Université de Lille – Sciences et Technologies

École Doctorale Sciences de la Matière, du Rayonnement et de l'Environnement

Thèse de Doctorat pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE LILLE SCIENCES ET TECHNOLOGIES

Discipline

Sciences agronomiques et écologiques

Sous-discipline

Biologie de l'environnement, des populations, écologie

HISTOIRE EVOLUTIVE D'UN GROUPE MESOPOLYPLOIDE CHEZ LES BRASSICACEAE: APPROCHES TRANSCRIPTOMIQUES ET PHYLOGENOMIQUES POUR EVALUER LES CONSEQUENCES DE LA POLYPLOÏDIE SUR L'EVOLUTION DU SYSTEME D'AUTO-INCOMPATIBILITE

Soutenue publiquement le 19 Juin 2018 par

Laura HENOCQ

Unité de recherche : Unité Evolution, Ecologie et Paléontologie (Evo-Eco-Paléo)
UMR 8198 – Villeneuve d'Ascq, France.

Malika Ainouche Professeure, UMR-CNRS 6553 ECOBIO, Rennes	Rapportrice
Guillaume Besnard Chargé de Recherche, UMR-CNRS 5174 EDB, *Toulouse*	Rapporteur
Tatiana Giraud Directrice de Recherche, UMR-CNRS 8079, ESE, *Paris-Sud AgroParisTech*	Examinatrice
Vincent Castric Directeur de Recherche, UMR-CNRS 8198 EEP, *Lille*	Examinateur
Xavier Vekemans Professeur, UMR-CNRS 8198 EEP, Lille	Directeur de thèse
Céline Poux Maître de Conférence, UMR-CNRS 8198 EEP, Lille	Co-encadrante de thèse

« Ce que j'aime chez les chercheurs, c'est leur curiosité. Le monde s'est construit sur l'attention plus forte de certains d'entre nous : celui qui découvre la pénicilline, celui qui conçoit le vaccin contre la rage, celui qui invente la bombe atomique, celui qui imagine l'aviation... Tous les progrès sont le résultat du travail d'individus un peu plus curieux que les autres, qui ont gratté un peu plus. »

Claude Lelouch ; Le dictionnaire de ma vie (2016)

« Celui qui pose une question est bête cinq minutes, celui qui n'en pose pas l'est toute sa vie. »

Bernard Werber ; Les fourmis (2002)

REMERCIEMENTS

Tout a commencé il y a de cela bien des années, alors que je n'étais encore qu'une enfant. Il serait bien trop long de retracer toute l'histoire et d'évoquer l'influence de toutes les personnes qui m'ont amenée jusqu'ici, je tiens toutefois à remercier chacune de ces personnes en tâchant de n'oublier personne. Tout d'abord, **REYNALD**, le directeur de l'association de protection de la faune sauvage le C.H.E.N.E (Centre d'Hébergement et d'Etude sur la Nature et l'Environnement) dont le discours et l'engagement m'ont profondément touchée à tel point que j'ai décidé de m'impliquer dans cette association en tant que bénévole dès le mois de Juillet 2007. J'y ai passé trois semaines consécutives durant l'été puis des jours à droite et à gauche sur mon temps libre durant les années qui suivirent. J'avais le sentiment d'être utile et d'apporter une contribution à une cause qui m'était chère. Merci infiniment à **ALAIN**, **CHRISTELLE**, **ERIC**, **MARC** et **LAURE** pour toutes les connaissances que vous m'avez transmises et pour tous les bons moments passés ensemble. C'est en partie grâce à vous que j'ai continué dans cette voie. En parallèle de ces activités de bénévolat, je poursuivais la préparation de mon baccalauréat en essayant de préciser mon projet professionnel qui était encore un peu flou à l'époque. Merci à toi **EDITH** pour avoir nourri ma curiosité scientifique et éveillé mon goût pour la recherche en cours de biologie et d'écologie, tes enseignements m'auront été précieux. Merci à toi **NICOLAS** d'avoir partagé ton expérience de thèse avec moi lorsque j'étais lycéenne. Tes journées de doctorant devaient être bien chargées mais tu as pris la peine de répondre à chacune de mes questions et de me parler de tes recherches concernant la prédation du chat sur une espèce d'oiseau endémique de l'île de La Réunion. Nos échanges ont encore plus attisé mon goût pour l'écologie et la recherche. Mes stages d'étude et mes activités de bénévolat au sein du M.E.R.L.E m'ont permis de découvrir d'autres thématiques de recherche toutes aussi passionnantes. C'est ainsi que j'ai suivi le master GEB (gestion et évolution de la biodiversité) à Lille, au cours duquel j'ai effectué mon stage de master 1 avec **MATHILDE** et **EMNA** sur le succès reproducteur des individus gynomonoïques de *Silene nutans* et mon stage de master 2 avec **JEAN-FRANÇOIS** et **LESLIE** sur l'histoire évolutive du crapaud calamite dans le Nord Pas-de-Calais. Ces deux stages ont été très enrichissants et ont attisé ma curiosité en matière de biologie de l'évolution. Mon stage de M2 avait une dimension de biologie de la conservation qui m'a permis de renouer un peu avec mes premières aspirations. Je remercie vivement mes encadrants de stage pour leur pédagogie, leur sympathie et pour toutes les choses qu'ils m'ont apprises.

