dans l'établissement d'un pont disulfure particulier (voir la Figure 1 de CHOOKAJORN *et al.* (2004)), elle ne possède par conséquent que trois de ces liens. De manière intéressante, les deux sites comprenant ces résidus de cystéine qui sont perdus chez AlSCR39 (les sites 48 et 82) sont détectés comme coévoluant avec divers sites dans toutes les analyses effectuées ici. On peut ainsi supposer que les contraintes physiques liées à l'absence de l'un des ponts disulfures engendrent la nécessité d'une coévolution avec d'autres sites afin de conserver la fonctionnalité de l'interaction des protéines du pollen et du pistil.

D - PERSPECTIVES

L'analyse préliminaire de la coévolution entre les protéines du pistil et du pollen dans le cadre de l'auto-incompatibilité a permis de mettre en évidence l'existence d'un signal de coévolution fort entre le domaine S de SRK et SCR, traduisant les contraintes qui régissent leur interaction. Dans le cadre de

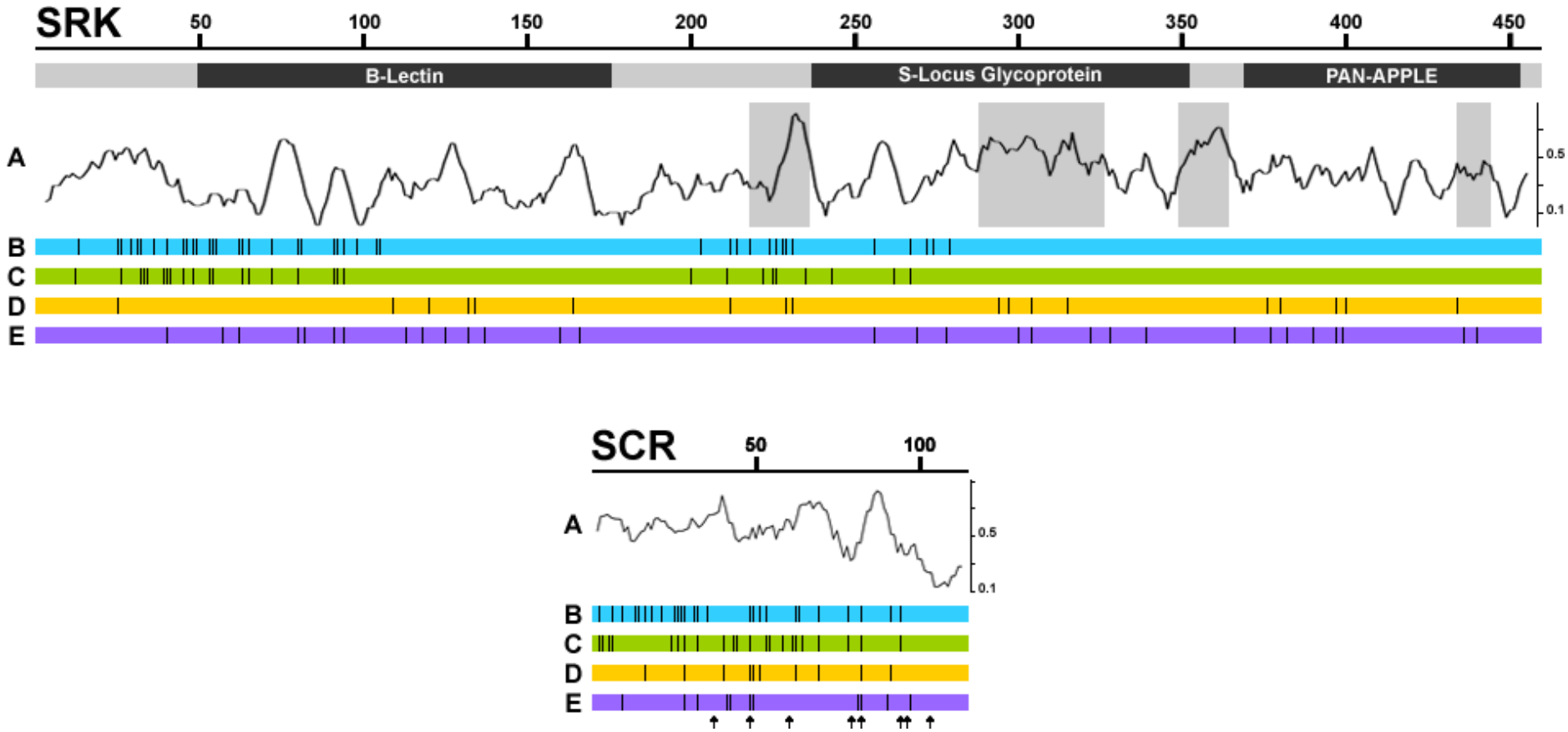


Figure 18. Récapitulatif des sites détectés dans les protéines SRK (uniquement le domaine S) et SCR par les différents types d'analyses.

Cette figure regroupe, pour les deux protéines analysées :

A. Une analyse en fenêtre coulissante de la diversité (π) le long de la protéine, avec une taille de fenêtre de 7 acides aminés et un pas de 1 acide aminé. Les régions hypervariables détectées pour SRK dans le genre *Brassica* (NISHIO & KUSABA 2000) sont représentées par des rectangles grisés et ne semblent pas correspondre à des régions hypervariables chez *Arabidopsis lyrata* et *A. halleri*.
B. Les sites détectés sous corrélation dans le cadre de l'analyse par paires.
C. Les sites détectés sous compensation dans le cadre de l'analyse par paires.
D. Les sites détectés sous corrélation dans le cadre de l'analyse par clusters.
E. Les sites détectés sous compensation dans le cadre de l'analyse par clusters.

l'analyse par paires, la totalité de la protéine SRK a été testée. Les sites détectés sont exclusivement restreints au domaine S de la protéine, c'est-à-dire à la partie impliquée dans la reconnaissance du pistil et du pollen. Cette localisation précise supporte le fait que ces sites sont bien liés à une interaction spécifique des protéines du pollen et du pistil et non seulement à leur histoire partagée. A terme, l'obtention de nouveaux couples de protéines SCR-SRK pourrait permettre d'établir une carte précise des sites qui coévoluent les uns avec les autres. D'autres perspectives pourraient également se révéler intéressantes. Avec un nombre de couples de protéines plus important, il pourrait par exemple être envisagé de réaliser des analyses séparées au sein des quatre classes phylogénétiques, afin de potentiellement y mettre en évidence des groupes de sites spécifiques. Par ailleurs, des approches fonctionnelles pourraient également être couplées à ces analyses prédictives. Des positions candidates pourraient ainsi être testées par introduction de mutations, afin de déterminer si ces modifications suffisent à perturber la reconnaissance du pollen et du pistil.

SYNTHESE

I - APPORTS ET LIMITES DES APPROCHES GENOMIQUES DANS L'ETUDE DU SYSTEME D'AUTO-INCOMPATIBILITE

Parce qu'il est soumis à des contraintes sélectives fortes et originales, le système d'auto-incompatibilité chez les plantes hermaphrodites est un système génétique complexe. Avec douze séquences génomiques comprenant la région du locus S dans le genre *Arabidopsis*, cette thèse propose l'analyse d'un jeu de données important. Ainsi, alors que la génomique se fait de plus en plus présente en biologie, elle illustre les apports et les limites de ce type d'approche dans l'étude d'une telle région génomique.

A - CONSIDERER LA REGION DU LOCUS D'AUTO-INCOMPATIBILITE DANS SON ENSEMBLE

Si la plupart des études décrivant l'auto-incompatibilité sont basées sur des séquences partielles du locus S, le plus souvent provenant du déterminant femelle *SRK* uniquement, l'approche génomique a l'intérêt de considérer la région génomique dans son ensemble. Ce changement d'échelle a permis de prendre en compte plusieurs aspects de cette région génomique qui n'étaient que peu documentés.

Tout d'abord, la comparaison des différents haplotypes a permis de délimiter avec précision le locus d'auto-incompatibilité en lui-même. Nous avons ainsi choisi de le définir comme étant la région génomique située entre les deux points précis qui marquent la rupture de la zone recombinante à la zone non-recombinante. Selon cette définition, le locus d'auto-incompatibilité s'étend d'un point situé après les régions promotrices de la *U-box* jusqu'à un point situé au niveau du codon stop du gène *ARK3*. Une première estimation précise de la taille du locus S a par conséquent pu être établie en évaluant l'étendue de la zone non-recombinante. Cette estimation peut être mise en rapport avec les travaux concernant le fardeau génétique lié au locus d'auto-incompatibilité. En effet, en raison du temps de résidence important des haplotypes d'auto-incompatibilité en populations naturelles et de la restriction locale de la recombinaison, on s'attend à ce que le locus S soit soumis à une accumulation de mutations délétères récessives (UYENOYAMA 1997). Le taux d'homozygotie étant très faible au niveau de cette région génomique, ces mutations délétères n'auraient que très rarement l'occasion d'être exprimées et ne seraient par conséquent pas purgées. Un fardeau génétique particulier serait ainsi lié à chaque haplotype d'auto-incompatibilité. On s'attend par ailleurs à ce qu'il soit d'autant plus important que le niveau de dominance de l'haplotype concerné est élevé, parce que les haplotypes les plus dominants subissent une restriction de la recombinaison plus importante. L'existence de ce fardeau génétique lié au locus d'auto-incompatibilité a par exemple été mise en évidence chez *A. halleri* grâce à des expériences de pollinisation forcée (LLAURENS *et al.* 2009b). Cette

étude a en outre confirmé que ce phénomène était fort dans le cas des haplotypes les plus dominants (par exemple *Ah15*) et indécelable dans le cas de l'haplotype le plus récessif (*Ah01*). En connaissant à présent l'étendue de la région non-recombinante, on peut se demander comment une région génomique si localisée et contenant aussi peu de gènes que le locus d'auto-incompatibilité peut entraîner un tel fardeau génétique.

Dans un second temps, en tenant compte de la région génomique dans son ensemble, les éléments transposables ont pu être annotés et quantifiés. En effet, si leur présence au locus d'auto-incompatibilité avait déjà été mentionnée chez plusieurs espèces (CUI *et al.* 1999; FUJIMOTO *et al.* 2006; GUO *et al.* 2011; TOMITA *et al.* 2004; WHEELER *et al.* 2003), leur abondance par rapport à une région quelconque du génome n'avait encore jamais été analysée. Les données génomiques disponibles chez l'espèce auto-incompatible *A. lyrata* (HU *et al.* 2011) couplées à l'analyse des séquences BAC présentées ici, ont donc permis de souligner l'importance des éléments transposables au locus d'auto-incompatibilité, comme cela avait déjà été mis en évidence dans le cas d'autres régions génomiques impliquées dans le déterminisme du sexe (BACHTROG 2005; HOOD *et al.* 2004; MARAIS *et al.* 2008). De plus, le nombre important d'haplotypes analysés a permis de mettre en évidence des patrons différents selon le niveau de dominance, avec une densité en éléments transposables plus importante dans le cas des haplotypes les plus dominants. L'influence de ces éléments transposables sur la région génomique du locus d'auto-incompatibilité reste toutefois à déterminer. On pourrait en particulier supposer qu'ils sont impliqués dans le fardeau génétique précédemment décrit, mais il est cependant difficile d'imaginer que le coût de leur accumulation soit si fort qu'il en est mesurable empiriquement (LLAURENS *et al.* 2009b).

Enfin, l'obtention des séquences analysées dans cette thèse a permis de comparer le locus d'auto-incompatibilité en lui-même et ses régions flanquantes. Des patrons d'évolution moléculaire contrastés ont ainsi pu être mis en évidence, que ce soit au travers des phylogénies des gènes ou de la diversité structurale (ordre et orientation des gènes) de ces deux régions. Ces résultats illustrent notamment à quel point l'effet des contraintes sélectives agissant sur une région génomique telle que le locus S peut être à la fois important et localisé.

B - ANALYSE DE COUPLES DE SEQUENCES *SCR* ET *SRK*

L'obtention de la séquence complète du locus d'auto-incompatibilité a également permis d'accéder aux séquences du déterminant mâle *SCR* et de confronter ses patrons d'évolution moléculaire à ceux du déterminant femelle *SRK*. En plus de mettre en évidence chez *SCR* la conservation des classes phylogénétiques définies chez *SRK*, ces données constituent la première opportunité d'étudier directement le processus de coévolution liant les deux partenaires de ce système clé-serrure. Peu d'attendus théoriques existent sur les modalités exactes de ce type de processus. En particulier, on ignore si les mutations subies par l'un des deux gènes entraînent le passage par un intermédiaire

partiellement auto-compatible avant que le partenaire ne soit sujet à des mutations venant réactiver l'interaction, que ce soit au sein d'une spécificité donnée ou dans le cadre de l'évolution vers une spécificité nouvelle (CHARLESWORTH 2000). On ignore également le nombre de sites qui sont impliqués dans l'interaction moléculaire entre le pollen et le pistil. On peut en effet imaginer qu'un nombre important de mutations soit nécessaire pour déstabiliser l'interaction (CHOOKAJORN *et al.* 2004). A l'inverse, une mutation unique pourrait suffire à briser la reconnaissance entre les deux protéines, comme cela a par exemple été montré dans le cas d'un système gamétophytique de type S-RNase chez *Solanum chacoense* (MATTON *et al.* 1999). Par ailleurs, la vitesse d'évolution du processus de coévolution est encore inconnue. Dans un cas extrême, si les contraintes mutuelles sont extrêmement strictes sur les partenaires de l'interaction, on pourrait effectivement s'attendre à ce que la coévolution favorise le maintien des combinaisons telles qu'elles sont, tendant à figer les partenaires dans leur évolution. Dans le cas contraire, la coévolution pourrait être envisagée comme une sorte de course évolutive dans laquelle l'un des partenaires mute aussi rapidement que possible pour suivre toutes les variations qui se produisent sur le second. Dans ce contexte, nos données ont permis de mettre en évidence un signal de coévolution important entre les protéines SCR et SRK, reflétant les contraintes spécifiques qui les lient. Ce signal pourra par la suite être affiné grâce à l'obtention de nouvelles données et des analyses fonctionnelles pourront permettre, par introduction de mutations par exemple, de valider les positions mises en cause. Il reste cependant encore difficile d'envisager de quelle manière ce signal de coévolution pourra être mis en relation avec les analyses théoriques prédisant notamment que la dynamique de diversification allélique implique préférentiellement des mutations au niveau du pollen suivies de mutations compensatrices au niveau du pistil (GERVAIS *et al.* 2011; UYENOYAMA *et al.* 2001).

C - UTILISATION DE DONNEES NON EXPLOITEES

A une échelle plus large, le projet 1001 génomes (WEIGEL and MOTT 2009) vise à obtenir la séquence du génome complet d'un millier d'accessions chez *A. thaliana*. Il offre par conséquent la possibilité de décrire la diversité d'un locus d'auto-incompatibilité après que le système ait été rompu. Cependant, certaines régions des génomes ne peuvent pas être assemblées par les méthodes classiques, et les données qui les concernent ne peuvent être exploitées. C'est le cas notamment du locus d'auto-incompatibilité, en raison de sa forte divergence, et son étude nécessite l'utilisation de méthodes génomiques complémentaires. Nous avons donc réalisé de nouveaux assemblages du locus d'auto-incompatibilité, en prenant chacun des haplogroupes principaux comme référence. Les résultats obtenus illustrent l'intérêt de la méthode génomique ciblée qui a été la nôtre. En effet, en l'absence de la séquence de l'haplogroupe *C* décrite dans le chapitre II, les lectures issues du projet 1001 génomes et concernant la région génomique du locus d'auto-incompatibilité seraient restées inexploitées, faute d'avoir les références auxquelles les comparer. La mise à disposition de nouvelles séquences du projet

1001 génomes pourra ainsi permettre par la suite d'affiner le scénario de dégénérescence du locus d'auto-incompatibilité qui a été proposé dans ce même chapitre.

II - PERSPECTIVES

Les résultats et les données présentés dans le cadre de cette thèse ouvrent également de nouvelles pistes de recherche dans l'étude du locus d'auto-incompatibilité, que ce soit par l'analyse des séquences disponibles grâce à des méthodes non utilisées ici ou par l'acquisition de nouvelles données.

A - DETERMINATION DE LA DOMINANCE

Tout d'abord, les phénomènes moléculaires responsables de la dominance des gènes d'auto-incompatibilité restent encore mal connus. Concernant le gène pollen *SCR* dans le genre *Brassica*, il a été montré que dans le cas d'un individu hétérozygote, l'absence d'expression d'un allèle donné en présence d'un second plus dominant était induite par la méthylation de son promoteur (SHIBA *et al.* 2006). La méthylation de ce même allèle ne s'observe pas lorsqu'il est en présence d'un allèle plus récessif, ou chez un individu homozygote. Un allèle particulier ayant toujours le même niveau de dominance, on peut avancer que l'élément provoquant la méthylation du promoteur est lié au locus S. Néanmoins, cet élément n'est pas l'allèle dominant lui-même. En effet, chez un individu hétérozygote, la suppression de la séquence codante de l'allèle dominant ne rétablit pas l'expression de l'allèle récessif (FUJIMOTO and NISHIO 2007).

Les petits ARN sont de courtes molécules d'ARN d'environ 20 à 30 nucléotides. Très abondants chez les plantes (VOINET 2009), ils sont impliqués dans de nombreux mécanismes de contrôle de l'expression des gènes, soit transcriptionnels soit post-transcriptionnels selon qu'ils interagissent directement avec le gène concerné ou son ARN_m (CARTHEW and SONTHEIMER 2009). Récemment, Tarutani *et al.* (2010) ont montré que, chez *B. rapa*, la méthylation des allèles appartenant à la classe récessive était contrôlée par un petit ARN non-codant, nommé S_{mi}. Exprimé dans les anthères, cet élément se trouve sur les haplotypes dominants à une distance de 15 à 25 kb du gène *SCR*. La question de la généralisation de ce mécanisme de contrôle à d'autres espèces, et en particulier aux espèces du genre *Arabidopsis*, pose des difficultés intéressantes. En effet, les haplotypes d'auto-incompatibilité dans le genre *Arabidopsis* se groupent en au moins quatre classes de dominance (PRIGODA *et al.* 2005), alors que deux seulement coexistent dans le genre *Brassica* (NASRALLAH *et al.* 1991) et on peut par conséquent s'attendre à des patrons plus complexes, nécessitant potentiellement la coexistence de plusieurs petits ARN non codants. Dans ce contexte, les séquences génomiques du locus d'auto-incompatibilité chez *Arabidopsis* analysées dans le cadre de ce travail peuvent permettre d'identifier des petits ARN non codants ciblant les régions promotrices du gène *SCR*, à l'aide de programmes

informatiques prédictifs (ALEXIOU *et al.* 2009). L'expression de petits ARN candidats pourrait dans un second temps être vérifiée par des expériences de séquençage massif des petits ARN.

De plus, si le rôle des petits ARN dans le contrôle de la dominance au sein du genre *Arabidopsis* était confirmé, la question de leur origine pourrait alors être posée. Ces éléments pourraient en effet être engendrés à partir de structures tige-boucle, elles-mêmes formées par le biais des séquences répétées des éléments transposables (PIRIYAPONGSA and JORDAN 2007), et recrutés dans un second temps dans le cadre du contrôle de la dominance. Puisque nous avons montré que l'accumulation d'éléments transposables était plus importante pour les haplotypes dominants, on peut supposer que ces derniers sont plus susceptibles de recruter de nouveaux petits ARN. Ainsi, une fois acquis un certain niveau de dominance, ils pourraient être entraînés dans un cycle qui les pousserait à devenir de plus en plus dominants.

B - DIVERSITE INTRA-HAPLOTYPIQUE AU SEIN ET ENTRE ESPECES

Les données de cette thèse comportent une douzaine d'haplotypes d'auto-incompatibilité répartis dans les quatre classes de dominance connues dans le genre *Arabidopsis*. L'analyse de nouvelles séquences pourrait permettre des comparaisons plus fines, par exemple à travers la comparaison de couples trans-spécifiques ou de plusieurs séquences montrant la même spécificité.

1 - Couples trans-spécifiques

Les haplotypes du locus d'auto-incompatibilité se maintenant sur de longues périodes de temps (VEKEMANS and SLATKIN 1994), les espèces proches telles que *A. halleri*, *A. lyrata* et *A. thaliana* présentent du polymorphisme trans-spécifique et partagent une partie de leurs spécificités. Cependant, si l'on sait que les couples trans-spécifiques sont issus de l'évolution d'un même haplotype ancestral et possèdent des séquences *SRK* fortement similaires (CASTRIC *et al.* 2008), on ne connait rien de la diversité de leur région génomique. Cette diversité d'un même haplotype chez plusieurs espèces pourrait en particulier être évaluée en comparant leur taille, l'orientation relative de leurs gènes mais aussi leur contenu en éléments transposables. De plus, deux types de couples trans-spécifiques pourraient être analysés. En effet, des comparaisons pourraient tout d'abord être effectuées entre des haplotypes fonctionnels chez *A. halleri* et *A. lyrata*. Par exemple, l'haplotype récessif *Al01*, présent dans nos données, pourrait être comparé à l'haplotype *Ah01*. De la même manière dans les autres classes phylogénétiques, les haplotypes *Al14, Al13* et *Al39*, que nous avons analysés dans le cadre de cette thèse pourraient être comparés à leurs haplotypes trans-spécifiques respectifs, à savoir *Ah09, Ah29* et *Ah18*. D'autre part, des comparaisons pourraient être entreprises entre les haplotypes non-fonctionnels chez *A. thaliana* et leurs haplotypes trans-spécifiques fonctionnels chez *A. halleri* et *A. lyrata*, afin de comprendre les patrons de dégénérescence d'un locus d'auto-incompatibilité à la suite de la perte du système. Ainsi, l'haplotype *Al16* ayant été récemment

séquencé (GUO *et al.* 2011), on sait que l'orientation de ses gènes est identique à celle de son haplotype trans-spécifique non-fonctionnel chez *A. thaliana* (haplogroupe *B* (TANG *et al.* 2007)) mais que *Al16* est largement plus étendu (GUO *et al.* 2011). Il serait donc intéressant à présent de comparer précisément leurs contenus respectifs en éléments transposables. De la même manière, les séquences des haplogroupes *A* (Col-0 (THE *ARABIDOPSIS* GENOME INITIATIVE 2000)) et *C* (Ita-0, voir le Chapitre II) chez *A. thaliana* pourraient être comparées à leurs haplotypes trans-spécifiques : *Ah04* et *Al37* pour le premier, et *Al36* pour le second.

2 - Evaluation de la diversité intra-haplotype au sein d'une espèce

A une échelle plus fine encore, la diversité du locus d'auto-incompatibilité pourrait également être évaluée au sein d'une même spécificité chez une espèce donnée. On ignore en effet à quel point la région génomique encadrant les gènes *SCR* et *SRK* est divergente dans une même lignée, et deux cas extrêmes peuvent être envisagés. Deux séquences portant la même spécificité peuvent présenter une séquence identique entre les deux points que nous avons définis comme les limites du locus d'auto-incompatibilité. A l'inverse, ces deux séquences peuvent ne partager que les éléments précis qui confèrent la spécificité, à savoir les gènes *SCR* et *SRK,* et potentiellement les petits ARN non codants s'ils sont effectivement impliqués dans le contrôle de la dominance. Dans ce contexte, il pourrait être intéressant d'obtenir des séquences portant la même spécificité mais provenant de régions géographiques différentes. De la même manière que pour les couples trans-spécifiques, leur diversité pourrait être analysée par leur taille, l'organisation de leurs gènes ou leur abondance en éléments transposables. De plus, les patrons de diversité observés pourraient être confrontés entre les lignées d'haplotypes récessifs, qui ont la possibilité de recombiner et d'atteindre des fréquences relativement fortes, et les lignées d'haplotypes dominants, qui ne peuvent au contraire pas recombiner et se maintiennent à des fréquences très faibles (BILLIARD *et al.* 2007; SCHIERUP *et al.* 1997).

III - CONCLUSION GENERALE

Le système d'auto-incompatibilité chez les plantes hermaphrodites a longtemps été étudié au travers d'approches ciblées sur des fragments d'ADN de petite taille (quelques centaines de paires de bases). Depuis quelques années, les méthodes génomiques se font de plus en plus présentes dans le domaine de la biologie évolutive. Cependant, les nouvelles techniques de séquençage ne permettent pas encore d'accéder directement à des régions fortement polymorphes telles que le locus S. Dans ce contexte, cette thèse se place à une échelle intermédiaire entre les approches de type PCR et le séquençage de génomes complets. Elle allie pour cela des données issues du séquençage ciblé de la région génomique d'intérêt et les données génomiques disponibles chez les espèces étudiées. L'utilisation de ce type d'approche est pertinente dans l'étude d'une région complexe telle que le locus d'auto-incompatibilité

et a permis d'approfondir différents aspects présentés dans ce document : la comparaison entre la région en elle-même et ses régions flanquantes, la diversité structurale du locus S, la densité en éléments transposables, la dégénérescence de la région génomique après rupture du système ou encore les patrons de coévolution des déterminants mâle et femelle. A moyen terme, on peut supposer que les progrès incessants des nouvelles technologies de séquençage à haut débit et d'assemblage permettront d'accéder aux régions fortement polymorphes par le séquençage de génomes complets. Les données engendrées pourraient ainsi permettre de caractériser au niveau moléculaire d'autres types de systèmes d'auto-incompatibilité. Malgré les perspectives que ces données peuvent offrir, des analyses fonctionnelles resteront toutefois nécessaires afin de mettre en évidence les déterminants mâle et femelle et de saisir toute la complexité de leurs interactions.