A la fin de ma deuxième année de master, XAVIER et CELINE m'ont parlé d'un projet de thèse qui a suscité à la fois un grand intérêt mais aussi de grandes appréhensions de ma part. En effet, les questionnements de ce projet de recherche nécessitaient la maîtrise d'outils que je n'avais encore jamais utilisés jusqu'à présent et il me semblait difficile de mobiliser dans le cadre de cette thèse les compétences que j'avais acquises au cours de mes stages précédents. Après mûre réflexion et discussions (merci notamment à **PASCAL** pour ses conseils avisés) et après avoir regardé les projets de thèse proposés ailleurs, j'ai choisi de me lancer dans l'étude proposé par Xavier et Céline. Je savais que j'apprendrais beaucoup au cours de ma thèse et que je découvrirais tout un nouveau pan de la biologie évolutive. Et ce fût le cas.

Tout d'abord, merci à toi **XAVIER**, tu as toujours su te rendre disponible et tu as été très réactif malgré la charge incommensurable de travail qui pèse sur tes épaules. Tu as tenu, et ce dès le début, à instaurer un rendez-vous hebdomadaire ou bimensuel autant que faire se peut, et on s'y est tenu, sauf quand cela n'était pas nécessaire. Merci pour tous les échanges constructifs, de visu ou par mail, que l'on a pu avoir. Merci **CELINE** de m'avoir guidé dans les premiers temps de la thèse pour la mise en place des manipulations en serre et des premières manipulations en laboratoire. Merci aussi pour toutes les initiations à la phylogénie, moi qui n'en avais jamais fait auparavant en dehors de mes quelques cours à l'Université. Merci d'avoir eu les mots qu'il fallait quand parfois mon moral n'était pas au mieux après plusieurs semaines de manipulations avortées. J'ai souvent su trouver en toi une pointe de réconfort et une oreille attentive dans mes moments de doutes et de stress intense.
Un grand merci aux membres du jury qui ont accepté d'évaluer mes travaux et qui se sont déplacés pour assister à ma soutenance de thèse, **MALIKA AINOUCHE**, **GUILLAUME BESNARD**, **TATIANA GIRAUD** et **VINCENT CASTRIC**. J'ai été ravie d'avoir pu échanger avec vous sur mes recherches et de bénéficier de vos conseils avisés et de vos suggestions. Merci également à **CHRISTIAN PARISOD**, **SYLVAIN GLEMIN** et VINCENT CASTRIC qui ont participé à mon comité de thèse et qui m'ont apporté des conseils précieux.

Mes travaux de thèse ont requis de nombreux travaux de biologie moléculaire en laboratoire. A ce titre, je tiens à remercier grandement **ANNE-CATHERINE** avec qui j'ai réalisé énormément de manipulations (PCR, purification sur colonne ou sur gel, séquençage) dans la bonne humeur et la rigolade mais toujours dans le sérieux et la rigueur. J'ai trouvé en toi un soutien précieux et une bonne camarade et je te remercie vivement pour ça. Merci également à **CHRISTELLE** avec qui j'ai également réalisé des manipulations laborieuses (BioAnalyzer) et à **CECILE** qui m'a initiée aux extractions d'ADN et aux PCR lors de mes stages antérieurs, ce qui m'a été très utile. Mes travaux de thèse ont aussi nécessité de faire pousser et de maintenir pas mal de plantes en serre pour réaliser entre autres des croisements, je tiens donc à adresser un immense merci à **ERIC** qui a bichonné toutes mes belles petites plantes. C'était un plaisir de travailler à la serre avec toi et nos discussions scientifiques et

souvent moins scientifiques étaient vraiment appréciables. Merci également à **CEDRIC**, **NATHALIE**, **ANGELIQUE** ET **CHLOE**. Enfin, merci à **CHLOE C.** qui a effectué son stage de Master 1 à mes côtés et avec qui j'ai pu réaliser, entre autres, la détermination du système de reproduction des espèces que j'étudiais.

Mes travaux de thèse ont également requis de nombreuses analyses bio-informatiques, lesquelles ont été menées par mes soins mais aussi et surtout par de nombreux collègues que je tiens à remercier tout particulièrement. Merci à **SOPHIE** pour son aide précieuse et faramineuse dans la rédaction de centaines voire de milliers de lignes de commande nécessaires à l'élaboration de nos phylogénies de gènes puis nos phylogénies d'espèces. On en aura passé des heures à discuter longuement toutes les deux et à s'arracher les cheveux sur les difficultés qu'on peut rencontrer quand on cherche à faire de la phylogénomique à partir de données d'assemblages de transcriptome et d'espèces anciennement polyploïdes. Merci pour ton « acharnement », ta persévérance à toute épreuve. Merci à **MATHIEU**, que j'ai sollicité pour mes « Orychotrucmuches » mais pour des tas d'autres choses encore ! Merci également à **CLEMENT**, qui m'a beaucoup aidé sur la prise en main du pipeline WGDetect et d'autres analyses connexes. Enfin, merci à la plateforme de bio-informatique et de bio-analyse **BILILLE** de m'avoir fourni les ressources nécessaires pour réaliser bon nombre d'analyses longues et gourmandes.

Je ne saurais oublier de remercier aussi bon nombre de personnes qui m'ont accompagnée durant ma thèse, davantage sur le plan personnel que professionnel, bien qu'elles m'aient également apportée divers conseils et suggestions utiles à mes travaux. En premier, évidemment, merci à **NICOLAS**, mon « compagnon de galère ». On a commencé cette aventure ensemble et on l'a terminée ensemble. Comment aurais-je vécu ma thèse sans toi ? Je l'ignore, mais elle n'aurait certainement pas eu la même saveur. Merci à toi pour ton soutien, pour nos franches rigolades, pour nos délires complètement barrés mais tellement revigorants ! Nos chemins se sont croisés durant la thèse mais nos routes se croiseront encore je l'espère ! Merci à mes collègues de bureau que ce fut un plaisir de retrouver chaque jour : NICOLAS bien sûr, mais aussi **ANNE**, **RENATO**, notre regretté **ROMUALD**, **CLAUDIA** et à une époque plus ancienne, mon cher **CHRISTOPHE V**. L'ambiance était studieuse mais toujours ponctuée de petites parenthèses d'échanges moins professionnels, de rigolades ou de confessions. Ce fût vraiment un plaisir de travailler à vos côtés. Et merci pour les chocolats, les barres chocolatées et autres confiseries qui apportaient un peu de douceur dans un quotidien parfois maussade. Un grand merci tout particulier à **MATHILDE L**. en qui j'ai trouvé une amie et confidente plus qu'une collègue, et qui m'a rendu bon nombre de services en plus de m'apporter régulièrement, sans même que je ne le demande, de délicieux gâteaux et chocolats ! J'en aurais pour des années à te rendre la pareille ma belle. Un immense merci à **AMELIE** qui, grâce à mes demandes de dernière minute, connaît maintenant mieux que quiconque l'emplacement du service d'impression et du service postal de la fac, et merci pour les délicieuses et impressionnantes pâtisseries que tu m'as fait l'honneur de préparer pour mon pot de thèse. Tu es toujours prête à aider les autres, et avec plaisir en plus ! Ta générosité débordante et ta bienveillance m'impressionneront toujours. Si seulement je pouvais faire d'aussi bons gâteaux que les tiens…
Enfin, merci à **VINCENT C**. pour ses taquinages quotidiens et ses « reboostages » jusqu'à la dernière minute même après son départ du labo et à des milliers de kilomètres d'ici, merci à **CAMILLE**, **HELENE**, **DIMA**, **CLOTILDE**, **THIBAULT**, **THOMAS B.**, **THOMAS L.**, **MATHILDE P.**, **MARINA**, **CHRISTOPHE V. B.**, **LESLIE**, **MAXIME**, **CLEMENTINE**, **NATASHA**, **ALESSANDRO**, **ESTELLE**, **BENEDICTE**, **MARYSE**, **MARIE-JOE** et **VINCENT B**. Les moments que j'ai eu le plaisir de partager avec vous au labo ou en dehors étaient tous délicieux et je suis ravie de tous vous avoir rencontré, vous êtes des personnes formidables et que j'apprécie vivement. Merci également à **DIALA** et à **ROXANNE** que j'ai eu la chance de rencontrer et avec qui j'ai eu l'occasion d'échanger en master et en début de thèse, en cours, en salle café ou dans les bars animés de Lille. Merci à l'ensemble du laboratoire GEPV/EEP que je fréquente depuis plus de 6 ans maintenant. Merci à **TOUS LES COLLEGUES** toujours disponibles pour apporter de précieux conseils. La convivialité exemplaire qui règne dans ce laboratoire de recherche a égayé chacune de mes journées. Je crois que je ne trouverais nulle part ailleurs une salle café où les débats sont aussi riches et animés. C'est l'esprit de ce laboratoire, et j'espère de tout cœur qu'il perdurera le plus longtemps possible. « Un salarié heureux en vaut deux » ! Cette expérience de thèse m'a permis de constater à quel point cela peut être vrai.