RÉFÉRENCES BIBLIOGRAPHIQUES

ALEXIOU, P., M. MARAGKAKIS, G. L. PAPADOPOULOS, M. RECZKO and A. G. HATZIGEORGIOU, 2009 Lost in translation: an assessment and perspective for computational microARN target identification. Bioinformatics **25:** 3049-3055.

ALMEIDA, R., and R. C. ALLSHIRE, 2005 RNA silencing and genome regulation. Trends in Cell Biology **15:** 251-258.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. Journal of Molecular Biology **215:** 403-410.

BACHTROG, D., 2005 Sex chromosome evolution: Molecular aspects of Y-chromosome degeneration in *Drosophila*. Genome Research **15:** 1393-1401.

BACHTROG, D., M. KIRKPATRICK, J. E. MANK, S. F. MCDANIEL, J. C. PIRES *et al.*, 2011 Are all sex chromosomes created equal? TRENDS in Genetics **in press**.

BAKER, H. G., 1967 Support for Baker's law. Evolution **21:** 853-856.

BAKKEREN, G., G. JIANG, R. L. WARREN, Y. BUTTERFIELD, H. SHIN *et al.*, 2006 Mating factor linkage and genome evolution in basidiomycetes pathogens of cereals. Fungal Genetics and Biology **4:** 655-666.

BARRETT, S. C. H., 2002 The evolution of plant sexual diversity. Nature Review Genetics **3:** 274-284.

BARRIO, I. C., C. G. BUENO, P. B. BANKS and F. S. TORTOSA, 2010 Prey naiveté in an introduced prey species: The wild rabbit in Australia. Behavioral Ecology **21:** 986-991.

BARTON, N. H., and B. CHARLESWORTH, 1998 Why sex and recombination? Science **281:** 1986-1990.

BECHSGAARD, J. S., V. CASTRIC, D. CHARLESWORTH, X. VEKEMANS and M. H. SCHIERUP, 2006 The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 Myr. Molecular Biology and Evolution **23:** 1741-1750.

BENNETT, M. D., I. J. LEITCH, H. J. PRICE and J. S. JOHNSTON, 2003 Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~ 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25 % larger than the *Arabidopsis* Genome Initiative estimate of ~125 MB. Annals of Botany **91:** 547-557.

BENNETZEN, J. L., 2000 Transposable elements contribution to plant genome and genome evolution. Plant Molecular Biology **42:** 251-269.

BERGERO, R., and D. CHARLESWORTH, 2009 The evolution of restricted recombination in sex chromosomes. Trends in Ecology and Evolution **24:** 94-102.

BERGERO, R., D. CHARLESWORTH, D. A. FILATOV and R. C. MOORE, 2008 Defining regions and rearrangements of the *Silene latifolia* Y chromosome. Genetics **178:** 2045-2053.

BERT, V., 2000 Tolérance aux métaux lourds et accumulation chez *Arabidopsis halleri* (Brassicaceae), pp. Université des Sciences et Technologies de Lille 1.

BILLIARD, S., V. CASTRIC and X. VEKEMANS, 2007 A general model to explore complex dominance patterns in plant sporophytic self-incompatibility systems. Genetics **175:** 1351-1369.

BILLIARD, S., M. LOPEZ-VILLAVIVENCIO, B. DEVIER, M. E. HOOD, C. FAIRHEAD *et al.*, 2011 Having sex, yes, but with whom? Inferences from fungi on the evolution of anisogamy and mating types. Biological reviews **86:** 421-442.

BOGGS, N. A., K. G. DWYER, P. SHAH, A. A. MCCULLOCH, J. BECHSGAARD *et al.*, 2009a Expression of distinct self-incompatibility specificities in *Arabidopsis thaliana*. Genetics **182:** 1313-1321.

BOGGS, N. A., J. B. NASRALLAH and M. E. NASRALLAH, 2009b Independent S-locus mutations caused self-fertility in *Arabidopsis thaliana*. PLoS Genetics **5:** e1000426.

BOMBLIES, K., and D. WEIGEL, 2010 *Arabidopsis* and relatives as models for the study of genetic and genomic incompatibilities. Philosophical Transactions of the Royal Society B **365:** 1815-1823.

BOYES, D. C., and J. B. NASRALLAH, 1993 Physical linkage of the SLG and SRK genes at the self-incompatibility locus of *Brassica oleracea*. Molecular and General Genetics **236:** 369-373.

BRUDNO, M., A. POLIAKOV, C. B. DO, I. DUBCHAK and S. BATZOGLOU, 2003 Glocal alignment: finding rearrangements during alignment. Bioinformatics **19S1:** i54-i62.

BURGE, C., and S. KARLIN, 1997 Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology **268:** 78-94.

BUSCH, J. W., and D. J. SCHOEN, 2008 The evolution of self-incompatibility when mates are limiting. TRENDS in Plant Science **13:** 128-136.

CAMPBELL, J. M., and M. J. LAWRENCE, 1981 The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. II. The number and frequency of S-alleles in a natural population (R106). Heredity **46:** 81-90.

CARTHEW, R. W., and E. J. SONTHEIMER, 2009 Origins and Mechanisms of miRNAs and si RNAs. Cell **136:** 642-655.

CASSELMAN, A. L., J. VREBALOV, J. A. CONNER, A. SINGHAL, J. GIOVANNONI *et al.*, 2000 Determining the physical limits of the Brassica S locus by recombinational analysis. Plant Cell **12:** 23-34.

CASTRIC, V., J. BECHSGAARD, S. GRENIER, R. NOUREDDINE, M. H. SCHIERUP *et al.*, 2010 Molecular evolution within and between self-incompatibility specificities. Molecular Biology and Evolution **27:** 11-20.

CASTRIC, V., J. BECHSGAARD, M. H. SCHIERUP and X. VEKEMANS, 2008 Repeated Adaptive Introgression at a Gene under Multiallelic Balancing Selection. PLoS Genetics **4:** e1000168.

CASTRIC, V., and X. VEKEMANS, 2004 Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Molecular Ecology **13:** 2873-2889.

CASTRIC, V., and X. VEKEMANS, 2007 Evolution under strong balancing selection: how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae? BMC Evolutionary Biology **7:** 132-146.

CHALHOUB, B., H. BELCRAM and M. CABOCHE, 2004 Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. Plant Biotechnology Journal **2:** 181-188.

CHARLESWORTH, D., 2000 How Can Two-Gene Models of Self-Incompatibility Generate New Specificities? The Plant Cell **12:** 309-310.

CHARLESWORTH, D., 2006 Evolution of plant breeding systems. Current Biology **16:** R726-735.

CHARLESWORTH, D., and P. AWADALLA, 1998 The molecular population genetics of flowering plant self-incompatibility polymorphisms. Heredity **81:** 1-9.

CHARLESWORTH, D., C. BARTOLOMÉ, M. H. SCHIERUP and B. K. MABLE, 2003 Haplotype Structure of the Stigmatic Self-Incompatibility Gene in Natural Populations of *Arabidopsis lyrata*. Molecular Biology and Evolution **20:** 1741-1753.

CHARLESWORTH, D., B. CHARLESWORTH and G. MARAIS, 2005 Steps in the evolution of heteromorphic sex chromosomes. Heredity **95:** 118-128.

CHARLESWORTH, D., and X. VEKEMANS, 2005 How and when did *Arabidopsis thaliana* become highly self-fertilising. Bioessays **27:** 472-476.

CHOOKAJORN, T., A. KACHROO, D. R. RIPOLL, A. G. CLARK and J. B. NASRALLAH, 2004 Specificity determinants and diversification of the *Brassica* self-incompatibility pollen ligand. PNAS **101:** 911-917.

CLAUSS, M. J., and M. A. KOCH, 2006 Poorly known relatives of *Arabidopsis thaliana*. TRENDS in Plant Science **11:** 449-459.

CUI, Y., N. BRUGIÈRE, L. JACKMAN, Y. BI and S. J. ROTHSTEIN, 1999 Structural and transcriptional comparative analysis of the S locus regions in two self-incompatible *Brassica napus* lines. The Plant Cell **11:** 2217-2231.

DE NETTANCOURT, D., 2001 *Incompatibility and incongruity in wild and cultivated plants*. Springer-Verlag, Berlin.

DE VIENNE, D. M., T. GIRAUD and O. C. MARTIN, 2007 A congruence index for testing topological similarity between trees. Bioinformatics **23:** 3119-3124.

DUTHEIL, J., and N. GALTIER, 2007 Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evolutionary Biology **7:** 242.

DUTHEIL, J., T. PUPKO, J. M. ALAIN and N. GALTIER, 2005 A Model-Based Approach for Detecting Coevolving Positions in a Molecule. Molecular Biology and Evolution **22:** 1919-1928.

DWYER, K. G., M. A. BALENT, J. B. NASRALLAH and M. E. NASRALLAH, 1991 DNA sequences of self-incompatibility genes from *Brassica campestris* and *B. oleracea*: polymorphism predating speciation. Plant Molecular Biology **16:** 481-486.

EDGAR, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research **32:** 1792-1797.

EDH, K., B. WIDEN and A. CEPLITIS, 2009a The evolution and diversification of S-locus haplotypes in the Brassicaceae family. Genetics **181:** 977-984.

EDH, K., B. WIDÉN and A. CEPLITIS, 2009b Molecular population genetics of the SRK and SCR self-incompatibility genes in the wild plant species *Brassica cretica* (Brassicaceae). Genetics **181:** 985-995.

EMERSON, S., 1939 A preliminary survey of the *Oenothera organensis* population. Genetics **24:** 524-537.

ENTANI, T., M. IWANO, H. SHIBA, F. S. CHE, A. ISOGAI *et al.*, 2003 Comparative analysis of the self-incompatibility (S-) locus region of *Prunus mume*: identification of a pollen-expressed F-box gene with allelic diversity. Genes Cells **8:** 203-213.

FERRIS, P., B. J. S. C. OLSON, P. L. DE HOFF, S. DOUGLASS, D. CASERO DIAZ-CANO *et al.*, 2010 Evolution of an expanded sex determining locus in *Volvox*. Science **328:** 351-354.

FERRIS, P. J., E. V. ARMBRUST and U. W. GOODENOUGH, 2002 Genetic structure of the mating-type locus of *Chlamydomonas reinhardtii*. Genetics **160:** 181-200.

FISHER, R. A., 1941 Average excess and average effect of a gene substitution. Annals of Eugenics **11:** 53-63.

FOBIS-LOISY, I., C. MIEGE and T. GAUDE, 2004 Molecular evolution of the S locus controlling mating in the Brassicaceae. Plant Biology **6:** 109-118.

FRANKLIN-TONG, N. V. E., and F. C. H. FRANKLIN, 2003 Gametophytic self-incompatibility inhibits pollen tube growth using different mechanisms. TRENDS in Plant Science **8:** 598-605.

FUJIMOTO, R., and T. NISHIO, 2007 Self-Incompatibility. Advances in Botanical Research **45:** 139-154.

FUJIMOTO, R., K. OKAZAKI, E. FUKAI, M. KUSABA and T. NISHIO, 2006 Comparison of the genome structure of the self-incompatibility (S) locus in interspecific pairs of S haplotypes. Genetics **173:** 1157-1167.

FUKAI, E., R. FUJIMOTO and T. NISHIO, 2003 Genomic organization of the S core region and the S flanking regions of a class-II S haplotype in *Brassica rapa*. Molecular Genetics and Genomics **269:** 361-369.

GERVAIS, C. E., V. CASTRIC, A. RESSAYRE and S. BILLIARD, 2011 Origin and diversification dynamics of self-incompatibility haplotypes. Genetics **188:** 625-636.

GISH, W., and D. J. STATES, 1993 Identification of protein coding regions by database similarity search. Nature Genetics **3:** 266-272.

GLÉMIN, S., 2007 Mating systems and the efficacy of selection at the molecular level. Genetics **177:** 905-916.

GLÉMIN, S., E. BAZIN and D. CHARLESWORTH, 2006 Impact of mating system on patterns of sequence polymorphism in flowering plants. Proceedings of the Royal Society B **273:** 3011-3019.

GLÉMIN, S., T. GAUDE, M. L. GUILLEMIN, M. LOURMAS, I. OLIVIERI *et al.*, 2005 Balancing Selection in the Wild: Testing Population Genetics Theory of Self-Incompatibility in the Rare Species *Brassica insularis*. Genetics **171:** 279–289.

GOLDBERG, E. E., J. R. KOHN, R. LANDE, K. A. ROBERTSON, S. A. SMITH *et al.*, 2010 Species selection maintains self-incompatibility. Science **22:** 493-495.

GONTHIER, L., A. BELLEC, C. BLASSIAU, E. PRAT, N. HELMSTETTER *et al.*, 2010 Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory (*Cichorium intybus* L., Asteraceae). BMC Research Notes **3:** 225.

GOODENOUGH, U. W., E. V. ARMBRUST, A. M. CAMPBELL and P. J. FERRIS, 1995 Molecular genetics of sexuality in *Chlamydomonas*. Annual Review of Plant Physiology and Plant Molecular Biology **46:** 21-44.

GOODWILLIE, C., 1999 Multiple Origins of Self-Compatibility in *Linanthus* Section *leptosiphon* (Polemoniaceae):Phylogenetic Evidence from Internal-Transcribed-Spacer Sequence Data. Evolution **53:** 1387-1395.

GOODWILLIE, C., S. KALISZ and C. G. ECKERT, 2005 The evolutionary enigma of mixed mating systems in plants: occurence, theoretical explanations, and empirical evidence. Annual Review of Ecology, Evolution and Systematics **36:** 47-79.

GOTOH, O., 1982 An improved algorithm for matching biological sequences. Journal of Molecular Biology **162:** 705-708.

GOTOH, O., 2008 Direct mapping and alignment of protein sequences onto genomic sequence. Bioinformatics **24:** 2438-2444.

GRANTHAM, R., 1974 Amino acid difference formula to help explain protein evolution. Science **185:** 862-864.

GUO, Y., X. ZHAO, C. LANZ and D. WEIGEL, 2011 Evolution of the S-locus region in *Arabidopsis thaliana* relatives. Plant Physiology.

HAFNER, M. S., and S. A. NADLER, 1988 Phylogenetic trees support the coevolution of parasites and their hosts. Nature **332:** 258-259.

HAGENBLAD, J., J. BECHSGAARD and D. CHARLESWORTH, 2006 Linkage Disequilibrium Between Incompatibility Locus Region Genes in the Plant *Arabidopsis lyrata*. Genetics **173:** 1057-1073.

HALL, C., J. WELCH, D. J. KOWBEL and N. L. GLASS, 2010 Evolution and diversity of a fungal self/nonself recognition locus. PLoS ONE **5:** e14055.

HASSELMANN, M., and M. BEYE, 2006 Pronounced differences of recombination activity at the sex determination locus of the Honeybee, a locus under strong balancing selection. Genetics **174:** 1469-1480.

HATAKEYAMA, K., M. WATANABE, T. TAKASAKI, K. OJIMA and K. HINATA, 1998 Dominance relationships between S-alleles in self-incompatible *Brassica campestris* L. Heredity **80:** 241-247.

HELLBORG, L., and H. ELLEGREN, 2004 Low levels of nucleotide diversity in mammalian Y chromosomes. Molecular Biology and Evolution **21:** 158-153.

HISCOCK, S. J., and S. M. MCINNIS, 2003 Pollen recognition and rejection during the sporophytic self-incompatibility response: Brassica and beyond. TRENDS in Plant Science **8:** 606-613.

HOLLISTER, J. D., and B. S. GAUT, 2009 Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Research **19:** 1419-1428.

HOOD, M. E., J. ANTONOVICS and B. KOSKELLA, 2004 Shared forces of sex chromosome evolution in haploid-mating and diploid-mating organisms. Genetics **168:** 141-146.

HU, T. T., P. PATTYN, E. G. BAKKER, J. CAO, J. CHENG *et al.*, 2011 The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nature Genetics **43:** 476-481.

HUA, Z., A. FIELDS and T. KAO, 2008 Biochemical models for S-RNase-based self-incompatibility. Molecular Plant **1:** 575-585.

IGIC, B., and R. LANDE, 2008 Loss of self incompatibility and its evolutionary consequences. International Journal of Plant Sciences **169:** 93-104.

IOERGER, T. R., A. G. CLARK and T.-H. KAO, 1990 Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. PNAS **87:** 9732-9735.

JOHNSON, S. D., and B. ANDERSON, 2010 Coevolution between food-rewarding flowers and their pollinators. Evolution: Education and Outreach **3:** 32-39.

JURKA, J., V. V. KAPITONOV, A. PAVLICEK, P. KLONOWSKI, O. KOHANY *et al.*, 2005 Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research **110:** 462-467.

KAPITONOV, V. V., and J. JURKA, 2008 A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Review Genetics **9:** 411-412.

KAWABE, A., B. HANNSON, A. FORREST, J. HAGENBLAD and D. CHARLESWORTH, 2006 Comparative gene mapping in *Arabidopsis lyrata* chromosomes 6 and 7 and *A. thaliana* chromosome IV: evolutionary history, rearrangements and local recombination rates. Genetical Research Cambridge **88:** 45-46.

KEMP, B. K., and J. DOUGHTY, 2007 S cysteine-rich (SCR) binding domain analysis of the *Brassica* self-incompatibility S-locus receptor kinase. New Phytologist **175:** 619-629.

KIDWELL, M. G., 2002 Transposable elements and the evolution of genome size in eukaryotes. Genetica **115:** 49-63.

KIM, J. M., S. VANGURI, J. D. BOEKE, A. GABRIEL and D. F. VOYTAS, 1998 Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Research **8:** 464-478.

KIMURA, R., K. SATO, R. FUJIMOTO and T. NISHIO, 2002 Recognition specificity of self-incompatibility maintained after the divergence of *Brassica oleraceae* and *Brassica rapa*. The Plant Journal **29:** 215-223.

KOCH, M., J. BISHOP and T. MITCHELL-OLDS, 1999 Molecular systematics of Arabidopsis and Arabis. Plant Biology **1:** 529-537.

KOCH, M. A., and M. MATSCHINGER, 2007 Evolution and genetic differentiation among relatives of Arabidopsis thaliana. PNAS **104:** 6272-6277.

KOHANY, O., A. J. GENTLES, L. HANKUS and J. JURKA, 2006 Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics **25:** 474.

KRÄMER, U., 2010 Metal hyperaccumulation in Plants. Annual review of Plant Biology **61:** 517-534.

KUBO, K., T. ENTANI, A. TAKARA, N. WANG, A. FIELDS *et al.*, 2010 Collaborative non-self recognition system in S-RNase-Based self-incompatibility. Science **330:** 796.

KUSABA, M., K. DWYER, J. HENDERSHOT, J. VREBALOV, J. B. NASRALLAH *et al.*, 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., C. TUNG, M. E. NASRALLAH and J. B. NASRALLAH, 2002 Monoallelic expression and dominance interactions in anthers of self-incompatible *Arabidopsis lyrata*. Plant Physiology **128:** 17-20.

LAHN, B. T., and D. C. PAGE, 1999 Four evolutionary strata on the human X chromosome. Science **286:** 964-967.

LAI, Z., W. MA, B. HAN, L. LIANG, Y. ZHANG *et al.*, 2002 An F-box gene linked to the self-incompatibility (S) locus of *Antirrhinum* is expressed specifically in pollen and tapetum. Plant Molecular Biology **50:** 29-42.

LANDE, R., and D. W. SCHEMSKE, 1985 The Evolution of Self-Fertilization and Inbreeding Depression in Plants. I. Genetic Models. Evolution **39:** 24-40.

LAWSON HANDLEY, L. J., H. CEPLITIS and H. ELLEGREN, 2004 Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. Genetics **167:** 367-376.

LEE, N., G. BAKKEREN, K. WONG, J. E. SHERWOOD and J. W. KRONSTAD, 1999 The mating-type and pathogenicity locus of the fungus *Ustilago hordei* spans a 500-kb region. Genetics **96:** 15026-15031.

LEMAITRE, C., M. D. V. BRAGA, C. GAUTIER, M. F. SAGOT, E. TANNIER *et al.*, 2009 Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. Genome Biology and Evolution **1:** 56-66.

LENGELER, K. B., D. S. FOX, J. A. FRASER, A. ALLEN, K. FORRESTER *et al.*, 2002 Mating-type locus of *Cryptococcus neoformans*: a step in the evolution of sex chromosomes. Eukaryotic Cell **1:** 704-718.

LI, H., and R. DURBIN, 2009 Fast and accurate short reads alignment with Burrows-Wheeler Transform. Bioinformatics **25:** 1754-1760.

LIM, J. K., and M. J. SIMMONS, 1994 Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. Bioessays **16:** 269-275.

LIU, P., S. SHERMAN-BOYLES, M. E. NASRALLAH and J. B. NASRALLAH, 2007 A Cryptic Modifier Causing Transient Self-Incompatibility in *Arabidopsis thaliana*. Current Biology **17:** 734-740.

LLAURENS, V., S. BILLIARD, V. CASTRIC and X. VEKEMANS, 2009a Evolution of dominance in sporophytic self-incompatibility systems: I. genetic load and coevolution of levels of dominance in pollen and pistil. Evolution **63:** 2427-2437.

LLAURENS, V., S. BILLIARD, J.-B. LEDUCQ, V. CASTRIC, E. K. KLEIN *et al.*, 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., L. GONTHIER and S. BILLIARD, 2009b 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.

LYNCH, M., and J. S. CONERY, 2003 The origins of genome complexity. Science **302:** 1401-1404.

MABLE, B. K., M. H. SCHIERUP and D. CHARLESWORTH, 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:** 422-431.

MARAIS, G. A. B., M. NICOLAS, R. BERGERO, P. CHAMBRIER, E. KEJNOVSKY *et al.*, 2008 Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. Current Biology **18:** 545-549.

MARCHLER-BAUER, A., S. LU, J. B. ANDERSON, F. CHITSAZ, M. K. DERBYSHIRE *et al.*, 2011 CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research **39:** 225-229.

MATTON, D. P., D. T. LUU, Q. XIKE, G. LAUBLIN, M. O'BRIEN *et al.*, 1999 Production of an S RNase with dual specificity suggests a novel hypothesis for the generation of new S alleles. Plant Cell **11:** 2087-2098.

MAYOR, C., M. BRUDNO, J. R. SCHWARTZ, A. POLIAKOV, E. M. RUBIN *et al.*, 2000 VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics **16:** 1046-1047.

MCCLURE, B. A., and V. FRANKLIN-TONG, 2006 Gametophytic self-incompatibility understanding the cellular mechanisms involved in "self" pollen tube inhibition. Planta **224**.

MENKIS, A., D. J. JACOBSON, T. GUSTAFSSON and H. JOHANNESSON, 2008 The mating-type chromosome in the filamentous ascomycete Neurospora tetrasperma represents a model for early evolution of sex chromosomes. PLoS Genetics **4:** e1000030.

MOTT, R., 1997 EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Computer Applications in Biosciences **13:** 477-478.

MOYA, A., J. PERETO, R. GIL and A. LATORRE, 2008 Learning how to live together: genomic insights into procaryote-animal symbioses. Nature Review Genetics **8:** 218-229.

NAITHANI, S., T. CHOOKAJORN, D. R. RIPOLL and J. B. NASRALLAH, 2007 Structural modules for receptor dimerization in the S-locus receptor kinase extracellular domain. PNAS **104:** 12211-12216.

NASRALLAH, J. B., T. H. KAO, M. L. GOLDBERG and M. E. NASRALLAH, 1985 A cDNA clone encoding an S-locus specific glycoprotein from *Brassica oleracea*. Nature **318:** 263-267.

NASRALLAH, J. B., T. NISHIO and M. E. NASRALLAH, 1991 The self-incompatibility genes of *Brassica*: expression and use in genetic ablation of floral tissues. Annual Review of Plant Physiology and Plant Molecular Biology **42**.

NASRALLAH, M. E., P. LIU, S. SHERMAN-BOYLES, N. A. BOGGS and J. B. NASRALLAH, 2004 Natural variation in expression of self-incompatibility in *Arabidopsis thaliana*: Implications for the evolution of selfing. PNAS **101:** 16070-16074.

NEHER, E., 1994 How frequent are correlated changes in families of protein sequences? PNAS **91:** 98-108.

NISHIO, T., and M. KUSABA, 2000 Sequence diversity of *SLG* and *SRK* in *Brassica oleracea* L. Annals of botany **85:** 141-146.