Merci à tou·te·s mes ami·e·s Lillois·e·s et Normand·e·s et d'autres contrées pour m'avoir accompagnée durant cette belle aventure et, pour certain·e·s, de m'y avoir conduit. Merci **BORIS** et **JULIEN** pour votre soutien et les innombrables moments de décompression que vous m'avez offerts. Merci **INGRID**, **CHDUC**, **SEUMCHOU** et tous les autres. Merci **RAPHAËL** pour tes encouragements et tous les efforts que tu as fait pour comprendre tant bien que mal mes recherches. Merci à **MES PARENTS** qui ont éveillé mon goût intarissable pour la nature et qui ont toujours cru en moi, dont le soutien a été sans faille et qui m'ont toujours encouragée à poursuivre ma voie malgré les difficultés que cela pouvait impliquer. Merci de même à ma sœur **MARIE** et mon frère **PIERRE** qui m'ont soutenu de A à Z et qui ont toujours grandement estimé ce que je faisais. Merci à **LA FAMILLE FLOURENS** qui m'a vivement encouragée. Enfin, merci **LUCAS** pour ta patience, ton soutien quotidien dans les bons comme dans les mauvais moments, d'avoir su m'arracher des sourires même quand le stress et l'inquiétude étaient au maximum. Je remercie vivement la **REGION NORD-PAS-DE-CALAIS** pour la bourse de thèse et le **CPER-CLIMIBIO** pour les financements accordés afin de réaliser ce travail de thèse.

SOMMAIRE

« Entre ce que je pense, ce que je veux dire, ce que je crois dire, ce que je dis, ce que vous avez envie d'entendre, ce que vous croyez entendre, ce que vous entendez, ce que vous avez envie de comprendre, ce que vous croyez comprendre, ce que vous comprenez, il y a dix possibilités qu'on ait des difficultés à communiquer. Mais essayons quand même… »

Bernard Werber ; l'Encyclopédie du savoir relatif et absolu (2003)