NORDBORG, M., T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The Pattern of Polymorphism in *Arabidopsis thaliana*. PLoS Biology **3:** e196.

NOU, I. S., M. WATANABE, A. ISOGAI and K. HINATA, 1993 Comparison of S-alleles and S-glycoproteins between two wild populations of Brassica campestris in Turkey and Japan. Sexual Plant Reproduction **6:** 79-86.

O'KANE, S. L., and I. AL-SHEHBAZ, 1997 A synopsis of Arabidopsis (Brassicaceae). Novon **7:** 323-327.

OCKENDON, D. J., 2000 The S-allele collection of *Brassica oleracea*. Acta Horticulturae **539:** 25-30.

PAUWELS, M., H. FRÉROT, I. BONNIN and P. SAUMITOU-LAPRADE, 2006 A broad-scale analysis of population differenciation for Zn tolerance in an emerging model species for tolerance study: Arabidopsis halleri. Journal of Evolutionary Biology **19:** 1838-1850.

PETERSON, D. G., J. P. TOMKINS, D. A. FRISCH, R. A. WING and A. H. PATERSON, 2000 Construction of plant bacterial artificial chromosome (BAC) libraries: An illustrated guide. Journal of Agricultural Genomics **5**.

PIRIYAPONGSA, J., and I. K. JORDAN, 2007 A family of human microRNA genes from miniature inverted repeat transposable elements. PLoS ONE **2:** e203.

PRIGODA, N. L., A. NASSUTH and B. K. MABLE, 2005 Phenotypic and genotypic expression of self-incompatibility haplotypes in *Arabidopsis lyrata* suggests unique origin of alleles in different dominance classes. Molecular Biology and Evolution **22:** 1609–1620.

RAHMAN, M. H., K. SUWABE, J. KOHORI, R. N. TOMITA, K. KAKEDA *et al.*, 2007 Physical size of the S locus region defined by genetic recombination and genome sequencing in *Ipomoea trifida*, Convolvulaceae. Sexual Plant Reproduction **20:** 63-72.

RASPÉ, O., and J. R. KOHN, 2002 S-allele diversity in *Sorbus aucuparia* and *Crataegus monogyna* (Rosaceae: Maloideae). Heredity **88:** 458-465.

RASPÉ, O., and J. R. KOHN, 2007 Population structure at the S-locus of Sorbus aucuparia L. (Rosaceae: Maloideae). Molecular Ecology **16:** 1315-1325.

REA, A., and J. B. NASRALLAH, 2008 Self-incompatibility systems: barriers to self-fertilization in flowering plants. International Journal of Developmental Biology **52:** 627-636.

RICE, W. R., 1996 Evolution of Y sex chromosome in animals. Bioscience **46:** 331-343.

RICHMAN, A. D., S. SCHAEFFER and M. K. UYENOYAMA, 1995 S-allele sequence diversity in natural populations of *Solanum carolinense* (Horsenettle). Heredity **75:** 405-415.

ROSS-IBARRA, J., S. I. WRIGHT, J. P. FOXE, A. KAWABE, L. DEROSE-WILSON *et al.*, 2008 Patterns of polymorphism and demographic history in natural populations of Arabidopsis lyrata. PLoS ONE **3:** e2411.

ROSS, M. T., D. V. GRAFHAM, A. J. COFFEY, S. SCHERER, K. MCLAY *et al.*, 2005 The DNA sequence of the human X chromosome. Nature **434:** 325-337.

ROUX, C., V. CASTRIC, M. PAUWELS, S. I. WRIGHT, P. SAUMITOU-LAPRADE *et al.*, 2011 Does speciation between *Arabidopsis halleri* and *Arabidopsis lyrata* coincide with major changes in a molecular target of adaptation? PLoS ONE **6:** e26872

ROZAS, J., and R. ROZAS, 1995 DnaSP, DNA sequence polymorphism: an interactive program for estimating Population Genetics parameters from DNA sequence data. Computer Applications in Biosciences **11:** 621-625.

ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496-2497.

ROZEN, S., and H. J. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by M. S. KRAWETZ S. Humana Press, Totowa, NJ.

RUGGIERO, M. V., B. JACQUEMIN, V. CASTRIC and X. VEKEMANS, 2007 Hitch-hiking to a locus under balancing selection : high sequence diversity and low population subdivision at the S-locus genomic region in Arabidopsis halleri. Genetical Research **89:** 1-13.

SALAMOV, A., and V. SOLOVYEV, 2000 Ab initio gene finding in *Drosophila genomic* DNA. Genome Research **10:** 516-522.

SAMPSON, D. R., 1974 Equilibrium frequencies of sporophytic self-incompatibility alleles. Canadian Journal of Genetics and Cytology **16:** 611-618.

SANMIGUEL, P., A. TICKHONOV, Y. K. JIN, A. MELAKE-BERHAN, P. S. SPRINGER *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-768.

SATO, K., T. NISHIO, R. KIMURA, M. KUSABA, T. SUZUKI *et al.*, 2002 Coevolution of the S-locus genes SRK, SLG and SP11/SCR in *Brassica oleracea* and *B. rapa*. Genetics **162:** 931-940.

SATO, Y., R. FUJIMOTO, K. TORIYAMA and T. NISHIO, 2003 Commonality of self-recognition specificity of S haplotypes between *Brassica oleracea* and *Brassica rapa*. Plant Molecular Biology **52:** 617-626.

SCHIERUP, M. H., J. BECHSGAARD and F. B. CHRISTIANSEN, 2008 Selection at work in self-incompatible *Arabidopsis lyrata*. II. spatial distribution of S haplotypes in Iceland. Genetics **180:** 1051-1059.

SCHIERUP, M. H., J. S. BECHSGAARD, L. H. NIELSEN and F. B. CHRISTIANSEN, 2006 Selection at work in self-incompatible *Arabidopsis lyrata*: mating patterns in a natural population. Genetics **172:** 477–484.

SCHIERUP, M. H., A. M. MIKKELSEN and J. HEIN, 2001 Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genetics **159:** 1833-1844.

SCHIERUP, M. H., X. VEKEMANS and F. B. CHRISTIANSEN, 1997 Evolutionary dynamics of sporophytic self-incompatibility alleles in plants. Genetics **147:** 835-846.

SCHNEEBERGER, K., S. OSSOWSKI, F. OTT, J. D. KLEIN, X. WANG *et al.*, 2011 Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. PNAS **108:** 10249-10254.

SCHOEN, D. J., M. O. JOHNSTON, A. L'HEUREUX and J. V. MARSOLAIS, 1997 Evolutionary history of the mating system in *Amsinckia* (Boraginaceae). Evolution **51:** 1090-1099.

SCHOPFER, C. R., M. E. NASRALLAH and J. B. NASRALLAH, 1999 The male determinant of self-incompatibility in *Brassica*. Science **286:** 1697-1700.

SHERMAN-BROYLES, S., N. BOGGS, A. FARKAS, P. LIU, J. VREBALOV *et al.*, 2007 S locus genes and the evolution of self-fertility in *Arabidopsis thaliana*. The Plant Cell **19:** 94-106.

SHIBA, H., T. KAKIZAKI, M. IWANO, Y. TARUTANI, M. WATANABE *et al.*, 2006 Dominance relationships between self-incompatibility alleles controlled by DNA methylation. Nature Genetics **38:** 297-299.

SHIBA, H., M. KENMOCHI, M. SUGIHARA, M. IWANO, S. KAWASAKI *et al.*, 2003 Genomic organization of the S-locus region of Brassica. Bioscience Biotechnology and Biochemistry **67:** 622-626.

SHIMIZU, K. K., J. M. CORK, A. L. CAICEDO, C. A. MAYS, R. C. MOORE *et al.*, 2004 Darwinian Selection on a Selfing Locus. Science **306:** 2081-2084.

SHIMIZU, K. K., R. SHIMIZU-INATSUGI, T. TSUCHIMATSU and M. D. PURUGGANAN, 2008 Independent origins of self-compatibility in *Arabidopsis thaliana*. Molecular Ecology **17:** 704-710.

SKALETSKY, H., T. KURODA-KAWAGUCHI, P. J. MINX, H. S. CORDUM, L. HILLIER *et al.*, 2005 The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423:** 825-837.

SMIT, A. F., R. HUBLEY and P. GREEN, 1996-2006 RepeatMasker at http://repeatmasker.org.

STEIN, J. C., B. HOWLETT, D. C. BOYES, M. E. NASRALLAH and J. B. NASRALLAH, 1991 Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus from *Brassica oleracea*. PNAS **88:** 8816-8820.

STEINBACHS, J. E., and K. E. HOLSINGER, 2002 S-RNase-mediated gametophytic self-incompatibility is ancestral in eudicots. Molecular Biology and Evolution **19:** 825-829.

STOECKEL, S., E. K. KLEIN, S. ODDOU-MURATORIO, B. MUSCH and S. MARIETTE, 2011 Microevolution of S-allele frequencies in wild cherry populations: respective impacts of negative frequency dependent selection and genetic drift. Evolution **online in advance of print**.

SUZUKI, G., N. KAI, T. HIROSE, K. FUKUI, T. NISHIO *et al.*, 1999 Genomic organization of the S locus: identification and characterization of genes in SLG/SRK region of S^9 haplotype of *Brassica campestris* (syn. *rapa*). Genetics **153:** 391-400.

TAKASAKI, T., K. HATAKEYAMA, G. SUZUKI, H. WATANABE, A. ISOGAI *et al.*, 2000 The S receptor kinase determines self-incompatibility in Brassica stigma. Nature **403:** 913-916.

TAKAYAMA, S., and A. ISOGAI, 2005 Self-incompatibility in plants. Annual Review of Plant Biology **56:** 467-489.

TAKEBAYASHI, N., and P. L. MORRELL, 2001 Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. American Journal of Botany **88:** 1143-1150.

TAKUNO, S., R. FUJIMOTO, T. SUGIMURA, K. SATO, S. OKAMOTO *et al.*, 2007 Effects of recombination on hittchhiking diversity in the Brassica self-incompatibility locus complex. Genetics **177:** 949-958.

TAMURA, K., D. PETERSON, N. PETERSON, G. STECHER, M. NEI *et al.*, 2011 MEGA5: molecular evolutionary genetics analysis using likelihood, distance, and parsimony methods. Molecular Biology and Evolution.

TANG, C., C. TOOMAJIAN, S. SHERMAN-BROYLES, V. PLAGNOL, Y. GUO *et al.*, 2007 The evolution of selfing in *Arabidopsis thaliana*. Science **317:** 1070-1072.

TARUTANI, Y., H. SHIBA, M. IWANO, T. KAKIZAKI, G. SUZUKI *et al.*, 2010 *Trans*-acting small RNA determines dominance relationships in *Brassica* self-incompatibility. Nature **466:** 983-986.

THE *ARABIDOPSIS* GENOME INITIATIVE, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

THOMAS, S. G., and V. E. FRANKLIN-TONG, 2004 Self-incompatibility triggers programmed cell death in *Papaver* pollen. Nature **429:** 305-309.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research **22:** 4673-4680.

TOMITA, R. N., G. SUZUKI, K. YOSHIDA, Y. YANO, T. TSUCHIYA *et al.*, 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.

TOTH, G., G. DEAK, E. BARTA and G. KISS, 2006 PLOTREP: a web tool for defragmentation and visual analysis of dispersed genomic repeats. Nucleic Acids Research **34:** W708-W713.

TSUCHIMATSU, T., K. SUWABE, R. SHIMIZU-INATSUGI, S. ISOKAWA, P. PAVLIDIS *et al.*, 2010 Evolution of self-compatibility in Arabidopsis by a mutation in the male specificity gene. Nature **464:** 1342-1346.

USHIJIMA, K., H. SASSA, M. TAMURA, M. KUSABA, R. TAO *et al.*, 2001 Characterization of the S-locus region of almond (*Prunus dulcis*): analysis of a somaclonal mutant and a cosmid contig for an S haplotype. Genetics **158:** 379-386.

UYENOYAMA, M. K., 1995 A generalized least-squares estimate for the origin of self-incompatibility. Genetics **139:** 975-992.

UYENOYAMA, M. K., 1997 Genealogical structure among alleles regulating self-incompatibility. Genetics **147:** 1389-1400.

UYENOYAMA, M. K., 2000 Evolutionary Dynamics of Self-Incompatibility Alleles in Brassica. Genetics **156:** 351-359.

UYENOYAMA, M. K., Y. ZHANG and E. NEWBIGIN, 2001 On the origin of self-incompatibility haplotypes: transition through self-compatible intermediates. Genetics **157:** 1805-1817.

VEKEMANS, X., and M. SLATKIN, 1994 Gene and allelic genealogies at the gametophytic self-incompatibility locus Genetics **137:** 1157-1165.

VIEIRA, C. P., D. CHARLESWORTH and J. VIEIRA, 2003 Evidence for rare recombination at the gametophytic self-incompatibility locus. Heredity **91:** 262-267.

VITTE, C., and O. PANAUD, 2005 LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenetic and Genome Research **110:** 91-107.

VOINET, O., 2009 Origin, biogenesis, and activity of plant MicroRNAs. Cell **136:** 669-687.

VORECHOVSKY, I., 2010 Transposable elements in disease-associated cryptic exons. Human genetics **127:** 135-154.

WATANABE, M., A. ITO, Y. TAKADA, C. NIMOMIYA, T. KAKIZAKI *et al.*, 2000 Highly divergent sequences of the pollen self-incompatibility (S) gene in class-I S haplotypes of *Brassica campestris* (syn. *rapa*) L. FEBS Letter **473:** 139-144.

WATERHOUSE, A. M., J. B. PROCTER, D. M. A. MARTIN, M. CLAMP and G. J. BARTON, 2009 Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench. Bioinformatics **25:** 1189-1191.

WEIGEL, D., and R. MOTT, 2009 The 1001 genome project for *Arabidopsis thaliana*. Genome Biology **10:** 107.

WESSLER, S. R., 2006 Transposable elements and the evolution of eukaryotic genomes. PNAS **103:** 17600-17601.

WHEELER, M. J., S. A. ARMSTRONG, V. E. FRANKLIN-TONG and F. C. H. FRANKLIN, 2003 Genomic organization of the *Papaver rhoeas* self-incompatibility S1 locus. Journal of Experimental Botany **54:** 131-139.

WHEELER, M. J., B. H. D. DE GRAAF, N. HADJIOSIF, R. M. PERRY, N. S. POULTER *et al.*, 2009 Identification of the pollen self-incompatibility determinant in Papaver rhoeas. Nature **459:** 992-995.

WHITTLE, C. A., and H. JOHANNESSON, 2011 Evidence of the accumulation of allele-specific non-synonymous substitutions in the young region of recombination suppression within the mating-type chromosomes of *Neurospora tetrasperma*. Heredity **107:** 305-314.

WICKER, T., F. SABOT, A. HUA-VAN, J. L. BENNETZEN, P. CAPY *et al.*, 2007 A unified classification system for eukaryotic transposable elements. Nature Review Genetics **8:** 973-982.

WRIGHT, S., 1939 The distribution of self-sterility alleles in populations. Genetics **24:** 538-552.

WRIGHT, S. I., N. AGRAWAL and T. E. BUREAU, 2003 Effects of recombinaison rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Research **13:** 1897-1903.

YU, Q., S. HOU, R. HOBZA, F. A. FELTUS, X. WANG *et al.*, 2007 Chromosomal location and gene paucity of the male specific region on papaya Y chromosome. Molecular Genetics and Genomics **278:** 177-185.

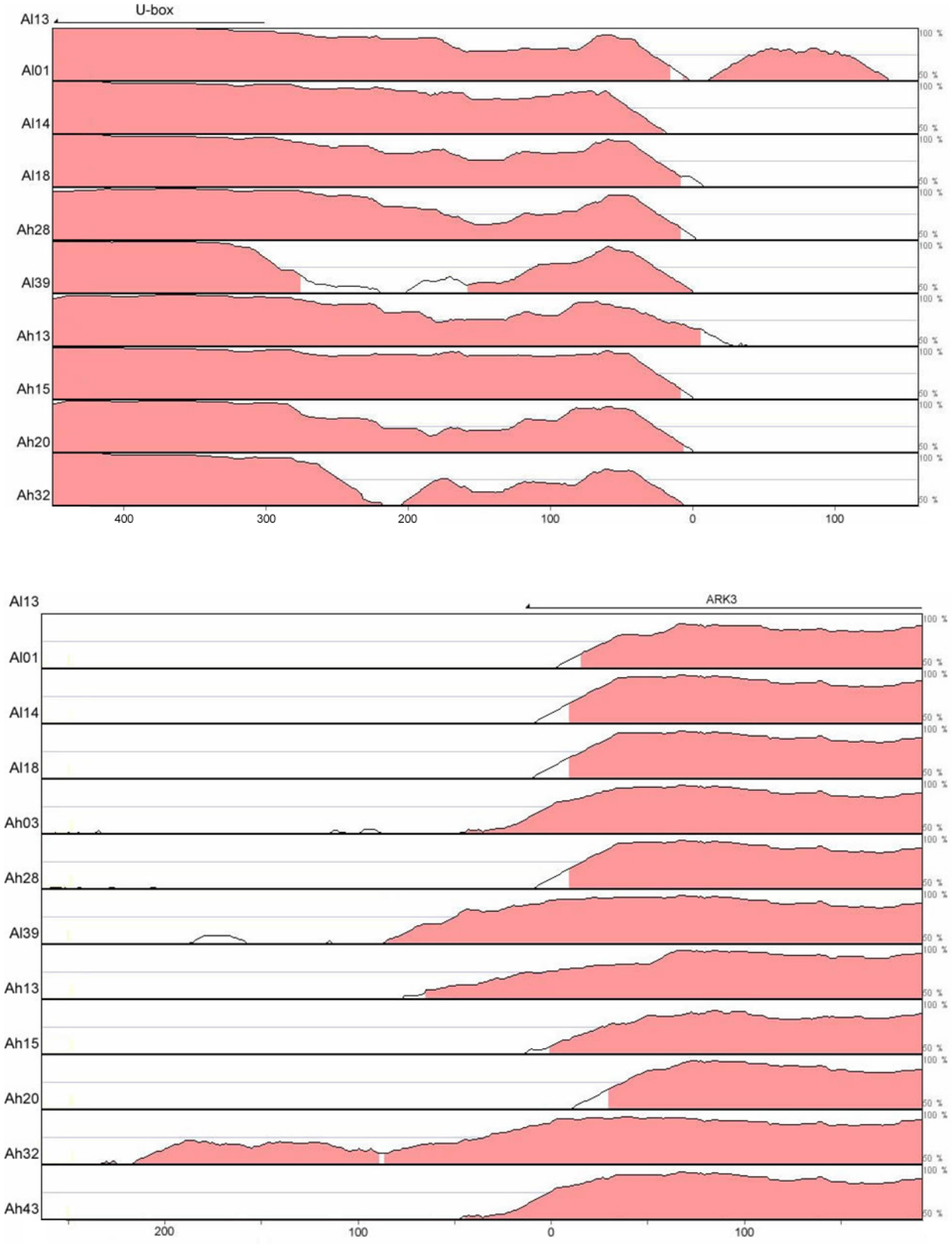
ZHOU, J., F. WANG, W. MA, Y. ZHANG, B. HAN *et al.*, 2003 Structural and transcriptional analysis of S-locus F-box genes in *Antirrhinum*. Sexual Plant Reproduction **16:** 165-177.

ANNEXES ET MATÉRIEL SUPPLÉMENTAIRE

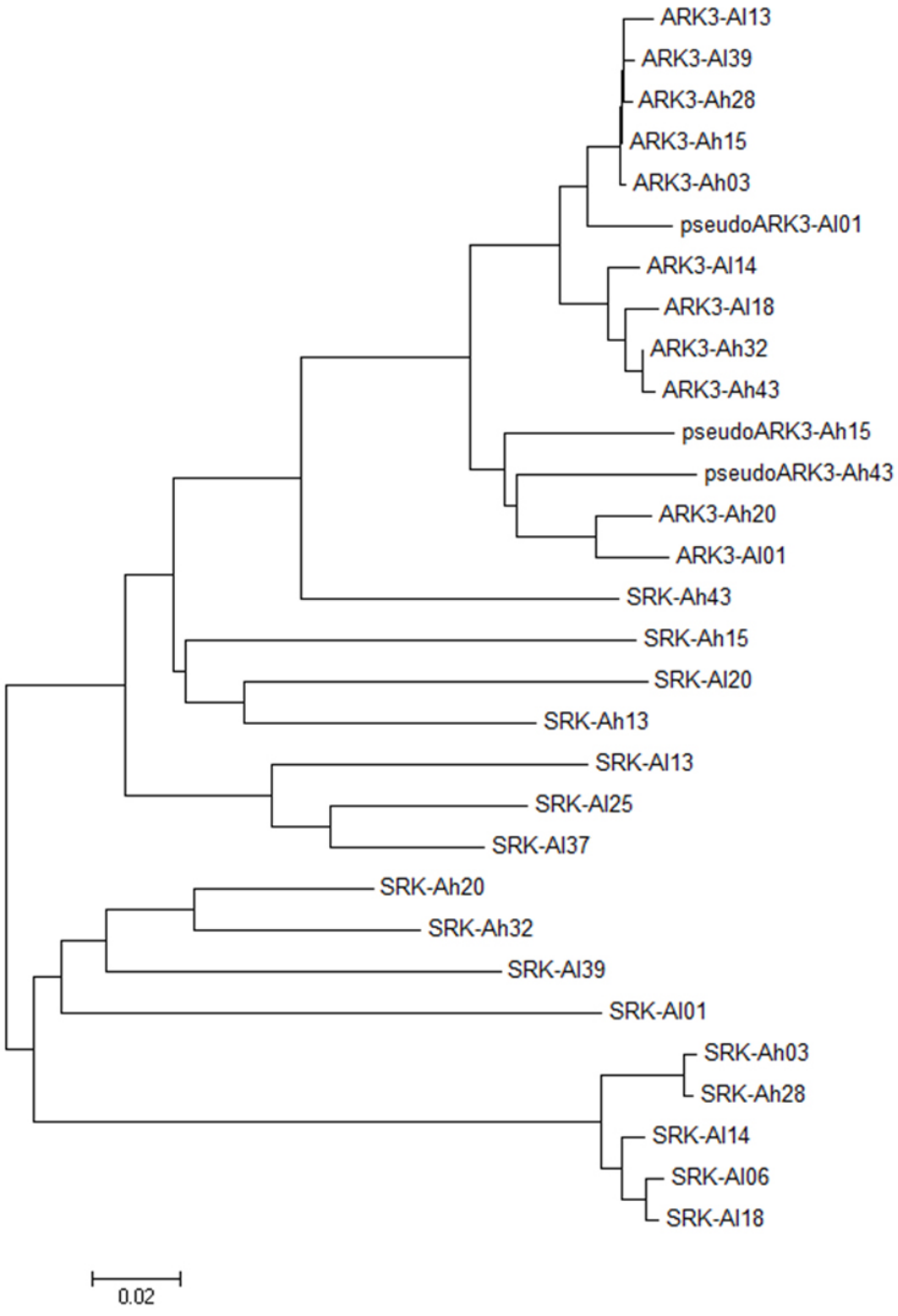
Supplementary table 1. Description of the different clones.
Two clones were necessary to cover the entire S-locus for three haplotypes : *Al18, Ah28* and *Ah15*.

BAC	BAC size	Number of reads	Average size of reads	Coverage	Number of contigs	Genes at extremities
Al01	100 937	10 900	331	35.74	4	*At4g21270*[a] - *At4g21430*
Al14	117 539	54 361	344	159.09	3	*At4g21300* - *At4g21480*
Al18 BAC1	88 062	15 066	372	63.64	3	*At4g21300*[a] - *SRK*[a]
Al18 BAC2	96 061	5 126	346	18.46	4	*SRK* - *At4g21500*
Ah03	84 197	6 848	359	29.20	8	*SCR* - *At4g21430*
Ah28 BAC1	101 609	9 569	362	34.09	2	*At4g21323* - *SRK*[a]
Ah28 BAC2	94 078	18 635	349	69.13	3	*SCR* - *At4g21470*
Al39	90 060	17 367	376	75.22	5	*At4g21320* - *At4g21410*[a]
Ah13	88 292	11 931	348	47.03	5	*At4g21326* - *SRK*
Ah15 BAC1	109 343	8 877	351	28.50	9	*SCR* - *At4g21440*
Ah15 BAC2	85 357	10 636	345	42.99	4	*At4g21310* - *SRK*
Ah20	105 142	27 578	350	91.80	5	*At4g21300*[a] - *At4g21430*[a]
Ah32	115 243	15 378	351	46.84	5	*At4g21326* - *At4g21440*
Ah43	95 096	12 656	378	47.54	8	*SCR*[a] - *ARK3*[a]

[a] The sequence is incomplete because the BAC sequence ends into the gene

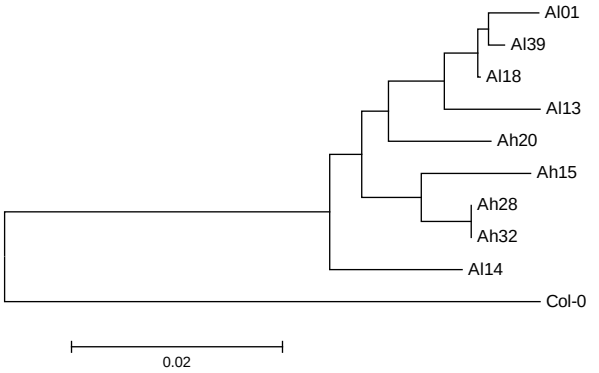


Supplemental figure 1. Sequence conservation at the S-locus boundaries between *Al13* (the reference *A. lyrata* genome) and each of the other haplotypes. Sequences not available for the *U-box* side (*Ah03* and *Ah43*) were not represented. Distance from the homology breakpoint is indicated under each graph.

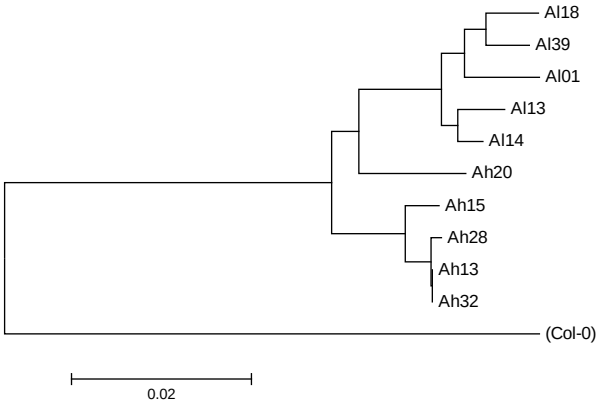


Supplementary figure 2. Phylogeny of *pseudo-ARK3* sequences, *SRK* and *ARK3*. Phylogeny was constructed using a Minimum Evolution analysis.

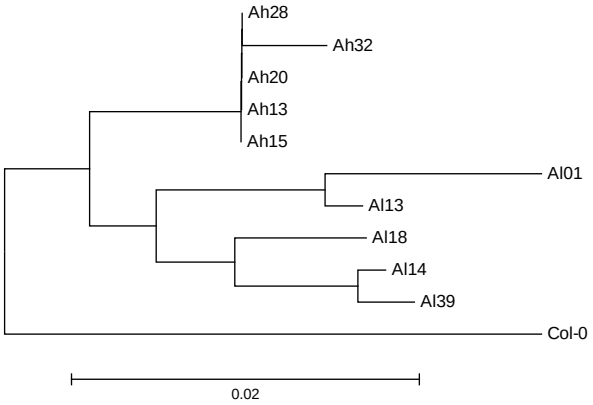
Supplementary figure 3. Separate phylogenies of the S-locus Region genes. Phylogenies were obtained by the Minimum Evolution method, and are based on protein sequences, with the *A. thaliana* reference sequences (Col-0) as outgroup.



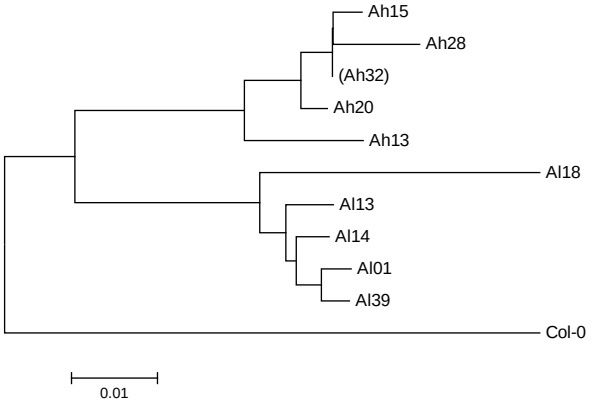
At4g21323



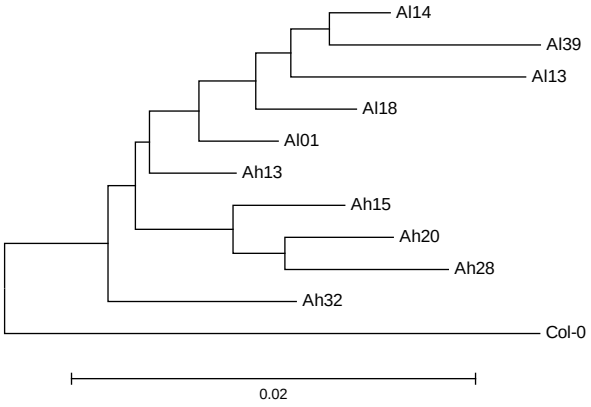
At4g21326



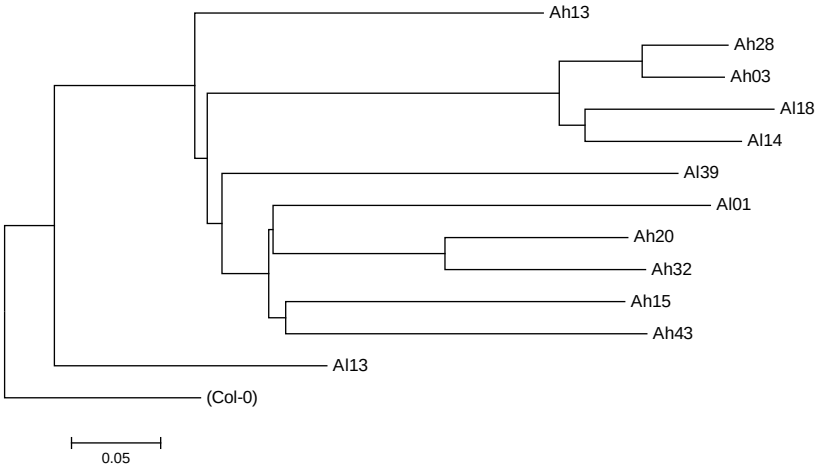
At4g21330



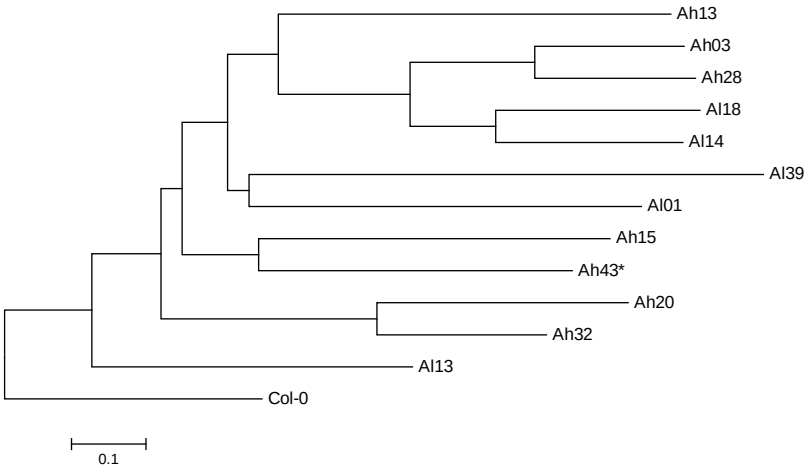
At4g21340



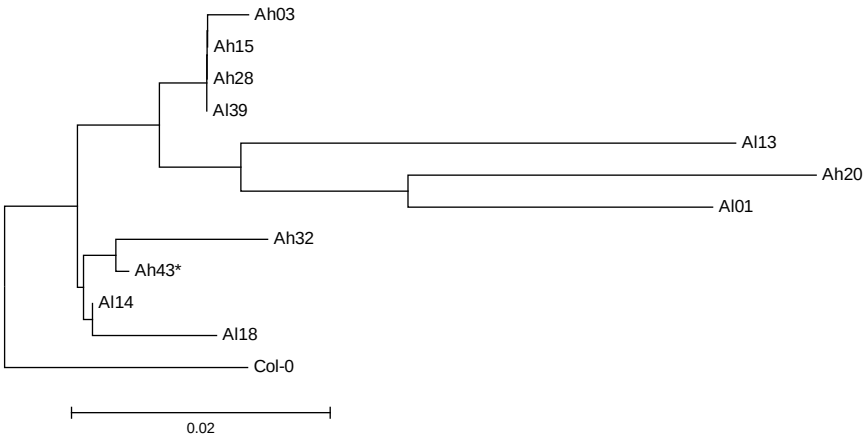
U-box



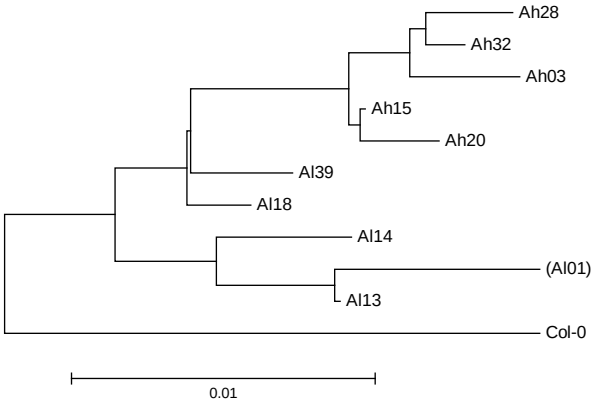
SRK



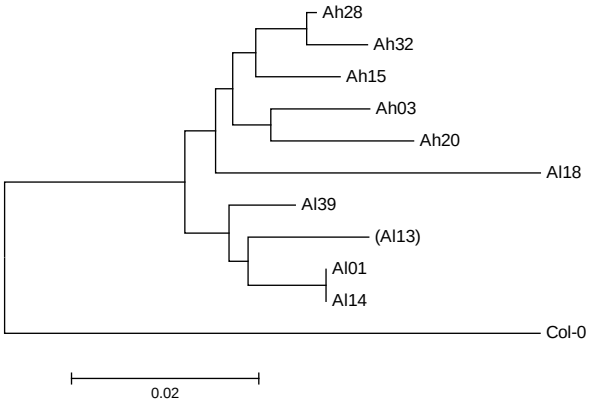
SCR



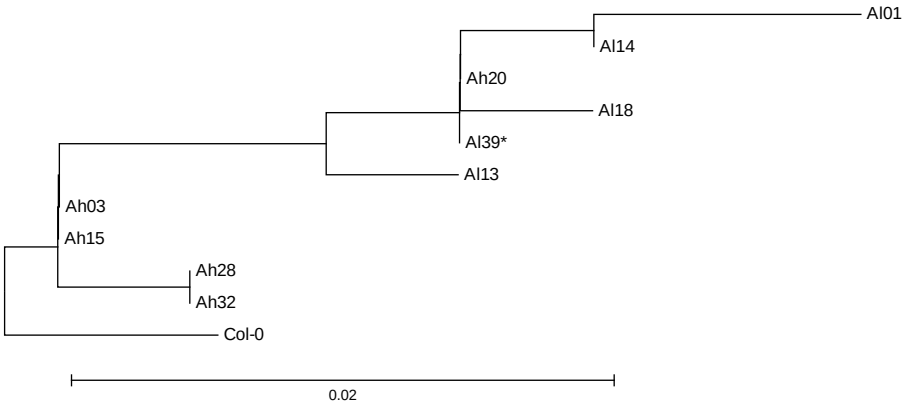
ARK3



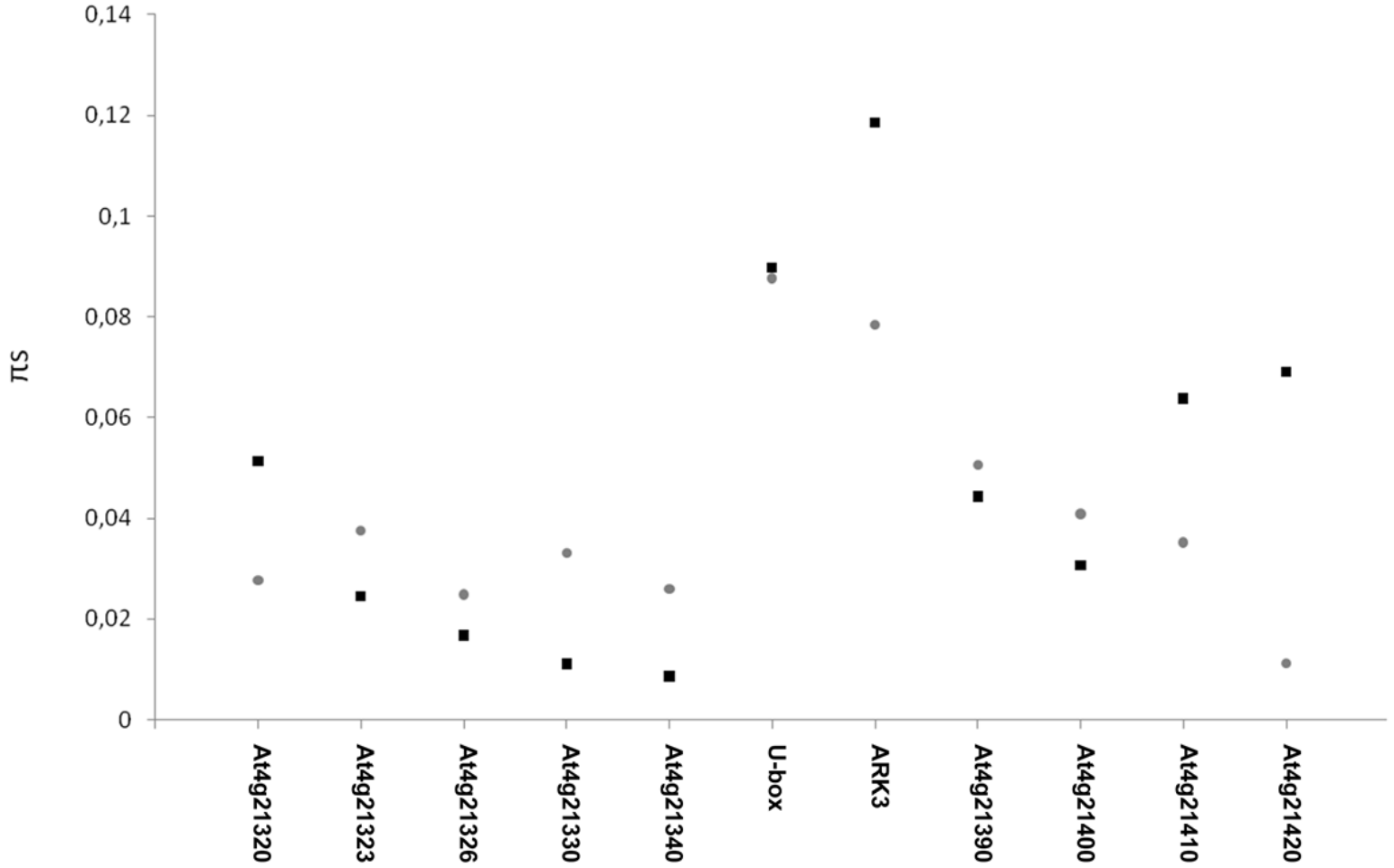
At4g21390



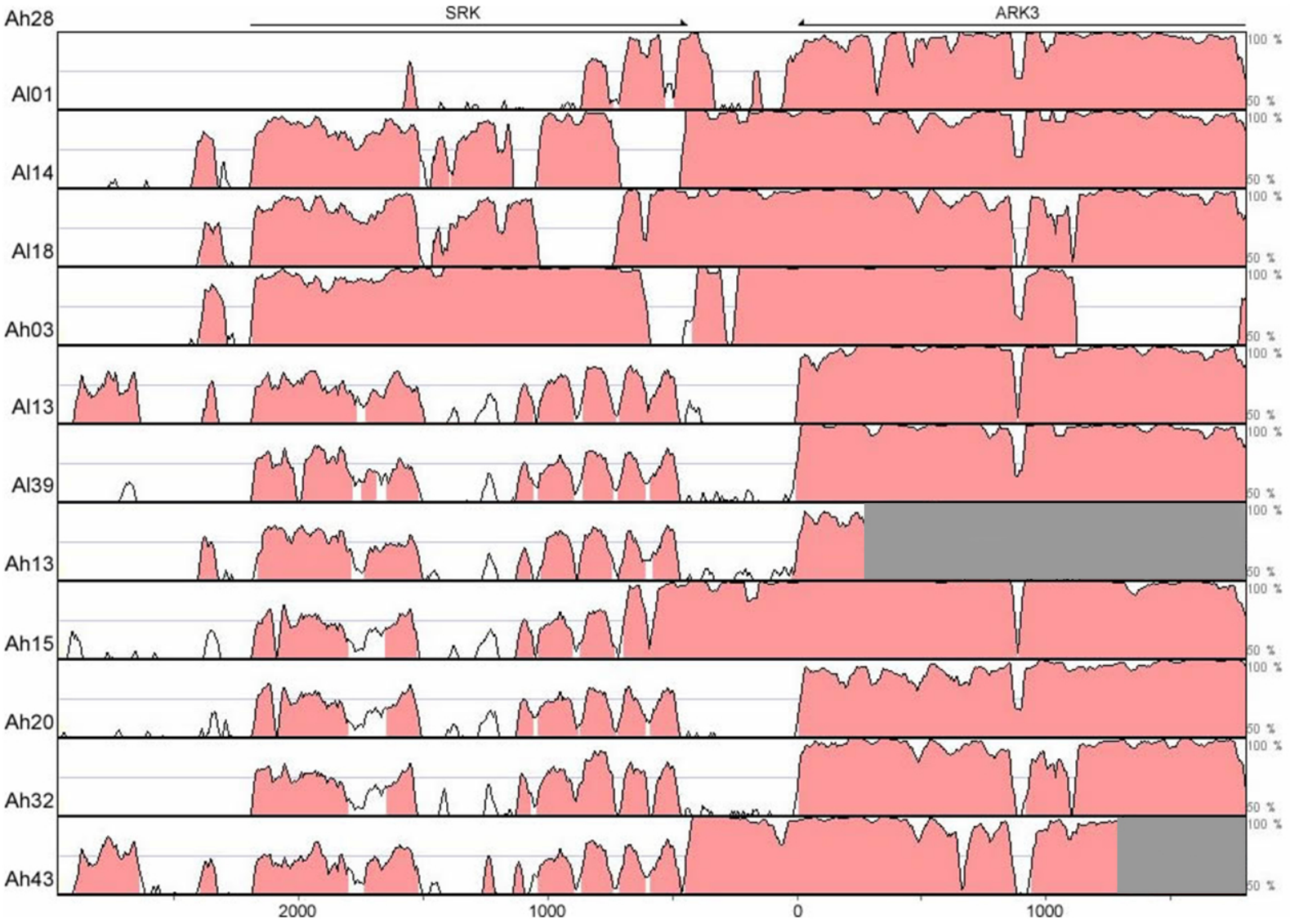
At4g21400



At4g21410

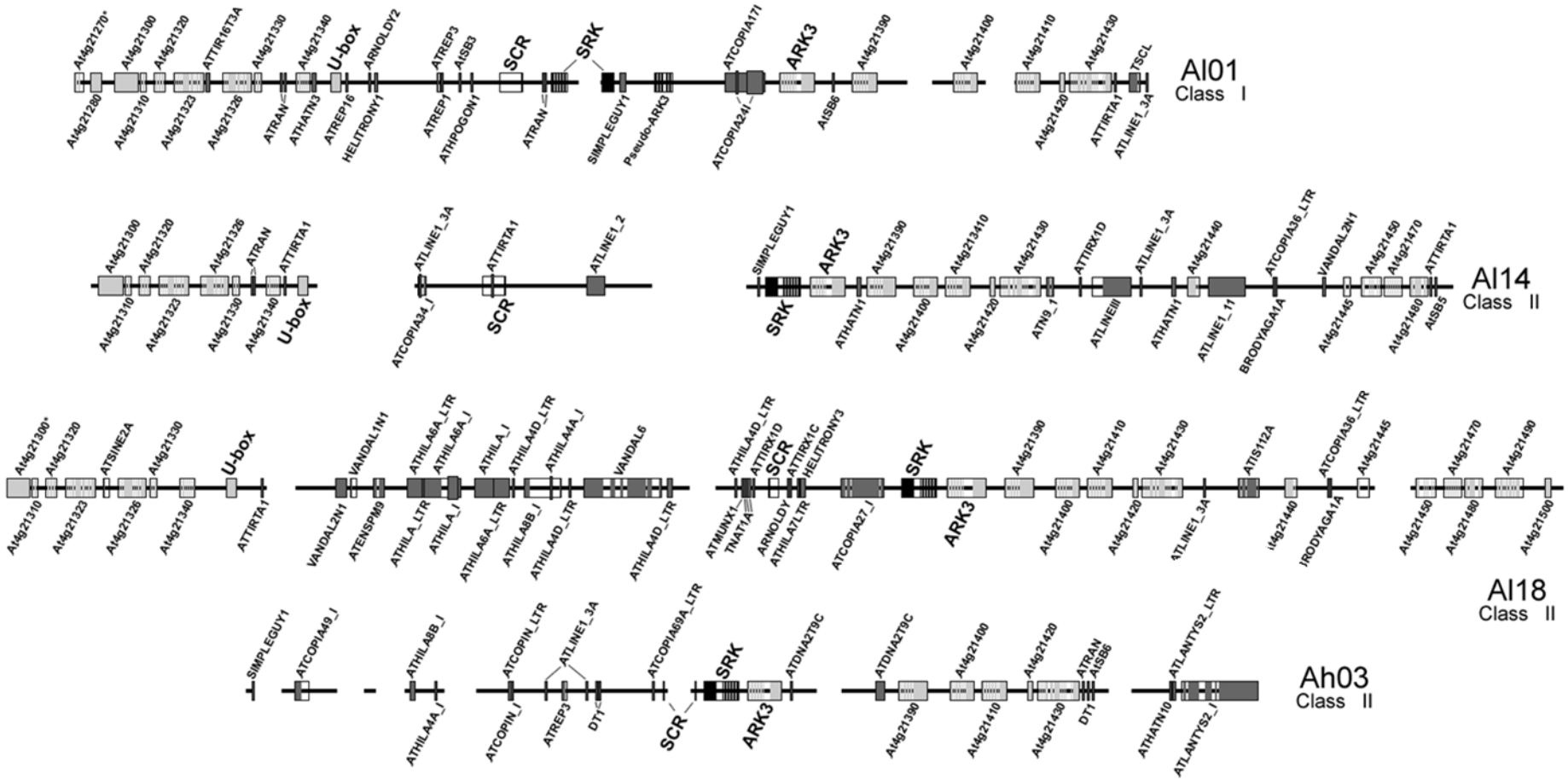


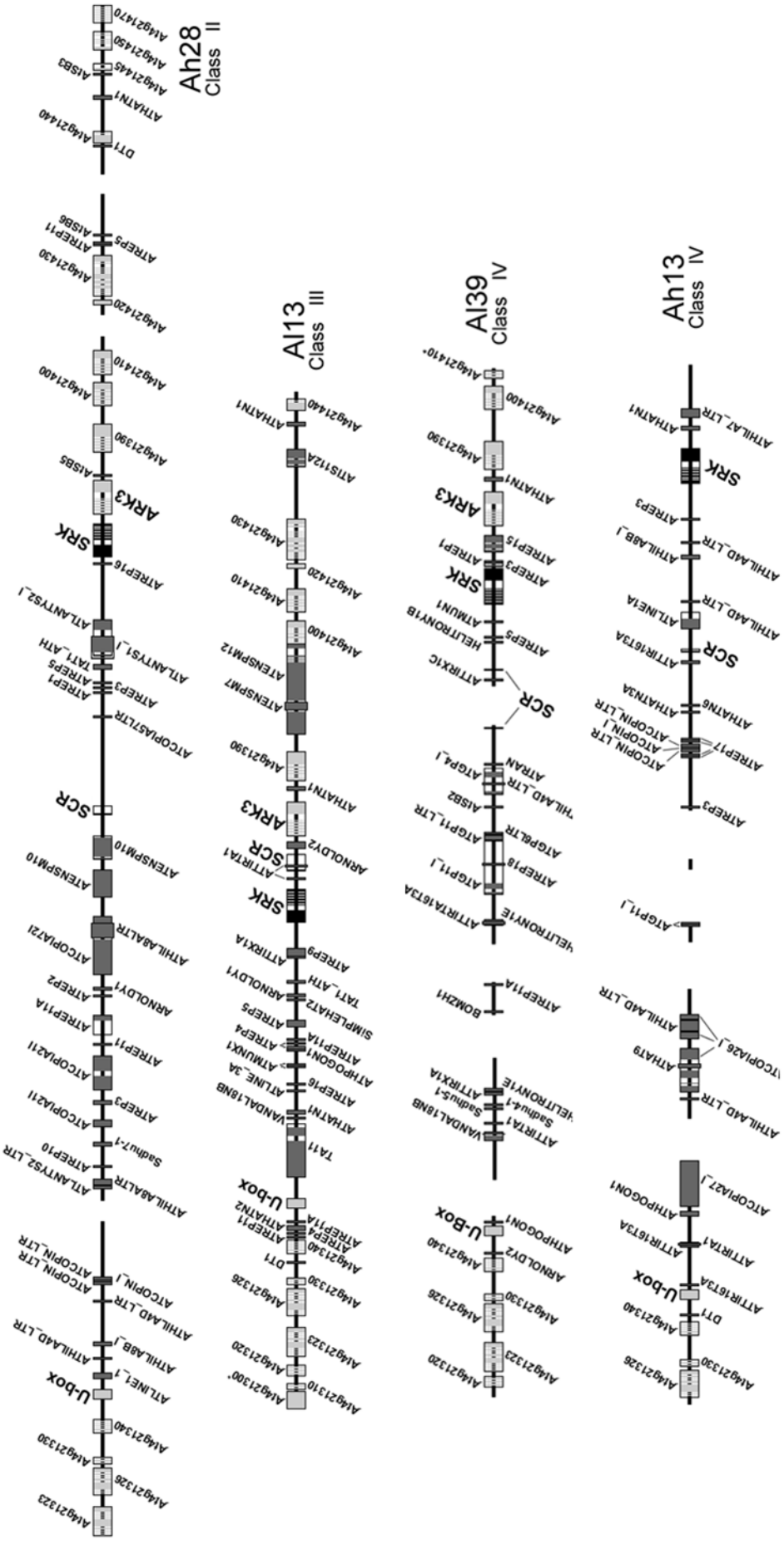
Supplementary figure 4. Synonymous nucleotide diversity (Π_S) at S-locus flanking genes for *A. halleri* (black) and *A. lyrata* (gray), estimated using DnaSP (ROZAS and ROZAS 1995).