INTRODUCTION GÉNÉRALE

Les espèces polyploïdes sont extrêmement répandues chez les plantes vasculaires (Jiao et al. 2011; Van de Peer et al. 2017), et dans une moindre mesure chez les métazoaires (poissons et amphibiens), bien que la plupart des génomes eucaryotes révèlent des traces de polyploïdies ancestrales (Wolfe and Shields 1997; Yu et al. 2005; Marcussen et al. 2014; Marcet-Houben and Gabaldón 2015). L'objet de cette introduction consiste dans un premier temps à définir la notion d'espèce polyploïde et les modalités de formation de telles espèces et, dans un second temps, à évoquer l'occurrence de la polyploïdie dans le règne végétal et plus particulièrement dans la famille des Brassicaceae. Dans un troisième temps, les effets supposés à moyen et à court terme des évènements de polyploïdisation sur le génome et la diversification des lignées, en documentant tout particulièrement le cas des Angiospermes et notamment celui de la famille des Brassicaceae, seront présentés. Après une brève présentation du système d'auto-incompatibilité chez les Brassicaceae, la troisième partie de l'introduction évoquera aussi les conséquences attendues des évènements de duplication de génome sur ce système de reproduction. Le modèle d'étude de cette thèse sera exposé en quatrième et dernière partie de l'introduction. Enfin, le contexte ainsi que les objectifs de la thèse seront présentés.

1. Les événements de duplication de génome (WGD)

Un organisme polyploïde comporte par définition, et ce de façon héritable, plus de deux lots de chromosomes homologues (au sens large), à la différence d'un organisme diploïde. En général, on parle de polyploïdisation pour évoquer la formation de nouvelles espèces polyploïdes (néopolyploïdes) mais on peut aussi parler d'événement de duplication (ou triplication) de génome entier (WGD ou WGT pour « Whole Genome Duplication » et « Whole genome Triplication »). Les espèces polyploïdes sont séparées en différentes catégories selon leur composition chromosomique et leur mode de formation. En effet, une espèce néopolyploïde peut soit être (i) le résultat d'une hybridation entre deux espèces diploïdes (espèces A et B transmettant leur génome diploïde AA et BB, respectivement), plus ou moins proches, associée à la fusion de leurs jeux de chromosomes non homologues (homéologues, AABB) ; on parle alors d'événement d'allopolyploïdie et d'espèce allopolyploïde, soit (ii) le résultat d'une reproduction entre deux individus de la même espèce, associée à une augmentation de la ploïdie par fusion de deux jeux de chromosomes strictement homologues (AAAA) ; on parle dans ce cas d'autopolyploïdie et d'espèce autopolyploïde (*cf.* FIGURE 1).

La fréquence de l'autopolyploïdie semble marginalement plus importante que la fréquence de l'allopolyploïdie, longtemps considérée comme largement prédominante (Barker et al. 2016). En réalité, la polyploïdie correspond davantage à un continuum entre la fusion de génomes strictement identiques et celle de génomes fortement différenciés (Tayalé and Parisod 2013). En termes de ségrégation des segments de chromosomes lors de la méiose chez une espèce néopolyploïde, on observe en général une ségrégation disomique (par paires de chromosomes homologues) chez les allotétraploïdes, et une ségrégation tétrasomique chez les autotétraploïdes (par tétrade de chromosomes homologues, toutes les combinaisons alléliques possibles étant produites en fréquence égale) (Soltis and Soltis 1993; Ramsey and Schemske 2002). Les modalités de formation des polyploïdes sont multiples mais la fusion de gamètes non réduits (*i.e.* diplogamètes) et dans une moindre mesure, le doublement spontané du jeu de chromosomes dans les cellules somatiques sont en général à l'origine de la formation des polyploïdes (Otto and Whitton 2000; Comai 2005; Tayalé and Parisod 2013).

La production de diplogamètes est très variable chez les plantes entre et au sein même des espèces mais suffisamment fréquente pour expliquer la formation de polyploïdes à un taux comparable à celui des mutations géniques, environ 1 par 100,000 chez les plantes à fleurs (Ramsey and Schemske 1998; Brownfield and Köhler 2011). Certains travaux suggèrent que ce processus pourrait être induit par un certain nombre de stress environnementaux tels que le froid, l'herbivorie, les blessures et le manque d'eau ou de nutriments (Mason et al. 2011) mais les mécanismes moléculaires sous-jacents restent peu documentés (mais voir Brownfield and Köhler (2011)).