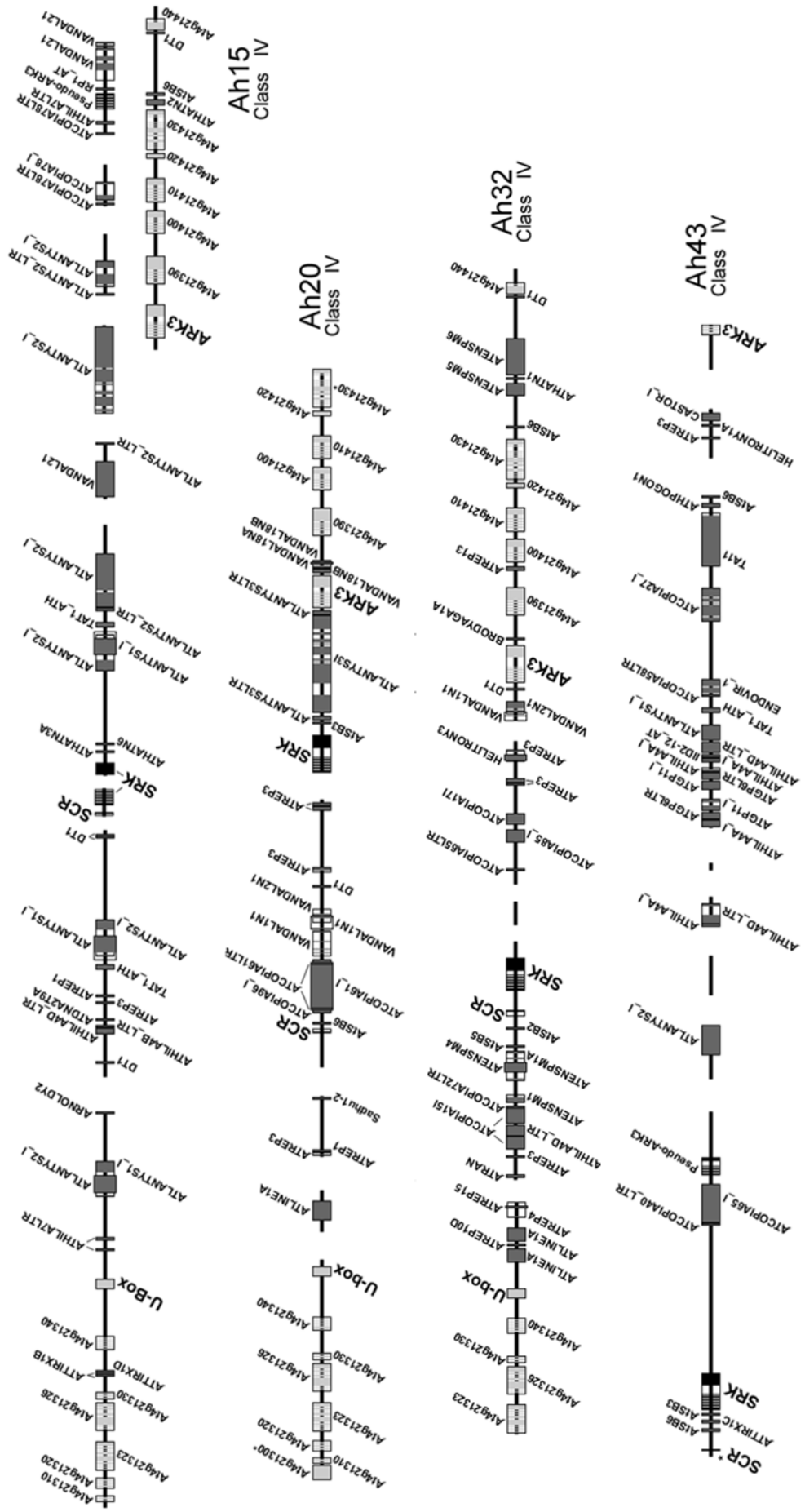


Supplementary figure 5. Sequence conservation in the *SRK-ARK3* region between *Ah28* (Class II) and each of the other haplotypes. Distance from homology breakpoint is indicated under the graph.

Supplementary figure 6. Annotation of genes and transposable elements for the 12 S-haplotypes. The S-locus genes are represented in black rectangles, with delimitation of their exons. Other genes are depicted in light gray. Transposable elements are shown in dark gray, and their fragmentation is indicated by white gaps.







Supplementary table 2. List of the transposable elements detected in the BAC sequences.

Haplotype	Name	Type	Size
Al01	ARNOLDY2	MuDR	156
Al01	ATCOPIA17I	Copia	2000
Al01	ATCOPIA24I	Copia	1924
Al01	ATHATN3	hAT	457
Al01	ATHPOGON1	Mariner/Tc1	132
Al01	ATLINE1_3A	L1	308
Al01	ATRAN	DNA transposon	116
Al01	ATRAN	DNA transposon	142
Al01	ATRAN	DNA transposon	148
Al01	ATRAN	DNA transposon	241
Al01	ATREP1	Helitron	231
Al01	ATREP16	DNA transposon	158
Al01	ATREP3	Helitron	357
Al01	AtSB3	SINE	249
Al01	AtSB6	SINE	345
Al01	ATTIR16T3A	Mariner/Tc1	491
Al01	ATTIRTA1	Mariner/Tc1	243
Al01	HELITRONY1B	Helitron	366
Al01	SIMPLEGUY1	MSAT	706
Al01	TSCL	Non LTR retrotransposon	1153
Al14	ATCOPIA34_I	Copia	457
Al14	ATCOPIA36_LTR	Copia	68
Al14	ATHATN1	hAT	480
Al14	ATHATN1	hAT	504
Al14	ATLINE1_11	Non LTR retrotransposon	3849
Al14	ATLINE1_2	L1	1911
Al14	ATLINE1_3A	L1	80
Al14	ATLINE1_3A	L1	87
Al14	ATLINEIII	Non LTR retrotransposon	2939
Al14	ATN9_1	MuDR	858
Al14	ATRAN	DNA transposon	142
Al14	ATRAN	DNA transposon	116
Al14	AtSB5	SINE	198
Al14	ATTIRTA1	Mariner/Tc1	60
Al14	ATTIRTA1	Mariner/Tc1	246
Al14	ATTIRTA1	Mariner/Tc1	239
Al14	ATTIRX1D	DNA transposon	339
Al14	BRODYAGA1A	DNA transposon	116
Al14	SIMPLEGUY1	MSAT	115
Al14	VANDAL2N1	MuDR	438
Al18	ARNOLDY2	MuDR	307
Al18	ATCOPIA27_I	Copia	4217

Al18	ATCOPIA36_LTR	Copia	68
Al18	ATENSPM9	EnSpm	1008
Al18	ATHILA_I	Gypsy	1968
Al18	ATHILA_I	Gypsy	1141
Al18	ATHILA_LTR	Gypsy	162
Al18	ATHILA4A_I	Gypsy	306
Al18	ATHILA4D_LTR	Gypsy	197
Al18	ATHILA4D_LTR	Gypsy	564
Al18	ATHILA4D_LTR	Gypsy	212
Al18	ATHILA4D_LTR	Gypsy	212
Al18	ATHILA6A_I	Gypsy	2096
Al18	ATHILA6A_LTR	Gypsy	1731
Al18	ATHILA6A_LTR	Gypsy	1565
Al18	ATHILA7LTR	Gypsy	233
Al18	ATHILA8B_I	Gypsy	768
Al18	ATIS112A	Harbinger	1978
Al18	ATLINE1_3A	L1	332
Al18	ATMUNX1	MuDR	101
Al18	ATSINE2A	SINE	180
Al18	ATTIRTA1	Mariner/Tc1	1937
Al18	ATTIRX1C	DNA transposon	191
Al18	ATTIRX1D	DNA transposon	345
Al18	BRODYAGA1A	DNA transposon	116
Al18	HELITRONY3	Helitron	619
Al18	TNAT1A	DNA transposon	381
Al18	TNAT1A	DNA transposon	77
Al18	TNAT1A	DNA transposon	73
Al18	VANDAL1N1	MuDR	231
Al18	VANDAL2N1	MuDR	1190
Al18	VANDAL6	MuDR	5665
Ah03	ATCOPIA49_I	Copia	302
Ah03	ATCOPIA69A_LTR	Copia	784
Ah03	ATCOPIN_I	LTR Retrotransposon	2904
Ah03	ATCOPIN_LTR	LTR Retrotransposon	302
Ah03	ATDNA2T9C	MuDR	290
Ah03	ATDNA2T9C	MuDR	431
Ah03	ATHATN10	hAT	171
Ah03	ATHILA4A_I	Gypsy	552
Ah03	ATHILA8B_I	Gypsy	1727
Ah03	ATLANTYS2_I	Gypsy	6427
Ah03	ATLANTYS2_LTR	Gypsy	198
Ah03	ATLINE1_3A	L1	1937
Ah03	ATLINE1_3A	L1	65
Ah03	ATRAN	DNA transposon	1671
Ah03	ATREP3	Helitron	395
Ah03	AtSB6	SINE	1326

Ah03	DT1	Mariner/Tc1	335
Ah03	DT1	Mariner/Tc1	239
Ah03	DT1	Mariner/Tc1	148
Ah03	SIMPLEGUY1	MSAT	130
Ah28	ARNOLDY1	MuDR	431
Ah28	ATCOPIA15I	Copia	1727
Ah28	ATCOPIA21I	Copia	3145
Ah28	ATCOPIA21I	Copia	752
Ah28	ATCOPIA57LTR	Copia	126
Ah28	ATCOPIA72_I	Copia	784
Ah28	ATCOPIA72_I	Copia	4496
Ah28	ATCOPIN_I	LTR Retrotransposon	111
Ah28	ATCOPIN_LTR	LTR Retrotransposon	364
Ah28	ATCOPIN_LTR	LTR Retrotransposon	357
Ah28	ATENSPM10	EnSpm	2904
Ah28	ATENSPM10	EnSpm	2304
Ah28	ATENSPM10	EnSpm	2304
Ah28	ATENSPM10	EnSpm	2904
Ah28	ATHATN1	hAT	480
Ah28	ATHILA4D_LTR	Gypsy	222
Ah28	ATHILA4D_LTR	Gypsy	233
Ah28	ATHILA8ALTR	Gypsy	1550
Ah28	ATHILA8ALTR	Gypsy	1550
Ah28	ATHILA8ALTR	Gypsy	656
Ah28	ATHILA8B_I	Gypsy	540
Ah28	ATLANTYS1_I	Gypsy	3294
Ah28	ATLANTYS2_LTR	Gypsy	427
Ah28	ATLINE1_1	L1	635
Ah28	ATREP1	Helitron	192
Ah28	ATREP10	Helitron	142
Ah28	ATREP11	Helitron	115
Ah28	ATREP11	Helitron	207
Ah28	ATREP11A	Helitron	607
Ah28	ATREP16	DNA transposon	241
Ah28	ATREP2	Helitron	268
Ah28	ATREP3	Helitron	335
Ah28	ATREP3	Helitron	462
Ah28	ATREP5	Helitron	280
Ah28	ATREP5	Helitron	285
Ah28	AtSB3	SINE	301
Ah28	AtSB5	SINE	148
Ah28	AtSB6	SINE	335
Ah28	DT1	Mariner/Tc1	239
Ah28	Sadhu7-1	SINE	484
Ah28	TAT1_ATH	Gypsy	613
Al13	ARNOLDY1	MuDR	443

Al13	ARNOLDY2	MuDR	845
Al13	ATENSPM12	EnSpm	7343
Al13	ATENSPM7	EnSpm	878
Al13	ATHATN1	hAT	514
Al13	ATHATN1	hAT	472
Al13	ATHATN1	hAT	498
Al13	ATHATN2	hAT	569
Al13	ATHPOGON1	Mariner/Tc1	83
Al13	ATIS112A	Harbinger	1774
Al13	ATLINE1_3A	L1	49
Al13	ATMUNX1	MuDR	81
Al13	ATMUNX1	MuDR	228
Al13	ATREP11	Helitron	168
Al13	ATREP11A	Helitron	180
Al13	ATREP11A	Helitron	194
Al13	ATREP16	DNA transposon	183
Al13	ATREP4	Helitron	232
Al13	ATREP4	Helitron	370
Al13	ATREP4	Helitron	247
Al13	ATREP5	Helitron	888
Al13	ATREP9	Helitron	912
Al13	ATTIRTA1	Mariner/Tc1	109
Al13	ATTIRTA1	Mariner/Tc1	252
Al13	ATTIRX1A	DNA transposon	330
Al13	DT1	Mariner/Tc1	269
Al13	SIMPLEHAT2	MSAT	247
Al13	TA11	L2	4782
Al13	TAT1_ATH	Gypsy	423
Al13	VANDAL18NB	MuDR	263
Al39	ARNOLDY2	MuDR	156
Al39	ATGP11_I	Gypsy	1044
Al39	ATGP11_LTR	Gypsy	634
Al39	ATGP4_I	Gypsy	1134
Al39	ATGP6LTR	Gypsy	268
Al39	ATHATN1	hAT	462
Al39	ATHILA4D_LTR	Gypsy	204
Al39	ATHPOGON1	Mariner/Tc1	100
Al39	ATMUN1	MuDR	135
Al39	ATRAN	DNA transposon	61
Al39	ATREP1	Helitron	232
Al39	ATREP11A	Helitron	211
Al39	ATREP15	Helitron	1571
Al39	ATREP18	Interspersed repeat	183
Al39	ATREP3	Helitron	351
Al39	ATREP5	Helitron	239
Al39	AtSB2	SINE	170

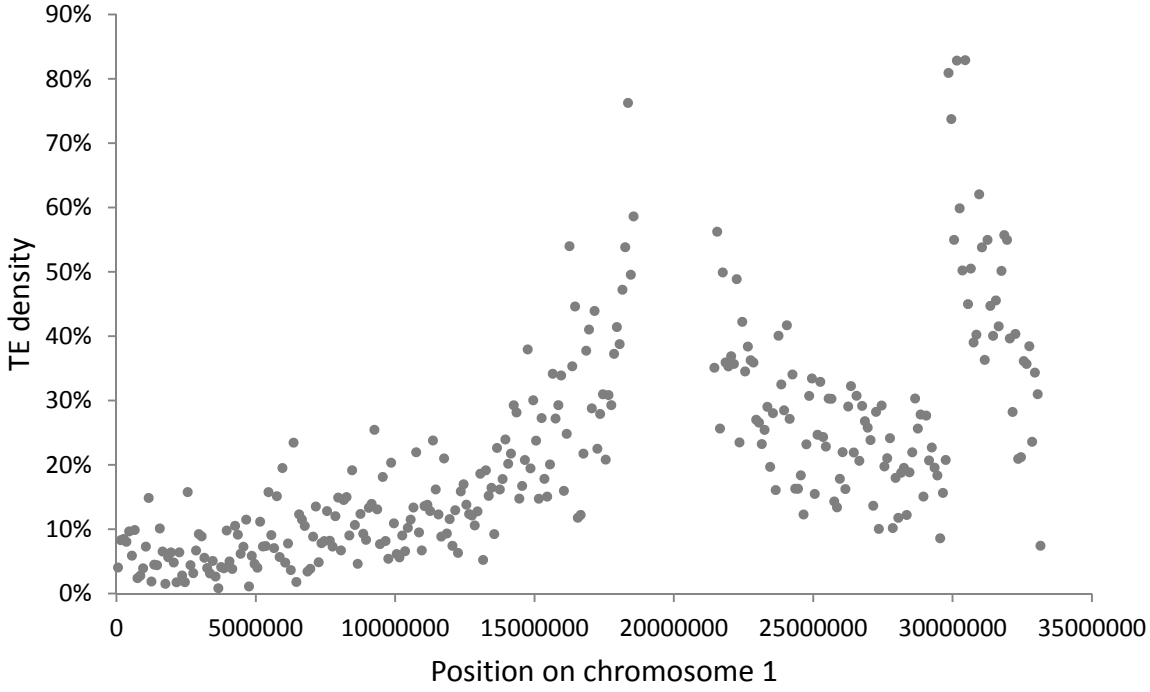
Al39	ATTIR16T3A	Mariner/Tc1	485
Al39	ATTIRTA1	Mariner/Tc1	244
Al39	ATTIRX1A	DNA transposon	368
Al39	ATTIRX1C	DNA transposon	309
Al39	BOMZH1	MuDR	286
Al39	HELITRONY1B	Helitron	139
Al39	HELITRONY1E	Helitron	445
Al39	HELITRONY1E	Helitron	200
Al39	Sadhu4-1	SINE	113
Al39	Sadhu5-1	SINE	347
Al39	VANDAL18NB	MuDR	602
Al39	VANDAL18NB	MuDR	108
Ah13	ATCOPIA26I	Copia	544
Ah13	ATCOPIA26I	Copia	828
Ah13	ATCOPIA26I	Copia	2807
Ah13	ATCOPIA27_I	Copia	4797
Ah13	ATCOPIN_I	LTR Retrotransposon	110
Ah13	ATCOPIN_LTR	LTR Retrotransposon	357
Ah13	ATCOPIN_LTR	LTR Retrotransposon	356
Ah13	ATGP11_I	Gypsy	355
Ah13	ATHAT9	hAT	491
Ah13	ATHATN1	hAT	473
Ah13	ATHATN3A	hAT	160
Ah13	ATHATN6	hAT	211
Ah13	ATHATN6	hAT	94
Ah13	ATHILA4_LTR	Gypsy	307
Ah13	ATHILA4D_LTR	Gypsy	151
Ah13	ATHILA4D_LTR	Gypsy	153
Ah13	ATHILA4D_LTR	Gypsy	1198
Ah13	ATHILA7LTR	Gypsy	966
Ah13	ATHILA8B_I	Gypsy	514
Ah13	ATHPOGON1	Mariner/Tc1	567
Ah13	ATLINE1A	Non LTR retrotransposon	1141
Ah13	ATREP17	DNA transposon	233
Ah13	ATREP17	DNA transposon	188
Ah13	ATREP17	DNA transposon	192
Ah13	ATREP17	DNA transposon	233
Ah13	ATREP3	Helitron	242
Ah13	ATREP3	Helitron	315
Ah13	ATTIR16T3A	Mariner/Tc1	433
Ah13	ATTIR16T3A	Mariner/Tc1	252
Ah13	ATTIR16T3A	Mariner/Tc1	176
Ah13	ATTIRTA1	Mariner/Tc1	262
Ah13	DT1	Mariner/Tc1	270
Ah15	ARNOLDY2	MuDR	185
Ah15	ATCOPIA78_I	Copia	920

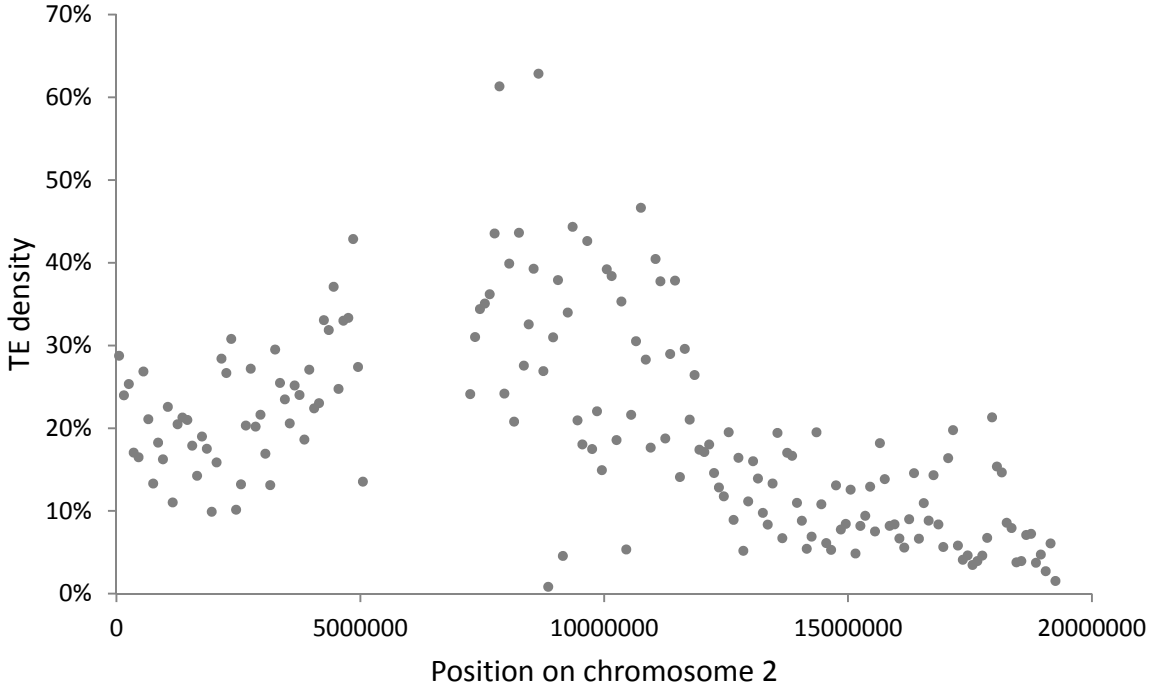
Ah15	ATCOPIA78LTR	Copia	302
Ah15	ATCOPIA78LTR	Copia	133
Ah15	ATDNA2T9A	MuDR	179
Ah15	ATHATN2	hAT	655
Ah15	ATHATN3A	hAT	160
Ah15	ATHATN6	hAT	122
Ah15	ATHILA4B_LTR	Gypsy	147
Ah15	ATHILA4D_LTR	Gypsy	748
Ah15	ATHILA7LTR	Gypsy	328
Ah15	ATHILA7LTR	Gypsy	378
Ah15	ATHILA7LTR	Gypsy	353
Ah15	ATLANTYS1_I	Gypsy	1902
Ah15	ATLANTYS1_I	Gypsy	1903
Ah15	ATLANTYS1_I	Gypsy	1820
Ah15	ATLANTYS2_I	Gypsy	1345
Ah15	ATLANTYS2_I	Gypsy	1153
Ah15	ATLANTYS2_I	Gypsy	1333
Ah15	ATLANTYS2_I	Gypsy	2135
Ah15	ATLANTYS2_I	Gypsy	1317
Ah15	ATLANTYS2_I	Gypsy	5540
Ah15	ATLANTYS2_I	Gypsy	7653
Ah15	ATLANTYS2_I	Gypsy	2135
Ah15	ATLANTYS2_LTR	Gypsy	333
Ah15	ATLANTYS2_LTR	Gypsy	440
Ah15	ATLANTYS2_LTR	Gypsy	88
Ah15	ATLANTYS2_LTR	Gypsy	333
Ah15	ATREP1	Helitron	194
Ah15	ATREP3	Helitron	312
Ah15	AtSB6	SINE	350
Ah15	ATTIRX1B	DNA transposon	139
Ah15	ATTIRX1B	DNA transposon	175
Ah15	ATTIRX1D	DNA transposon	339
Ah15	DT1	Mariner/Tc1	121
Ah15	DT1	Mariner/Tc1	66
Ah15	DT1	Mariner/Tc1	214
Ah15	DT1	Mariner/Tc1	239
Ah15	RP1_AT	DNA transposon	125
Ah15	TAT1_ATH	Gypsy	613
Ah15	TAT1_ATH	Gypsy	613
Ah15	VANDAL21	MuDR	3815
Ah15	VANDAL21	MuDR	2311
Ah20	ATCOPIA61_I	Copia	4665
Ah20	ATCOPIA61LTR	Copia	172
Ah20	ATCOPIA61LTR	Copia	178
Ah20	ATCOPIA96_I	Copia	710
Ah20	ATLANTYS3I	Gypsy	7890