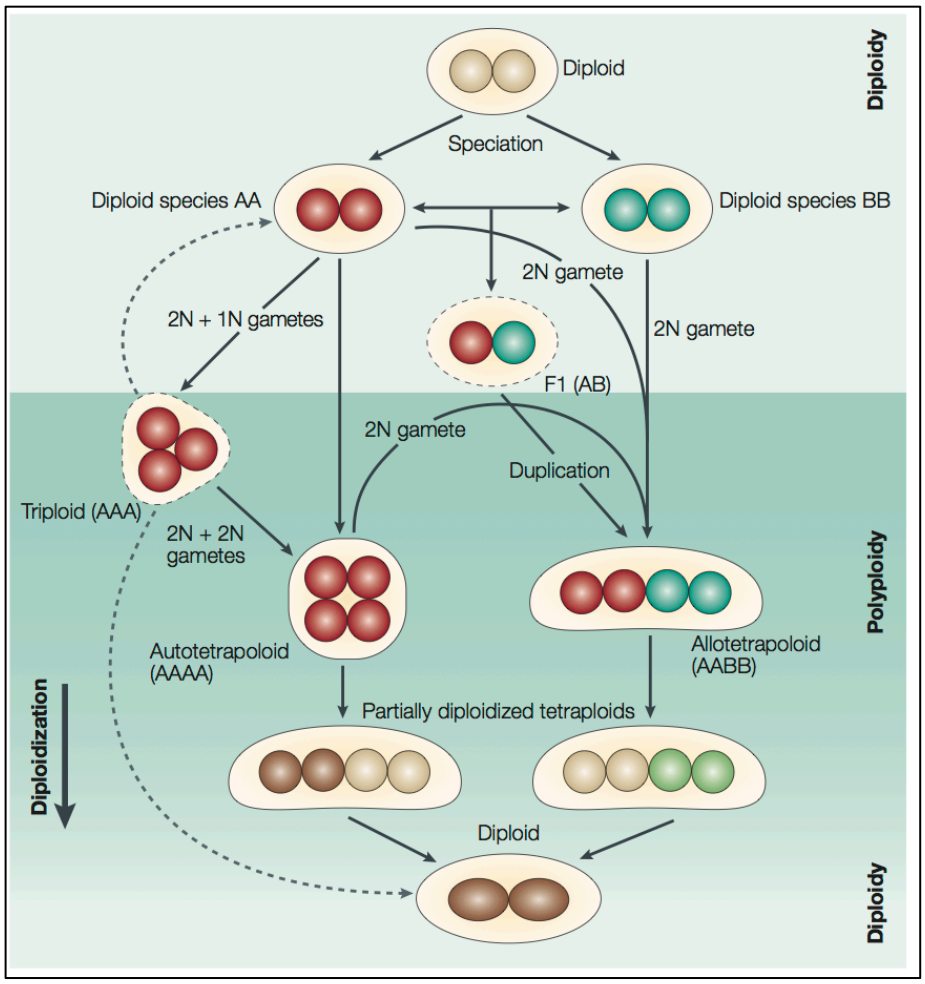


FIGURE 1 – Schéma des différentes modalités de formation des auto et des allopolyploïdes et transition graduelle de la polyploïdie vers la diploïdie [Figure extraite de Comai (2005)]. Toutes les modalités de formation des polyploïdes ne sont pas représentées. Les triploïdes, par exemple, contribuent uniquement à la formation des autotétraploïdes sur le schéma, mais ils peuvent aussi contribuer à la formation des allopolyploïdes. Chaque génome haploïde est représenté par un rond ou un ovale coloré à l'intérieur d'un contenant nucléaire beige. Certains génomes sont illustrés par un ovale pour représenter l'augmentation génomique associée à la rétention et à la sous-fonctionnalisation des gènes dupliqués qui ont lieu durant le processus de diploïdisation. Les cercles ou les ovales de couleur différente correspondent à des génomes divergents. Les états de ploïdie fortement instables sont entourés par des lignes nucléaires pointillées. A et B représentent le type de génome et N la ploïdie des gamètes.

Après l'événement de polyploïdisation, des changements génétiques, épigénétiques et structuraux réorganisent généralement le génome polyploïde en génome diploïde, de manière progressive, c'est ce que l'on appelle le processus de diploïdisation, qui sera brièvement expliqué ultérieurement dans ce document (Tayalé and Parisod 2013; Garsmeur et al. 2014; Panchy et al. 2016). Une distinction temporelle est ainsi généralement opérée entre les espèces paléopolyploïdes et les espèces néopolyploïdes. Les premières (paléo-) montrent une signature génomique d'évènement de polyploïdie ancien et leur génome est actuellement complètement "diploïdisé" (voir FIGURE 1), alors que les secondes (néo-) ont subi un évènement récent de polyploïdie et leur génome n'est pas encore diploïdisé (on peut alors clairement distinguer les deux sous-génomes parentaux) (Blanc and Wolfe 2004). Certains auteurs distinguent également une catégorie supplémentaire d'espèces polyploïdes, les mésopolyploïdes, pour lesquelles le processus de diploïdisation est en cours. Dans le génome des espèces mésopolyploïdes, il est alors possible de retrouver et distinguer les différents jeux de chromosomes ayant fusionnés par des approches de génétique et de cytogénomique comparatives (Mandáková et al. 2010).

2. Les duplications de génome entier : des évènements pas si rares que ça…

2.1. Polyploïdie chez les animaux

Chez les vertébrés, plusieurs évènements de duplication de génome ont été détectés, notamment un chez l'ancêtre commun des téléostéens actuels (Putnam et al. 2008; Glasauer and Neuhauss 2014). Des évènements plus récents ont été identifiés chez les Salmonidés (Berthelot et al. 2014) et chez les Cyprinidés (Xu et al. 2014), pour ne citer qu'eux (voir Le Comber and Smith 2004)). Certaines lignées d'amphibiens sont également concernées par les polyploïdies (Schmid et al. 2015; Session et al. 2016), tandis que la situation chez certains mammifères n'est pas encore tout à fait claire (Gallardo et al. 2004; Evans et al. 2017). De tels évènements ont également été signalés chez *Daphnia pulex,* un crustacé branchiopode (Adamowicz et al. 2002; Vergilino et al. 2009).

Cependant, il apparaît que la polyploïdie est bien moins répandue dans le règne animal que dans le règne végétal, probablement parce que les « barrières » à la polyploïdisation y sont plus importantes, telles que la présence de chromosomes sexuels, l'omniprésence de la reproduction croisée et la complexité histologique de la plupart des lignées animales (Mable 2004a). Toutefois, cette observation n'a jamais fait l'objet d'une étude en tant que telle.

2.2. Polyploïdie chez les végétaux

Chez les Angiospermes (plantes à fleur), les évènements de duplication de génome entier sont très largement répandus et sont très fréquents (Blanc and Wolfe 2004; Cui et al. 2006; Van de Peer et al. 2009; Kagale et al. 2014; Yang et al. 2015; Smith et al. 2017; Walker et al. 2017). En effet, on estime que 15% des évènements de spéciation sont associés à une augmentation de la ploïdie chez les Angiospermes (Wood et al. 2009). De plus, tous les représentants actuels des Angiospermes partagent au moins un événement de duplication de génome (nommé ε) précédant la diversification de ce clade (Jiao et al. 2011). En outre, la majorité des plantes à fleur présente des traces d'évènements anciens de duplication de génome (paléopolyploïdies) et les évènements de polyploïdie se succèdent dans le temps le long des lignées végétales (Van de Peer et al. 2009; Van de Peer et al. 2017), ce qui explique que l'on retrouve ainsi au sein des génomes, dans certaines lignées d'Angiospermes, des traces de plusieurs évènements de WGD et/ou WGT (FIGURE 2).