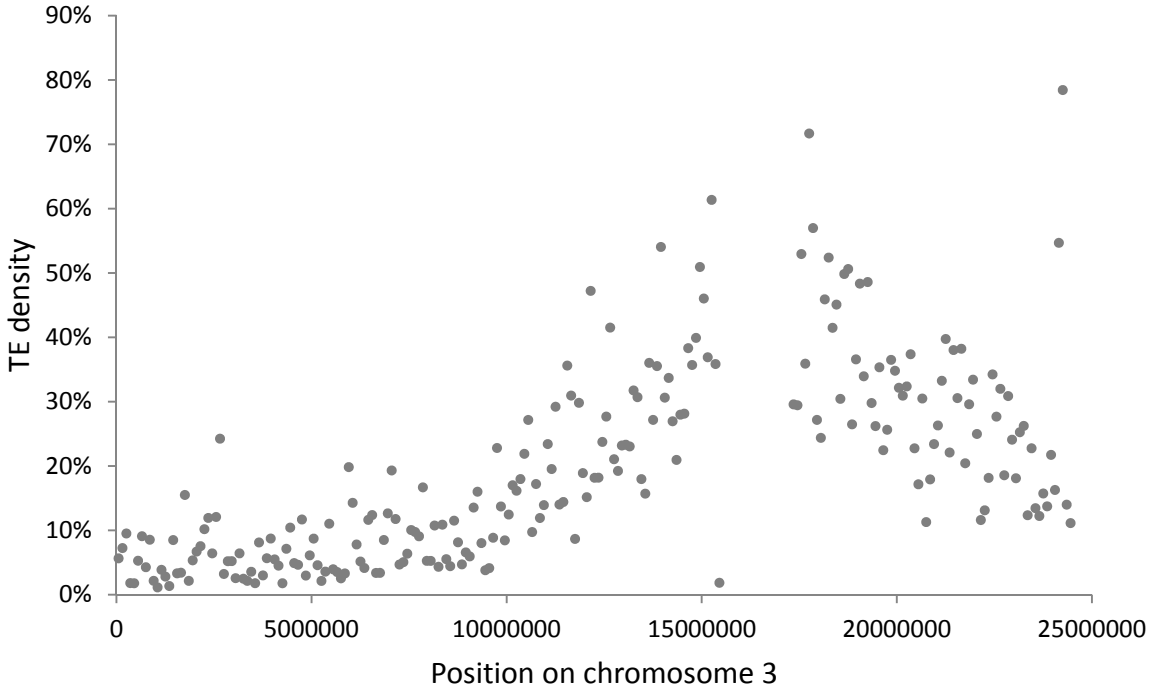
Ah20	ATLANTYS3LTR	Gypsy	493
Ah20	ATLANTYS3LTR	Gypsy	501
Ah20	ATLINE1A	Non LTR retrotransposon	2138
Ah20	ATREP1	Helitron	228
Ah20	ATREP3	Helitron	389
Ah20	ATREP3	Helitron	522
Ah20	ATREP3	Helitron	579
Ah20	ATREP3	Helitron	350
Ah20	AtSB3	SINE	224
Ah20	AtSB6	SINE	344
Ah20	DT1	Mariner/Tc1	104
Ah20	Sadhu1-2	SINE	241
Ah20	VANDAL18NA	MuDR	300
Ah20	VANDAL18NB	MuDR	423
Ah20	VANDAL18NB	MuDR	528
Ah20	VANDAL1N1	MuDR	539
Ah20	VANDAL1N1	MuDR	807
Ah20	VANDAL2N1	MuDR	455
Ah32	ATCOPIA15I	Copia	1422
Ah32	ATCOPIA15I	Copia	1700
Ah32	ATCOPIA17I	Copia	1247
Ah32	ATCOPIA65LTR	Copia	252
Ah32	ATCOPIA72LTR	Copia	75
Ah32	ATCOPIA85_I	Copia	1376
Ah32	ATENSPM1	EnSpm	640
Ah32	ATENSPM1A	EnSpm	716
Ah32	ATENSPM4	EnSpm	1150
Ah32	ATENSPM5	EnSpm	1354
Ah32	ATENSPM6	EnSpm	3909
Ah32	ATHATN1	hAT	279
Ah32	ATHILA4D_LTR	Gypsy	1191
Ah32	ATLINE1A	Non LTR retrotransposon	1465
Ah32	ATLINE1A	Non LTR retrotransposon	1504
Ah32	ATRAN	DNA transposon	366
Ah32	ATREP10D	Helitron	109
Ah32	ATREP13	Helitron	527
Ah32	ATREP15	Helitron	178
Ah32	ATREP3	Helitron	262
Ah32	ATREP3	Helitron	340
Ah32	ATREP3	Helitron	494
Ah32	ATREP3	Helitron	295
Ah32	ATREP4	Helitron	291
Ah32	AtSB2	SINE	97
Ah32	AtSB5	SINE	99
Ah32	AtSB6	SINE	336
Ah32	BRODYAGA1A	DNA transposon	115

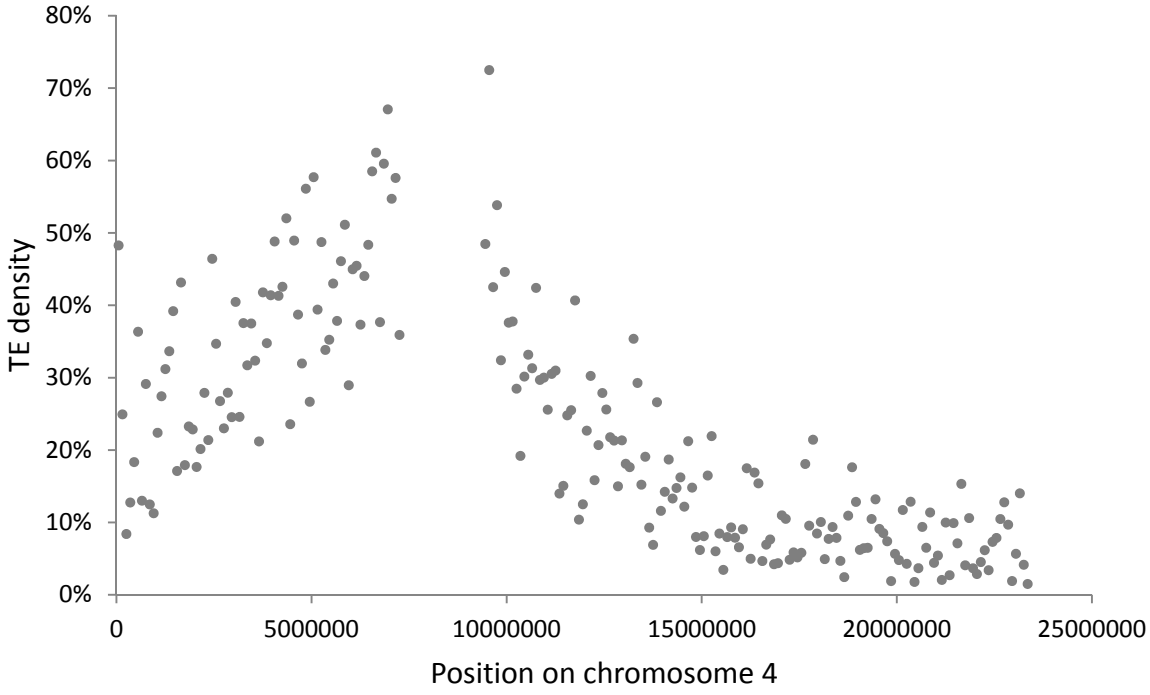
Ah32	DT1	Mariner/Tc1	236
Ah32	DT1	Mariner/Tc1	143
Ah32	HELITRONY3	Helitron	661
Ah32	VANDAL1N1	MuDR	268
Ah32	VANDAL2N1	MuDR	1228
Ah43	ATCOPIA27_I	Copia	3399
Ah43	ATCOPIA40_LTR	Copia	181
Ah43	ATCOPIA58LTR	Copia	312
Ah43	ATCOPIA65_I	Copia	4093
Ah43	ATGP11_I	Gypsy	943
Ah43	ATGP11_I	Gypsy	953
Ah43	ATGP6LTR	Gypsy	819
Ah43	ATGP6LTR	Gypsy	819
Ah43	ATHILA4A_I	Gypsy	363
Ah43	ATHILA4A_I	Gypsy	830
Ah43	ATHILA4A_I	Gypsy	369
Ah43	ATHILA4A_I	Gypsy	525
Ah43	ATHILA4D_LTR	Gypsy	1195
Ah43	ATHILA4D_LTR	Gypsy	1175
Ah43	ATHPOGON1	Mariner/Tc1	551
Ah43	ATLANTYS1_I	Gypsy	1704
Ah43	ATLANTYS2_I	Gypsy	3206
Ah43	ATREP3	Helitron	281
Ah43	AtSB3	SINE	278
Ah43	AtSB6	SINE	347
Ah43	AtSB6	SINE	345
Ah43	ATTIRX1C	DNA transposon	394
Ah43	CASTOR_I	Copia	980
Ah43	ENDOVIR1_I	Copia	1585
Ah43	HELITRONY1A	Helitron	154
Ah43	IID2-12_AT	Interspersed repeat	396
Ah43	TA11	L1	5452
Ah43	TAT1_ATH	Gypsy	613

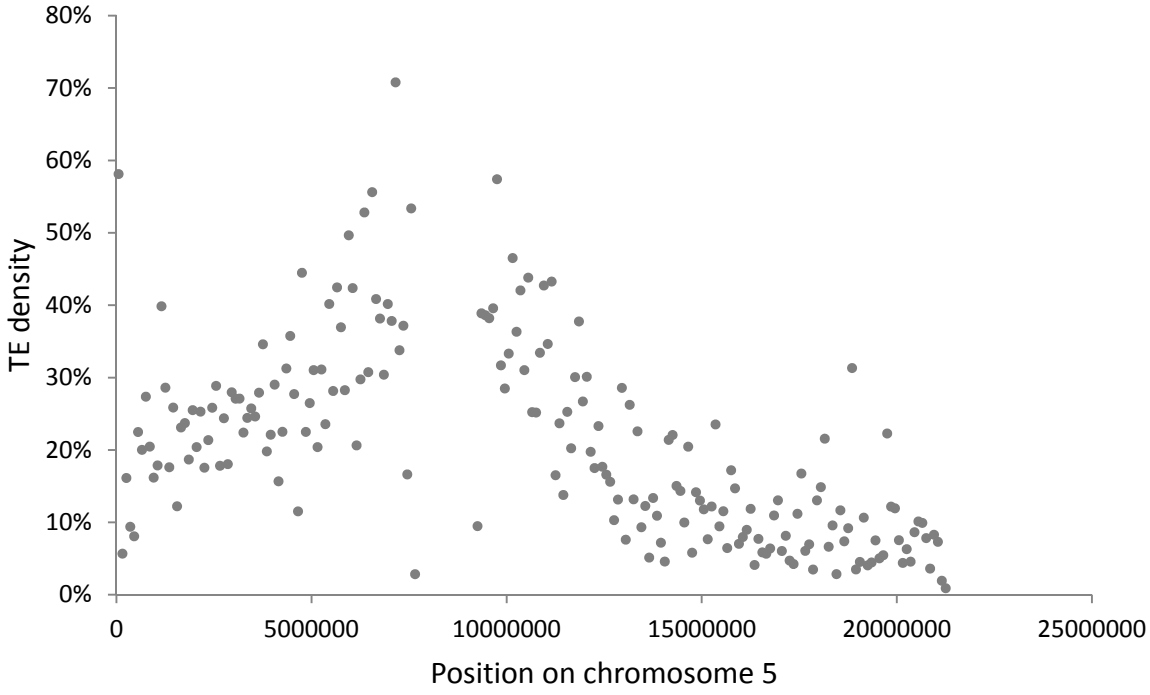
Supplementary figure 7. TE density along *A. lyrata* chromosomes 1 to 6 and chromosome 8. Transposable elements contents were calculated using CENSOR (KOHANY *et al.* 2006) for non overlapping windows of 100 kb.

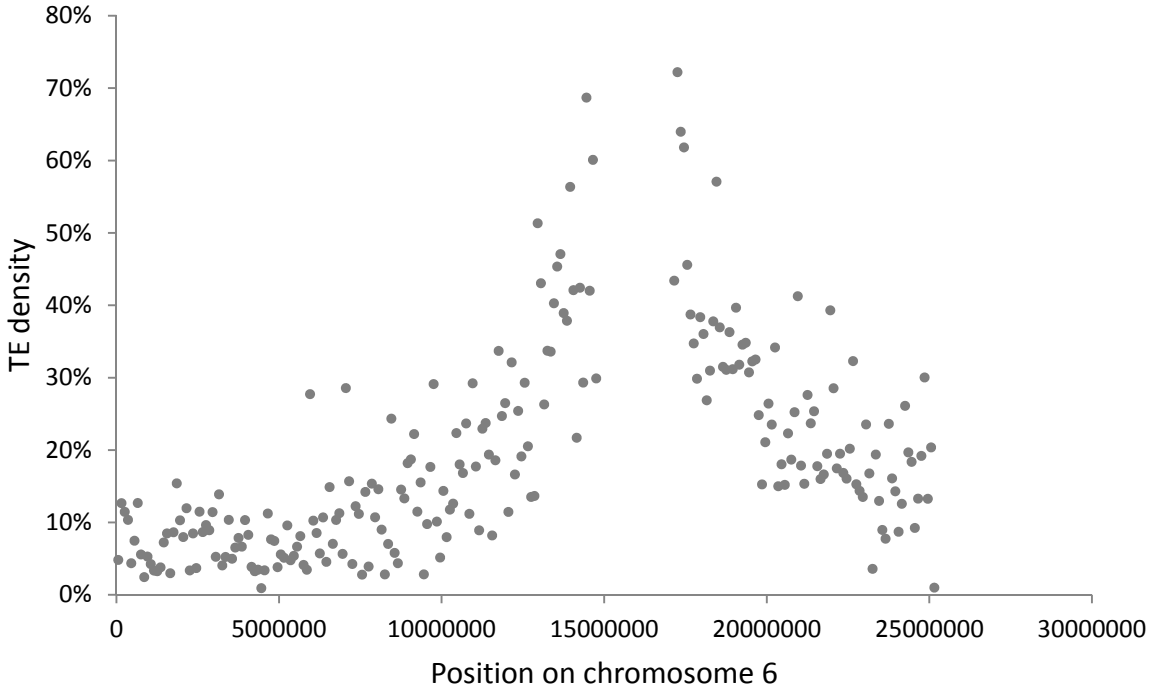


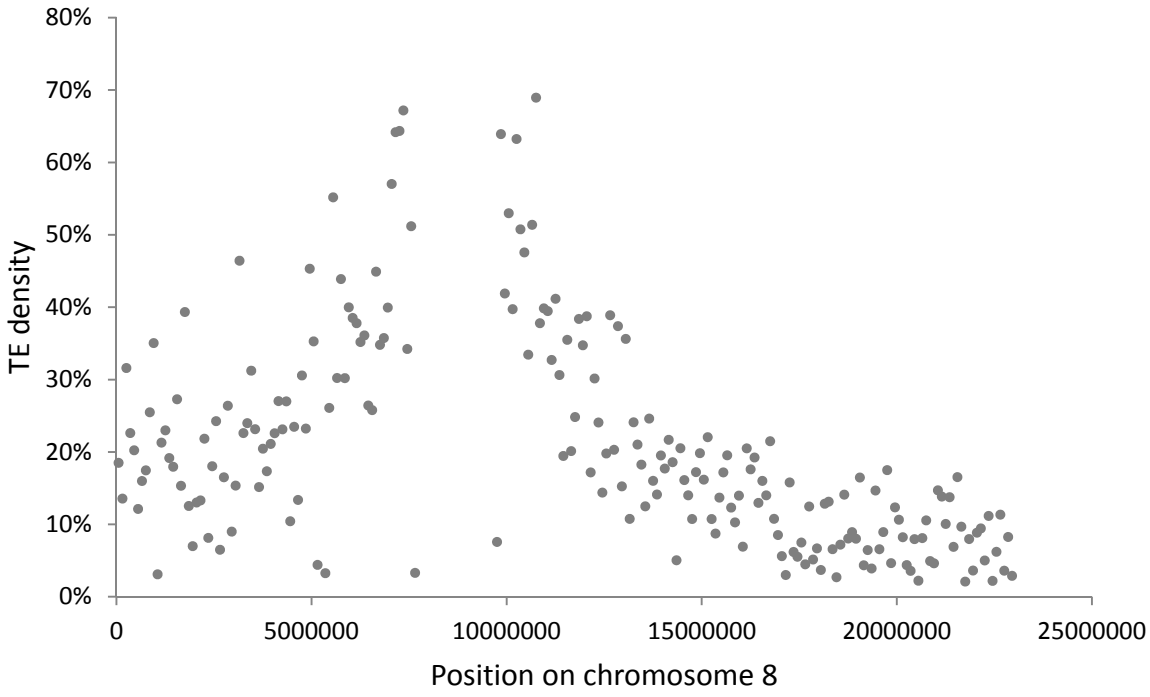










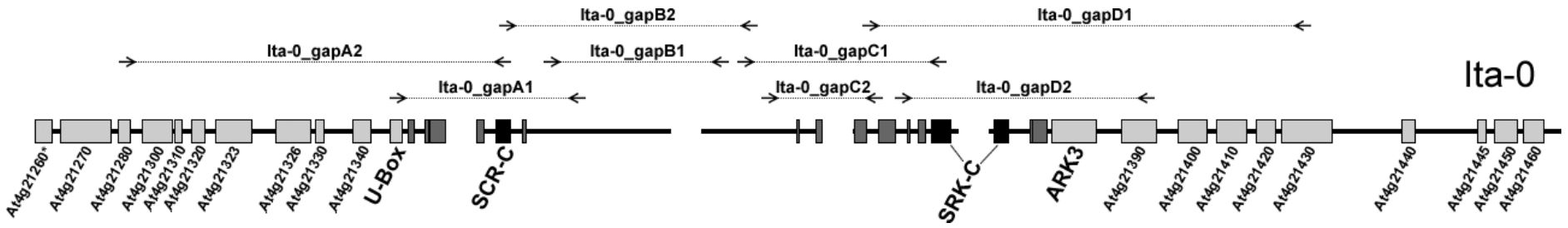


Supplementary table 3. Primers pairs used to validate BAC clones by PCR amplification. These primers were defined to amplify *SRK* (primer Sh04), *U-box* (primer B80) and *ARK3* (primer ARK3) genes.

Name	Forward primer	Reverse primer	Product size
Sh04	ACGCCGTACAGTTACAATGT	GACAATAAATACCGTAGACG	373
B80	TGGRTCRTATCCTGATGCAA	TCCGCTACAACAACACAAGC	348
ARK3	TTGTGCGGTTGAAGAAGATG	GGAGAAGGAACTAACCGAKA	255

Supplementary table 4. Primers used to confirmed the gaps in the S-locus sequence of Ita-0, *via* long range PCR.

Name	Forward primer	Reverse primer	Product size
Ita-0_gapA1	AGCAAAGACAGAGACTTTTCCTGTA	TAATGTGTGCCTACAAGAGTTGGTA	13675
Ita-0_gapA2	TAACTCAGACCAAGCAAGAGACTTT	ATTTGCTAAAACATACGGGATATGA	29377
Ita-0_gapB1	TGATGAAGATATTTGGACTCTAGCC	TTAATCGTATAAAAATGGGAGACCA	12868
Ita-0_gapB2	AAATCATATCCCGTATGTTTTAGCA	CTATTTGTTGCGATTTACTTGGTCT	18904
Ita-0_gapC1	AGACCAAGTAAATCGCAACAAATAG	ACAGTGTATCAAGAGAAAACCCAAG	12686
Ita-0_gapC2	ATGTGCTTGAGTATGATGACGTAAA	AAGTGACCAAGTCTTACATTGAAGG	4882
Ita-0_gapD1	ACAGTCTACAAACGGTCTAAAATGC	TTGACAAGCAAGTTAGAGATCAGTG	33655
Ita-0_gapD2	CTTCATCGACATTCTGTTACATCAC	ATTACTCATTAGGGGTTGATCCTTC	18567



Supplementary figure 8. Localization of primers used for gaps confirmation in the Ita-0 sequence.

Supplementary table 5. Detected haplotypes in the 107 accessions.

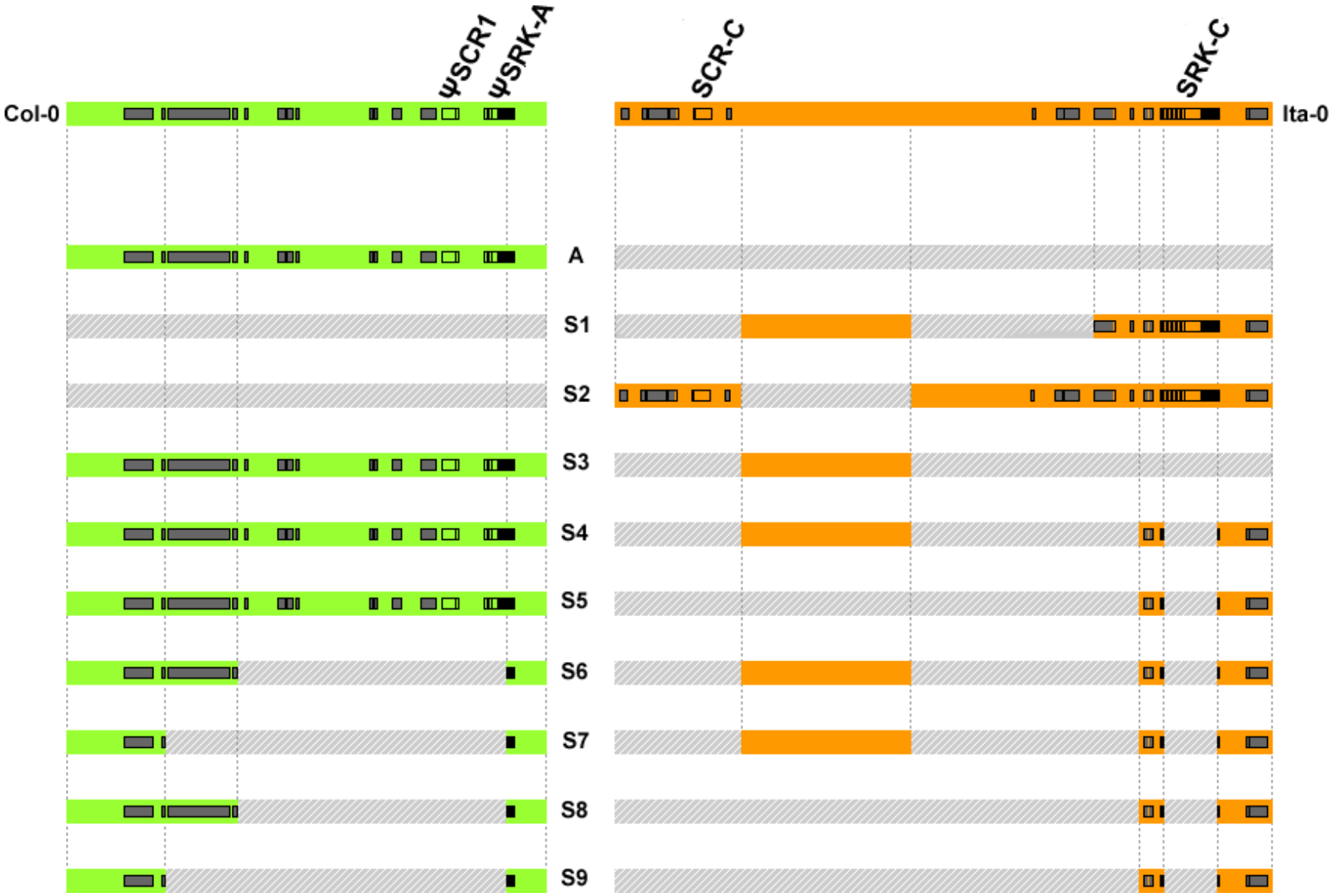
Name	Haplotype	Name	Haplotype	Name	Haplotype
algustrum	*A*	lis_2	*R1*	t540	*A*
app1_12	*R3*	lis_3	*A*	t550	*A*
app1_14	*R1*	lom1_1	*R1*	t570	*C2*
ba1_2	*R1*	lov_1	*R1*	t710	*A*
ba3_3	*R1*	lov_5	*R1*	t720	*A*
ba4_1	*R2*	lu_1	*R1*	t740	*A*
bil_5	*R2*	lund	*A*	t790	*A*
bil_7	*R2*	nyl_2	*R2*	t800	*A*
bro1_6	*R1*	nyl_7	*R2*	t840	*R1*
bur_0	*C2*	omn_1	*A*	t850	*A*
dor_10	*R2*	omn_5	*A*	t880	*R1*
dra2_1	*R1*	omo1_7	*A*	t900	*A*
dra3_1	*R1*	omo2_1	*R1*	t930	*A*
eden_1	*R2*	or_1	*R2*	t960	*R1*
eden_2	*A*	rev_1	*A*	t980	*A*
eden_7	*R2*	rev_2	*R3*	taa_04	*A*
eden_9	*A*	san_2	*A*	taa_14	*A*
eds_1	*A*	sanna_2	*A*	tad_01	*A*
fab_2	*A*	sparta_1	*R2*	tad_04	*A*
fab_4	*R1*	spr1_2	*A*	tad_06	*A*
fja1_1	*R2*	spr1_6	*A*	tal_07	*A*
fja1_2	*R2*	sr_3	*A*	tamm_2	*A*
fja1_5	*R2*	sr_5	*R2*	tdr_1	*A*
fly2_2	*A*	st_0	*A*	tdr_16	*A*
gron_5	*A*	stu1_1	*R1*	tdr_17	*A*
hov1_10	*A*	t1000	*C2*	tdr_2	*R1*
hov1_7	*A*	t1070	*R1*	teden_03	*R3*
hov3_2	*R3*	t1080	*R3*	tottarp_2	*A*
hov3_5	*A*	t1090	*A*	ull2_3	*R3*
hov4_1	*C2*	t1110	*A*	ull2_5	*A*
hovdala_2	*A*	t1130	*A*	var2_1	*A*
kavlinge_1	*A*	t1160	*A*	var2_6	*A*
kni_1	*A*	t460	*A*	vastervik	*A*
kulturen_1	*C2*	t470	*C2*	vimmerb	*R2*
liarum	*R1*	t480	*C3*	vinslov	*R1*
lillo_1	*R2*	t530	*A*		

Supplementary results

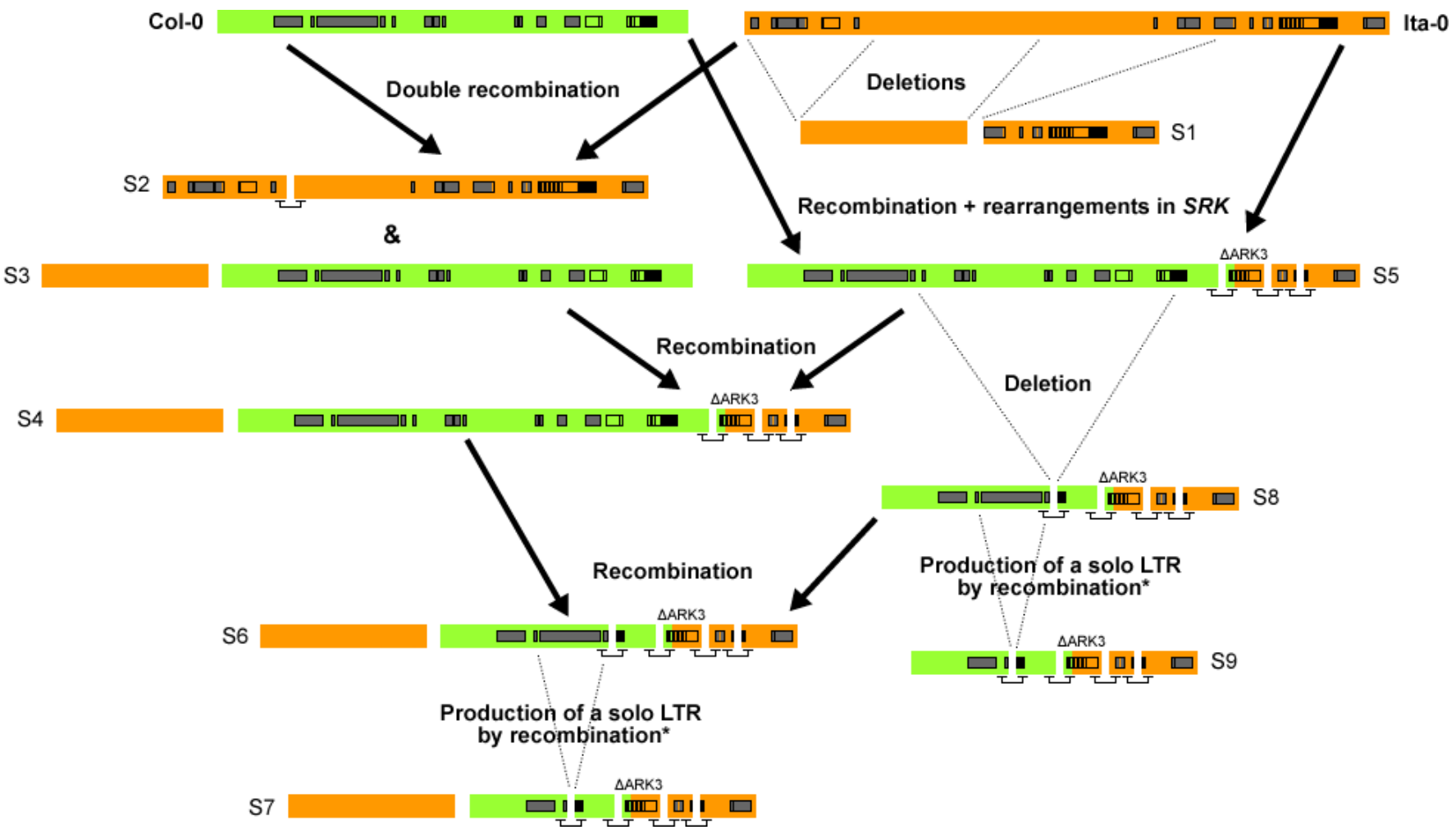
In a first analysis, ten distinct patterns were detected with respect to the fragments of Ita-0 or Col-0 they presented (Figure S2). Some analyzed accessions were similar to Col-0. Some others presented fragments from Ita-0 (haplotypes *S1* and *S2*) but none presented its complete sequence. All other accessions were detected as recombinants between haplogroups *A* and *C*, with both fragments from Ita-0 and Col-0.

From this point, a scenario for the generation of the haplotypes could be drawn (Figure S3). It is based on the reference sequences Col-0 and Ita-0 and assumes six recombination events, including two illegitimate recombinations between the LTRs of the Atcopia10 retrotransposon, and three deletion events. According to this scenario, Col-0 and Ita-0 would have recombined twice independently, in order to generate haplotypes *S2* and *S3* on the one hand, and haplotype *S5* on the other hand. This latter would then have recombined with haplotype *S3* to generate haplotype *S4* and would have produced haplotype *S8* as a result of a large deletion. Haplotype *S6* would have been generated by a recombination event between haplotypes *S8* and *S3*. Finally, haplotypes *S7* and *S9* would have independently been produced through an illegitimate recombination event between the two LTRs of the Atcopia10 retrotransposon.

Even if it allows to explain the formation of all haplotypes, the proposed scenario does not seem to be parsimonious. Indeed, it implies an important number of recombination events. In particular, it assumes a double recombination, from which the two daughter chromatides were maintained, and an illegitimate recombination between the two LTRs of Atcopia10, which would have occurred twice independently. Moreover, in contrast to all other fragments detected in the accessions with respect to the reference sequences, a 12 kb-fragment of Ita-0 could not be orientated as compared to others. We extracted reads, located at the extremities of this fragment and with one mate unmapped, from the BAM files of several accessions. Blast of unmapped reads revealed for the two extremities a strong homology with one particular position in the chromosome 2 (around position 1 825 600) of the reference genome of *A. thaliana*. This observation suggests the hypothesis that the concerned fragment is polymorphic, being present in chromosome 2 of some accessions (Haplotypes *S1, S3, S4, S6, S7*), and would have been inserted in the S-locus of the Ita-0 accession. Consequently, we decided to ignore its presence/absence in the analyzed accessions, and built a scenario without taking it into account.



Supplementary figure 9. Annotation of the reference sequences Col-0 (green) and Ita-0 (orange), and fragments detected in the ten types of accessions. Note that the relative position of the fragments is not respected in this representation.



Supplementary figure 10. Scenario for the generation of the different detected haplotypes. Asterisks indicate similar events thought to have occurred independently. Fragments, which were confirmed to be consecutive, are indicated by brackets. Fragments, which were confirmed to be consecutive by the identification of mates mapping to different fragments, are indicated by brackets.

Annexe 1. Liste détaillée des sites détectés sous corrélation et sous compensation dans le cas de l'analyse par paires et de l'analyse par clusters. Cette annexe comprend les 6 parties suivantes :

A. Alignement des protéines SCR ; B. Alignement des protéines SRK.

Ces alignements ont été réalisés en utilisant le programme Muscle (EDGAR 2004) implémenté sur Jalview (WATERHOUSE *et al.* 2009), puis ajustés à la main.

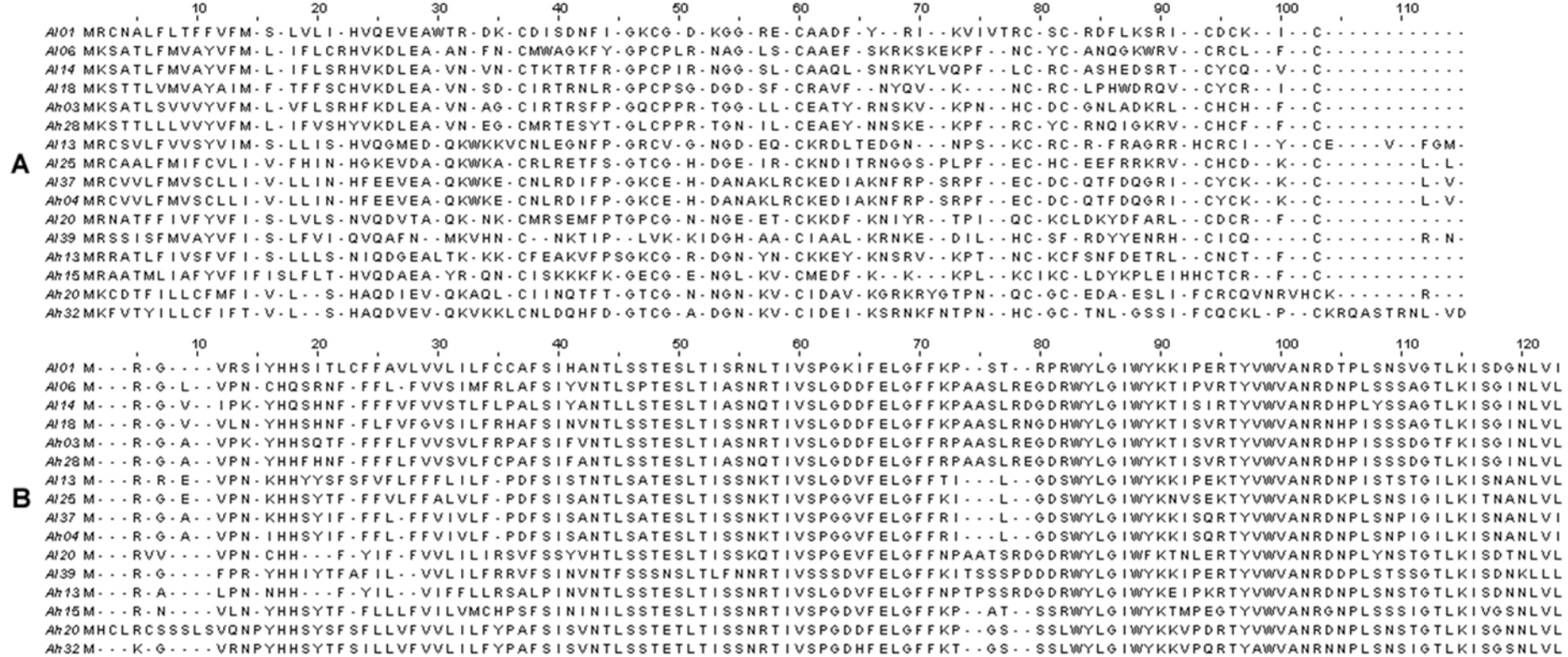
C. Liste des sites détectés sous corrélation dans le cas de l'analyse par paires.

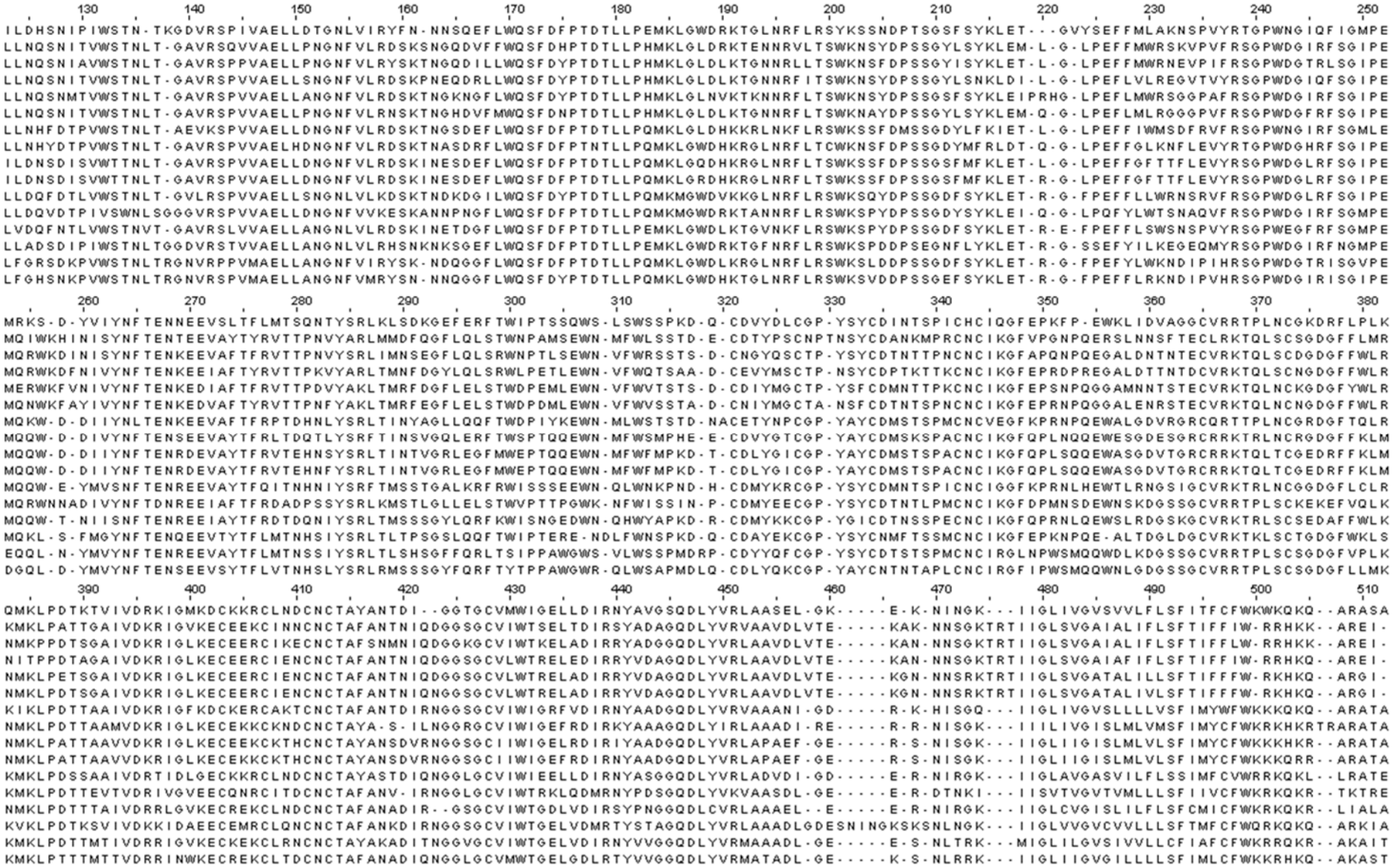
D. Liste des sites détectés sous compensation dans le cas de l'analyse par paires.

E. Liste des sites détectés sous corrélation dans le cas de l'analyse par clusters.

F. Liste des sites détectés sous compensation dans le cas de l'analyse par clusters.

Les parties C à F comprennent des tableaux regroupant pour chaque groupe de sites détecté sous coévolution les positions sur l'alignement de SCR et SRK, la statistique de coévolution calculée par comparaison des cartes de substitution, le nombre de sites détectés dans le groupe concerné, la p-value calculée par rapport à 100 000 simulations, et la propriété biochimique prise en compte dans la comparaison des cartes de substitution (« simple » pour les cartes attribuant le même poids à toutes les substitutions). La colonne FDR indique quant à elle si le groupe reste significatif après correction pour tests multiples.





```
         520       530       540       550       560       570       580       590       600       610       620       630       640
APNVNPERSPD-ILMDGMVIPSDIH---LST-ENIT-D-DLLLPSTDFEVIVRATNNFSVSNKLGEGGFGIVYKGRLHNGKEFAVKRLSDLSHQGSDEFKTEVKVISRLQHINLVRILGCCASGKEKMLI
A-Q-YTECGQ----R---V---GRQ--NLLD-TD-E-D-DLKLPLMEYDVVAMATDDFSITNKLGEGGFGTVYKGRLIDGEEIAVKKLSDVSTQGTNEFRTEMILIAKLQHINLVRLLGCFADADDKILV
A-Q-YTECGQ----R---V---GRQ--NLLE-TD-E-D-DLKLPLMEYDVVAMATDDFAITNKLGEGGFGTVYKGRLIDGEEIAVKKLSDVSTQGTNEFRTEMILIAKLQHINLVRLLGCFADADDKILV
A-Q-YTECGQ----R---V---GRQ--NLLD-TD-E-D-DLKLPLMEYDVVAMATDDFAITNKLGEGGFGTVYKGRLIDGEEIAVKKLSDVSTQGTNEFRTEMILIAKLQHINLVRLLGCFADADDKILV
A-L-YTECGQ----T---G---GRL--NLLDTTD-D-D-DLKLPLMEYDVVAMATNDFSISNKLGEGGFGTVYKGRLIDGEEIAVKKLSDVSTQGTNEFRTEMILIAKLQHINLVRLLGCFADEDDKILV
A-L-YTECGQ----T---G---GRL--NLLDTTD-D-D-DLKLPLMEYDVVAMATNDFSISNKLGEGGFGTVYKGRLIDGEEIAVKKLSDVSTQGTNEFRTEMILIAKLQHINLVRLLGCFADEDDKILV
APNVYRERTQH--LTNGVVISSGRH---LFG-ENKT-E-ELELPLTEFEAVVMATDNFSDSNILGQGGFGVVYMGRLPDGQEIAVKRLSMVSLQGVNEFKNEVKLIARLQHINLVRLFSCCIYADEKILI
STI---ERIQG-FLTNGYQVVSRRR--QLFE-ENKI-E-DLELPLTEFEAVVIATGNFSESNILGRGGFGMVYKGRLPDGQDTAVKRLSEVSAQGTTEFMNEVRLIARLQHINLVRLLSCCIYADEKILI
APIGYRDRIQESIITNGVVMSSGRR---LLG-EK---E-DLELPLTEFETVVMATDNFSDSDILGQGGFGIVYKGRLLDGQEIAVKRLSEMSSQGTNEFKNEVRLIARLQHINLVRLLSCCIYADEKILI
APIGYRDRIQESIITNGVVMSSGRR---LLG-EK---E-DLELPLTEFETVVMATDNFSDSNILGRGGFGIVYKGRLLDGQEIAVKRLSEMSSQGTNEFKNEVRLIARLQHINLVRLLSCCIYADEKILI
APIVYPTINQG-LLMNRLEISSGRH---LSE-DNQT-E-DLELPLVEFEAVVMATENFSNSNKLGEGGFGVVYKGRLLDGQEIAVKRLSTTSIQGICEFRNEVKLISKLQHINLVRLFGCCVDENEKMLI
API---ERNQN-LLMNGVVISSGRHLSDLSE-ESIA-E-DLELPLMEFNAIAMATKNFSDCNKLGEGGFGSVYKGRLANGQDIAVKRLSEMSHQGTNEFKNEVKLIARLQHINLVRILGCCFNAKEKVLI
APIVYHERNAE-LLMNGMVISSRRR---LSG-ENIT-E-DLELPLVELDAVVMATENFSNANKVGQGGFGIVYKGRLLDGQEIAVKRLSKTSLQGTNEFKNEVRLIAKLQHINLVRLLGCCVEVDEKMLI
APIGYQERNKD-LVINGIIMSSRRH---FTE-EFYT-E-DLELPLMEFEAVVIATNNFSDSNKLGQGGFGIVYKGTLLDGREIAVKRLSKMSLQGITEFKNEVKLIARLQHINLVRLIGCCVDMDEKILI
TPLVYQRRNQD-LLMNSVVIPSRRH---LSE-ESTE-E-DLELPLMDFKAVVMATSNFCESNKLGEGGFGIVYKGRLLDGQEIAVKRLSEMSHQGTREFKNEVNLIARLQHINLVRILGCCFDAKEKMLI
SPIVYQERNQD-LLTNRVVISSRKH---LFE-ENITEEDDLELPLMEFKAVAMATENFSHRNKLGEGGFGIVYKGRLINGQEIAVKRLSEMSHQGAREFRNEVKLIARLQHINLVRILGCCFDAKEKMLI
         650       660       670       680       690       700       710       720       730       740       750       760       770
YEYLENSSLDRHLFDKTRSS-NLNWQRRFDITNGIARGILYLHHDSRCRIIHRDLKASNILLDKNMIPKISDFGMARIFSDDVNEAITRRIVGTYGYMSPEYAMDGIYSEKSDVFSFGVMLLEIVTGMKN
YEYLENLSLDYYIFDETKSS-ELNWQTRFNIINGIARGLLYLHKDSRCKVIHRDLKTSNILLDKYMIPKISDFGLARIFARDEEEATTRRIVGTYGYMAPEYAMDGVYSEKSDVFSFGVVILEIVTGKKN
YEYLENLSLDYYIFDETKSS-DLNWQTRFNIINGIARGLLYLHKDSRCKVIHRDLKTSNILLDKDMIPKISDFGLARIFARDEEEATTRRIVGTYGYMAPEYAMDGVYSEKSDVFSFGVVILEIVTGKKN
YEYLENLSLDYYIFDETKSS-ELNWQTRFNIINGIARGLLYLHKDSRCKVIHRDLKTSNILLDKDMIPKISDFGLARIFARDEEEATTRRIVGTYGYMAPEYAMDGVYSEKSDVFSFGVVILEIVTGKKN
YEYLENLSLDYYIFDETKSS-ELNWQTRFNIINGIARGLLYLHKDSRCKVIHRDLKTSNILLDKDMIPKISDFGLARIFARDEEEATTRRIVGTYGYMAPEYAMDGVYSEKSDVFSFGVVILEIVTGKKN
YEYLENLSLDYYIFDETKSS-ELNWQTRFNIINGIARGLLYLHKDSRCKVIHRDLKTSNILLDKDMIPKISDFGLARIFARDEEEATTRRIVGTYGYMAPEYAMDGVYSEKSDVFSFGVVILEIVTGKKN
YEYLENGSLDSHLFKKVQSS-KLNWQKRFNIINGIARGLLYLHQDSRFKIIHRDLKASNVLLDKDMTPKISDFGMARIFEREETEASTKKVVGTYGYMSPEYAMDGIFSVKSDVFSFGVLVLEIVSGKRN
YEYLENGSLDSHLFKINQSS-KLNWQKRFNIINGIARGLLYLHQDSRFKIIHRDLKASNVLLDKNMTPKISDFGMARIFERDETEANTRKVVGTYGYMSPEYAMDGIFSVKSDVFSFGVLVLEIISGKRN
YEYLENGSLDSHLFETTQSSNKLNWQTRFNIINGIARGLLYLHQDSRFKIIHRDMKASNVLLDKNMTPKISDFGMARIFERDETEANTRKVVGTYGYMSPEYAMEGIFSVKSDVFSFGVLVLEIVSGKRN
YEYLENGSLDSHLFETTQSSNKLNWQTRFNIINGIARGLLYLHQDSRFKIIHRDMKASNVLLDKNMTPKISDFGMARIFERDETEANTRKVVGTYGYMSPEYAMEGIFSVKSDVFSFGVLVLEIVSGKRN
YEYLENLSLDSHLFNKSLSC-KLNWQMRFDITNGIARGLLYLHQDSRFRIIHRDLKASNVLLDKDMTPKISDFGMARIFGRDETEANTRKVVGTYGYMSPEYAMDGIFSVKSDVFSFGVLVLEIVSGKKN
YEYLENLSLDRYLFDKTRSS-KLNWQKRFDIANGIARGILYLHQDSRCKIIHRDLKASNVLLDKDMTPKISDFGMARIFARDDGEAITRRIVGTYGYMSPEYAMEGIFSVKSDVFSFGVLLLEIVSGKRN
YEYLENLSLDSYIFDKNRSW-KLNWQMRFNITNGIARGLLYLHQDSRCRIIHRDLKASNVLLDKDMTPKISDFGMARIFGREETEANTKKVVGTYGYMSPEYAMDGVFSMKSDVFSFGVLLLEIISGKRN
YEYLENRDLDSHIFDKTESH-KLNWQKRFDITNGIARGLLYLHQDSRFRIIHRDLKASNVLLDKDMTPKISDFGMARIFGRDETEVNTMKVVGTYGYMSPEYAMDGIYSIKSDVFSFGVLLLEIVSGKKN
YEYLENLSLDTYLFDKTQSS-KLNWQKRFEIINGIARGLLYLHQDSRCMIIHRDLKASNVLLDKDMIPKISDFGMARIFARDEEEATTRRIVGTYGYMSPEYAMDGIYSVKSDVFSFGVLVLEIVSGQRN
YEYLENLSLDTYLFDQTQSS-KLNWQKRFDITNGIARGLLYLHQDSRCRIIHRDLKASNILLDKDMIPKISDFGMARIFGRDEEEATTRRIVGTYGYMSPEYAMDGIFSIKSDVFSFGVLVLEIVCGKRN
         780       790       800       810       820       830       840       850       860       870       880       890       900
RGFFNSDLDSNLLSYVWRNMEEEKGLAVADPNIID-SSSLS--PT-FRPDEVLRCIKIALLCVQEYAEDRPTMLSVVSMLGSETAEIPKAKAPGYCVGRSLHDTDFSSS--L---TWTF-GFAFSDIEPR
RGFTSSDLDTNLLSYVWRNMEEGTGYKLLDPNMMD-SSS-Q---A-FKLDEILRCITIGLTCVQEYAEDRPMMSWVVSMLGSNT-DIPKPKPPGYCLAIS---SDPWT-------STTI-EYTTTEVEPR
RGFTSSDLDTNLLSYVWRNMEEGTGYKLLDPNMMD-SSS-Q---A-FKLDEILRCITIGLTCVQEYAEDRPMMSWVVSMLGSNT-DIPKPKPPGYCLAIS---SDPWT-------STTI-EYTTTEVEPR
RGFTSSDLDTNLLSYVWRNMEEGTGYKLLDPNMMD-SSS-Q---A-FKLDEILRCITIGLTCVQEYAEDRPMMSWVVSMLGSDT-DIPKPKPPGYCLAIS---SDPWT-------STTI-EYTTTEVEPR
RGFTSSDLDTNLLSYVWRNMEEGTGHKLLDPNMMD-SSS-Q---T-FKLNEILRCIAIGLTCVQEYAEDRPMMSWVVSMLGSDT-DIPKPKPPGYCLAIS---SDPWT-------STTI-EYTTTEVEPR
RGFTSSDLDTNLLSYVWRNMEEGTGYKLLDPNMID-SSS-Q---A-FKLDEILRCITIGLTCVQEYAEDRPMMSWVVSMLGSDT-DIPKPKPPGYCLAIS---SDPWT-------STTI-EYTTTEVEPR
RGFYNSNQDNNLLSYTWDHWKEGKWLEIADPIIVG-TSS-S--SSTFRPHEVLRCLQIGLLCVQERAEDRPKMSSVVFMLGNEKGEIPQPKPPGYCIGRSFLETDSSSSTQR-NESSTINQFTVSVINAR
RGFYNSNQDNNLLSYTWDNWKEGEGLKIVDPIIID-SSS-S--FSMFRPYEVLRCIQIGLLCVQERAEDRPKMSSVVLMLGSEKGDIPQPKPPGYCVGRSSLETDSSSSTQRGDESLTVNQITLSVINGR
RGFHNSGQDNNLLGYTWENWKEGKGLEIVDSIIVD-SSS-S--MSLFRPHEVLRCIQIGLLCVQERAEDRPKMSSVVLMLGSEKGEIPQPKRPGY--------------------------------
RGFHNSGQDNNLLGYTWENWKEGKGLEIVDSIIVD-SSS-S--MSLFRPHEVLRCIQIGLLCVQERAEDRPKMSSVVLMLGSEKGEIPQPKRPGYCVGRSSLDTDSSS-------------------
RGFYNSNQDNNLLGYAWRNWKEGKGLEILDPFIVD-SSS-S--PSAFRPHEVLRCIQIGLLCVQERAEDRPVMSSVVVMLRSETETIPQPKPPGYCVGRSPFETDSSTHEQR-DESCTVNQITISAIDPR
RGFDNSNHDSNLLGYTWRNWKEGKGLDILDPIIVD-PSSSS--P--LRTQEILRCIQIGLLCVQELSEDRPTMSSVILMLGSETALIPEPKPPGYWVGRSP-ETFSSSSKQQDDESFTVNQITLSIIDAR
KGFYNSDNDLNLLGCVWRNWTEGKGLEIVDPIILE-SSS-S--TV-I-LQEILKCMQIGLLCVQERAEDRPRMSSVVAMLGSETAVVPQPKLPGYCVGRSPLETDSSRSKQHDDESWTVNEITLSVIDAR
RGFYNSDHDNNLLGYTWRNWKEGKGLEIVDPVMIDVSSSLSPPPSSFRPHEVLRCIQIGLLCVQERAQDRPTMLSVVSMLGSETFVIPQPKPPGYCVGRSPLDTDPSSSEQTDEESWTINQVTISVIDPR
RGFYISNEENNLLGYVWRNWKEGKGPEIIDPTVVN-SSSSS--PSTFRPEDVLRCIQIGLLCVQEHAEDRPLMSSVVLMLGSETSTIPQPKQPGYCVGRSPLEHD-D-------ESRTVNQYTASVIEGR
RGFYNSSQENNLLSYAWRNWKEGKGLEIVDPIIID-SSSSS--VSTFRPHDVLKCIQIGLLCVQEFAEDRPSMSLVVLMLGSETATIPQPKPPGFCVGRSPFEHD-G-------DSWTVNQITASVIEAR
```

C

Position sur SRK	Position sur SCR	Statistique de coévolution	Taille	p-value	Propriété biochimique	FDR
13	90	0,834575	2	0,001363	Simple	non
25	16	0,732759	2	0,005167	Simple	non
25	51	0,703114	2	0,007700	Simple	non
26	53	0,889971	2	0,007361	Volume	non
29	13	0,80198	2	0,002938	Simple	non
31	32	0,964133	2	0,002239	Volume	non
31	32	0,962306	2	0,000896	Grantham	non
32	78	0,900552	2	0,008129	Grantham	non
32	78	0,971361	2	0,005368	Polarité	non
36	21	0,803428	2	0,006117	Grantham	non
40	28	0,967827	2	0,004874	Polarité	non
40	69	0,925318	2	0,009355	Polarité	non
45	48	0,999997	2	0,005679	Simple	non
45	48	0,999997	2	0,004915	Volume	non
45	82	0,999998	2	0,005467	Simple	non
45	82	0,999998	2	0,004584	Volume	non
46	78	0,830348	2	0,004739	Grantham	non
46	78	0,970552	2	0,003540	Polarité	non
48	48	0,999999	2	0,004769	Grantham	non
48	48	0,999999	2	0,005127	Simple	non
48	48	0,999999	2	0,003456	Volume	non
48	82	0,999999	2	0,004769	Grantham	non
48	82	0,999999	2	0,005127	Simple	non
48	82	0,999999	2	0,003456	Volume	non
49	48	0,999996	2	0,005426	Grantham	non
49	48	0,999997	2	0,005676	Simple	non
49	48	0,999996	2	0,004321	Volume	non
49	82	0,999996	2	0,005426	Grantham	non
49	82	0,999997	2	0,005676	Simple	non
49	82	0,999996	2	0,004321	Volume	non
53	48	0,999999	2	0,005126	Simple	non
53	82	1	2	0,008907	Polarité	non
53	82	1	2	0,004321	Simple	non
54	48	0,734099	2	0,009448	Grantham	non
54	48	0,932402	2	0,007450	Polarité	non
54	6	0,953482	2	0,001488	Volume	non
54	82	0,734094	2	0,009448	Grantham	non
54	82	0,932393	2	0,007450	Polarité	non
54	48	0,999905	2	0,000213	Volume	non
54	82	0,999905	2	0,000213	Volume	non
55	26	0,734661	2	0,005969	Simple	non

55	48	0,69387	2	0,008126	Simple	non
55	5	0,705145	2	0,007623	Simple	non
55	53	0,967338	2	0,002025	Volume	non
55	82	0,693866	2	0,008126	Simple	non
62	27	0,815273	2	0,001495	Simple	non
62	49	0,717477	2	0,007601	Simple	non
63	48	0,999999	2	0,003690	Grantham	non
63	48	0,999999	2	0,005126	Simple	non
63	48	0,999999	2	0,003455	Volume	non
63	6	0,970054	2	0,008372	Grantham	non
63	82	0,999999	2	0,003690	Grantham	non
63	82	0,999999	2	0,005126	Simple	non
63	82	0,999999	2	0,003455	Volume	non
65	14	0,693047	2	0,008747	Simple	non
65	2	0,962845	2	0,003444	Polarité	non
65	2	0,705782	2	0,007730	Simple	non
65	2	0,930921	2	0,000516	Grantham	non
77	32	0,763221	2	0,004875	Simple	non
80	48	0,938697	2	0,007767	Polarité	non
80	53	0,926093	2	0,003726	Volume	non
80	82	0,938707	2	0,007767	Polarité	non
81	9	0,724824	2	0,006288	Simple	non
91	89	0,74666	2	0,009645	Grantham	non
92	28	0,788142	2	0,007005	Grantham	non
92	28	0,972752	2	0,002097	Polarité	non
92	69	0,939363	2	0,003356	Polarité	non
94	32	0,788092	2	0,007567	Grantham	non
98	14	0,982769	2	0,005917	Grantham	non
98	14	0,971028	2	0,007424	Polarité	non
98	14	0,980188	2	0,003915	Volume	non
104	90	0,793766	2	0,008322	Charge	non
104	90	0,805367	2	0,001375	Simple	non
105	5	0,722799	2	0,006500	Simple	non
203	25	0,817968	2	0,001604	Simple	non
212	91	0,853518	2	0,000410	Simple	non
214	13	0,810612	2	0,002131	Simple	non
214	18	0,956198	2	0,008050	Polarité	non
218	63	0,723806	2	0,006386	Simple	non
224	31	0,700373	2	0,008565	Simple	non
226	48	0,999999	2	0,003229	Polarité	non
226	48	0,999999	2	0,005127	Simple	non
226	82	1	2	0,002220	Polarité	non
226	82	0,999999	2	0,005127	Simple	non
228	35	0,700028	2	0,006717	Simple	non
229	16	0,821565	2	0,008892	Simple	non
231	49	0,72203	2	0,009591	Simple	non

231	62	0,847333	2	0,001421	Simple	non
256	9	0,825912	2	0,001703	Simple	non
267	48	0,999993	2	0,006360	Simple	non
267	82	0,999994	2	0,006148	Simple	non
272	63	0,701392	2	0,008127	Simple	non
274	26	0,797248	2	0,001938	Simple	non
279	82	1	2	0,008740	Polarité	non
279	82	0,999999	2	0,005246	Simple	non
279	82	0,999999	2	0,004191	Volume	non

D

Position sur SRK	Position sur SCR	Statistique de coévolution	Taille	p-value	Propriété biochimique	FDR
12	6	0,612734	2	0,0085409	Volume	non
26	53	0,70027	2	0,0044109	Volume	non
32	78	0,609596	2	0,009406	Grantham	non
33	54	0,491127	2	0,0091226	Polarité	non
34	40	0,577304	2	0,0099826	Volume	non
39	64	0,543157	2	0,006732	Polarité	non
40	28	0,869579	2	0,0010623	Polarité	non
40	69	0,761865	2	0,0034525	Polarité	non
41	61	0,653153	2	0,0067971	Volume	non
45	43	0,580824	2	0,0067447	Volume	non
45	82	0,998002	2	0,0007767	Polarité	non
48	26	0,529718	2	0,009502	Grantham	non
48	48	0,998996	2	0,0001937	Volume	non
49	48	0,96325	2	0,0005937	Volume	non
53	48	0,857335	2	0,0086383	Polarité	non
54	6	0,766548	2	0,0028387	Polarité	non
54	6	0,832411	2	0,0009099	Volume	non
57	58	0,51474	2	0,0084041	Polarité	non
63	43	0,613163	2	0,0053377	Volume	non
65	44	0,551749	2	0,0084493	Charge	non
72	62	0,5684	2	0,0061033	Charge	non
80	26	0,635241	2	0,009191	Polarité	non
91	89	0,589631	2	0,0091625	Grantham	non
92	28	0,690079	2	0,0053505	Polarité	non
92	69	0,771048	2	0,0032103	Polarité	non
94	32	0,625023	2	0,004448	Grantham	non
104	90	0,658553	2	0,0004237	Charge	non
200	24	0,655822	2	0,0076534	Polarité	non
211	86	0,587853	2	0,0055094	Polarité	non

222	3	0,77031	2	0,0077702	Charge	non
225	32	0,699277	2	0,0025234	Polarité	non
226	6	0,572829	2	0,0064859	Polarité	non
235	5	0,540582	2	0,0099636	Polarité	non
243	58	0,604694	2	0,0013118	Charge	non
262	89	0,622017	2	0,009846	Polarité	non
267	48	0,997681	2	1,62E-05	Volume	non

E

Position(s) sur SRK	Position(s) sur SCR	Stat. de coévolution	Taille	p-value	Propriété biochim.	FDR
25, 229, 304	16	0,716373	4	0,0086374	Simple	oui
109, 297, 315	49	0,721087	4	0,0076485	Simple	oui
120, 132, 134, 376	48, 82	1	6	0,0106545	Grantham	oui
120, 132, 134, 376	48, 82	1	6	0,0193659	Volume	oui
164, 294	51, 69	0,693135	4	0,0326485	Simple	oui
212	91	0,871588	2	0,0399952	Simple	oui
231	62	0,868685	2	0,037379	Simple	oui
380, 400	40	0,863147	3	0,0015519	Simple	oui
397	28	0,999997	2	0,0067039	Charge	oui
434	91	1	2	0,0020759	Charge	oui
34, 44, 95, 104, 209, 382	-	0,957685	6	0,0164306	Volume	oui
37, 158, 186, 397, 401, 405	-	0,592999	6	0,0257161	Simple	oui
48, 53, 63, 122, 137, 226, 279	-	1	7	0,0342192	Simple	oui
48, 63, 122, 137, 279	-	1	5	0,0312341	Volume	oui
48, 63, 122, 137, 226, 279	-	1	6	0,0250074	Grantham	oui
194, 297, 298	-	0,975889	3	0,0109556	Polarité	oui
194, 298	-	0,979163	2	0,0029111	Grantham	oui
222, 449	-	0,999999	2	0,0350501	Volume	oui
237, 316	-	0,866881	2	0,0445603	Simple	oui
291, 323, 409, 410	-	0,718069	4	0,0107022	Simple	oui
375, 377	-	0,877129	2	0,0468973	Simple	oui
391, 443	-	0,945443	2	0,0044609	Simple	oui

F

Position(s) sur SRK	Position(s) sur SCR	Stat. de coévolution	Taille	p-value	Propriété biochim.	FDR
40, 113, 160, 256, 300, 366	28, 90	0,911244	8	0,03558	Polarity	oui
57, 82, 118, 304, 322, 377	49	0,870342	7	0,0136346	Charge	oui

62, 80, 256, 278, 390, 399	9	0,855544	7	0,0348837	Volume	oui
91, 94, 166	81	0,80642	4	0,0245296	Charge	oui
125, 339, 382, 436	41, 97	0,857275	6	0,0254623	Grantham	oui
132	82	0,944763	2	0,018209	Volume	oui
137	48	0,999602	2	0,0040699	Volume	oui
269, 440	42	0,745389	3	0,0467706	Charge	oui
328	32	0,905225	2	0,0450273	Volume	oui
397	28	0,998656	2	0,0015401	Charge	oui
7, 222, 226, 245, 311, 368, 391	-	0,860792	7	0,03125	Charge	oui
18, 175, 313	-	0,816895	3	0,0329913	Charge	oui
31, 94, 248, 311	-	0,858416	4	0,0120223	Grantham	oui
34, 54, 298, 337	-	0,882155	4	0,0014582	Grantham	oui
37, 213	-	0,727353	2	0,0478066	Grantham	oui
40, 93	-	0,993728	2	0,0031248	Volume	oui
43, 95, 129, 253, 366, 419, 454	-	0,863151	7	0,0238468	Charge	oui
44, 209	-	0,955499	2	0,021104	Polarity	oui
45, 120	-	0,999486	2	0,0078576	Polarity	oui
48, 279	-	0,99999	2	0,0001019	Grantham	oui
48, 63	-	0,999946	2	0,0002968	Volume	oui
49, 122, 421, 429	-	0,996849	4	0,004264	Volume	oui
49, 267	-	0,998452	2	0,0152701	Polarity	oui
82, 253	-	0,853008	2	0,0385655	Polarity	oui
91, 151, 164, 292	-	0,815919	4	0,0390464	Grantham	oui
104, 113, 120, 132, 133, 183, 243, 270, 340, 366	-	0,915442	10	0,0191799	Grantham	oui
104, 209	-	0,972575	2	0,0347644	Volume	oui
132, 279	-	0,999685	2	0,005372	Polarity	oui
133, 270	-	0,987178	2	0,0316515	Volume	oui
194, 297	-	0,866382	2	0,007731	Polarity	oui
208, 445	-	0,983676	2	0,0284162	Grantham	oui
216, 265	-	0,992284	2	0,0258866	Polarity	oui
222, 449	-	0,993317	2	0,0034589	Volume	oui
256, 352	-	0,800302	2	0,028364	Grantham	oui
270, 429	-	0,998045	2	0,0163712	Polarity	oui
421, 429	-	0,993647	2	0,0128036	Grantham	oui

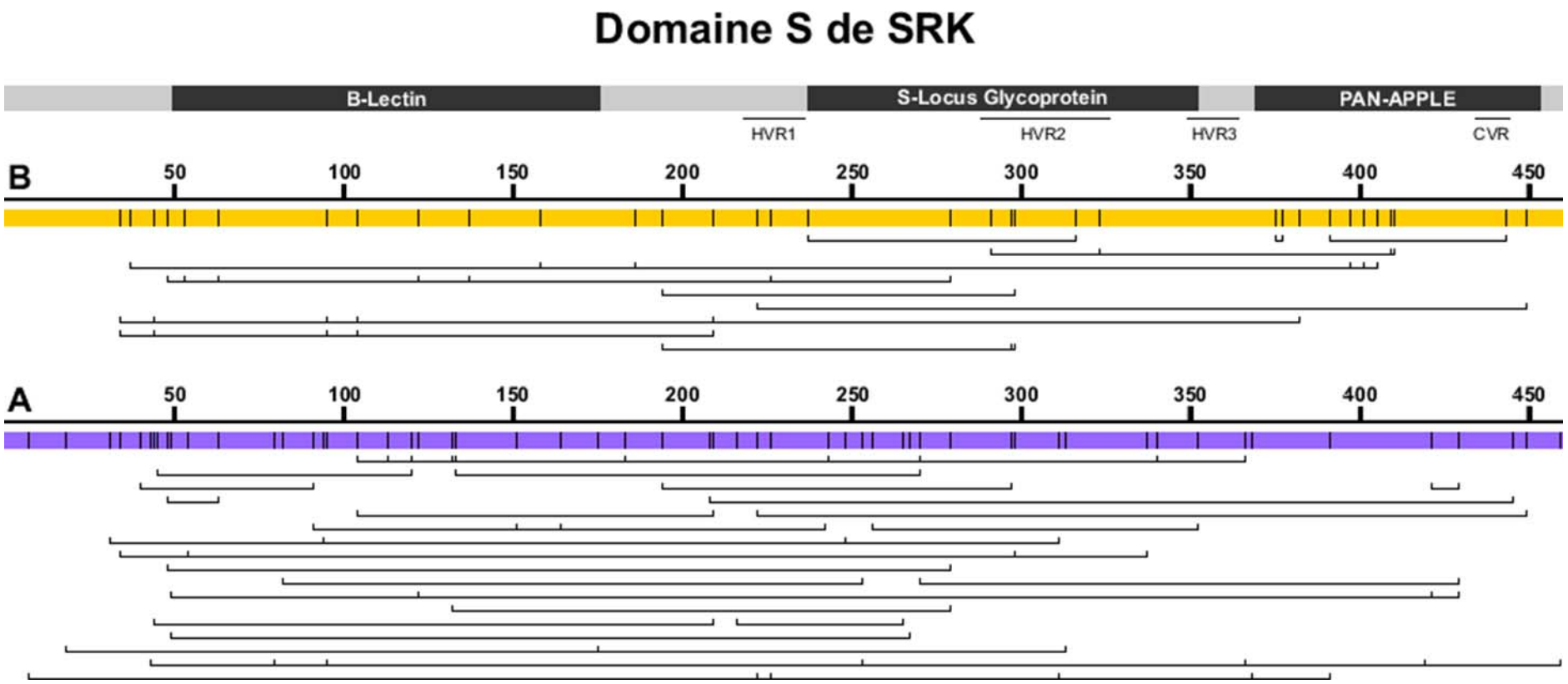


Figure supplémentaire 11. Groupes de positions montrant un signal de coévolution par corrélation (A) et par compensation (B) dans le domaine S de la protéine SRK, dans le cadre de l'analyse par clusters (DUTHEIL and GALTIER 2007). Une représentation schématique du domaine S de SRK est incluse en haut de la figure, indiquant les différentes régions fonctionnelles (MARCHLER-BAUER *et al.* 2011) et les régions hypervariables (NISHIO and KUSABA 2000). Les positions indiquées se réfèrent à l'alignement des séquences protéiques, et non aux positions absolues sur les séquences.

RESUME

Chez les plantes hermaphrodites, l'auto-incompatibilité est un système génétique permettant de limiter la dépression de consanguinité par évitement de l'autofécondation. Ce système est considéré en biologie évolutive comme l'un des caractères modèles d'une forme particulière de sélection naturelle, la sélection fréquence-dépendante. Chez les Brassicaceae, le système d'auto-incompatibilité est contrôlé par une région génomique appelée le locus S. Cette région comprend deux gènes fortement liés, dont un codant une protéine déposée à la surface du pollen et l'autre une protéine transmembranaire du pistil. La reconnaissance de type clé-serrure entre ces deux protéines provoque une cascade de réactions se traduisant par l'inhibition de la croissance du tube pollinique. Si les gènes impliqués sont relativement bien décrits, peu de données relatives à la diversité et à la dynamique de leur région génomique sont à jour disponibles. Dans ce contexte, douze séquences génomiques comprenant le locus S ont été obtenues dans le genre *Arabidopsis* par le séquençage de clones BAC. Ces séquences illustrent l'intérêt des données génomiques dans l'analyse d'une région telle que le locus d'auto-incompatibilité, soumise à de fortes contraintes sélectives. Dans un premier temps, l'annotation d'une douzaine de séquences fonctionnelles chez *A. lyrata* et *A. halleri* a permis d'examiner les patrons d'évolution moléculaire du locus d'auto-incompatibilité et de ses régions flanquantes. Une seconde partie se concentre quant à elle sur la perte du système d'auto-incompatibilité chez *A. thaliana*, et notamment sur l'occurrence de réarrangements et d'évènements de recombinaison entre séquences non fonctionnelles. Enfin, une analyse préliminaire de la coévolution entre les protéines du pollen et du pistil a pu être réalisée.

Mots clés : Systèmes de reproduction, auto-incompatibilité sporophytique, locus S, recombinaison, sélection fréquence-dépendante, dominance, éléments transposables.

ABSTRACT

Self-incompatibility is a common genetic system limiting inbreeding depression by preventing selfing. This system is considered in evolutionary biology as one of the models of frequency-dependant selection, a particular type of natural selection. In the Brassicaceae family, the self-incompatibility system is controlled by a genomic region called the S-locus and comprising two tightly linked genes. The first gene encodes a ligand deposited on the pollen surface and the second its transmembrane receptor. Molecular recognition between these two proteins leads to a cascade of reactions resulting in the reject of self-pollen. If the self-incompatibility genes are becoming well understood, the diversity and dynamics of their genomic region remains poorly described. In this context, twelve genomic sequences of the region comprising the S-locus were obtained in the genus *Arabidopsis* through sequencing of BAC clones. These sequences highlight the relevance of genomic data in the analysis of regions under such selective constraints. First, the annotation of twelve functional sequences in *A. lyrata* and *A. halleri* allows to study the patterns of evolution of the S-locus and its flanking regions. Second, the loss of the system was investigated in *A. thaliana*, in particular through the occurrence of rearrangements or recombination events in non-functional sequences. Finally, a preliminary analysis of coevolution between pollen and pistil proteins was achieved.

Keywords : Mating systems, sporophytic self-incompatibility, S-locus, recombination, frequency-dependent selection, dominance, transposable elements.