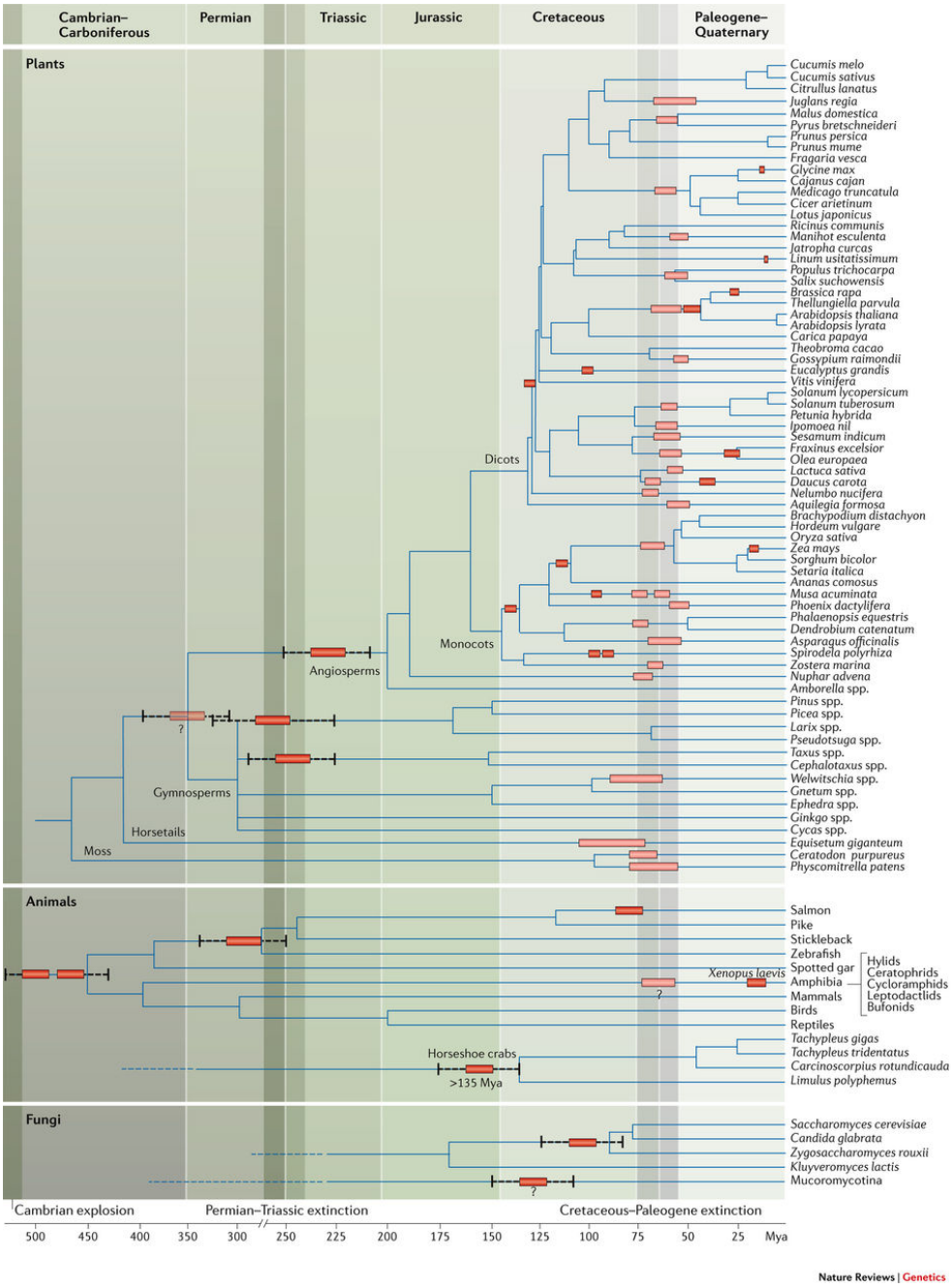


FIGURE 2 – Arbre tronqué pour les plantes, les animaux et les champignons présentant les relations phylogénétiques entre des espèces concernées par des évènements de polyploïdie et pour lesquelles un génome ou des données transcriptomiques sont disponibles [Figure extraite de Van de Peer et al. (2017)]. Les évènements de duplication de génome (WGDs) qui ont été décrits dans une ou plusieurs études sont indiqués sur l'arbre (rectangles), et l'incertitude sur l'âge de ces évènements est représentée par une ligne noire, pointillée et en gras. Les WGDs dont l'âge estimé se situe entre 55 et 75 millions d'années (zone ombragée au niveau de la frontière entre le Crétacé et le Paléogène) sont indiqués par des rectangles rouge clair. Les évènements d'extinction de masse sont représentés par des aires ombragées dont les limites correspondent à 10 millions d'années de chaque côté de l'âge estimé de l'événement.

Chez les monocotylédones, deux évènements successifs et propres à ce clade (nommés σ et ρ, dans leur ordre d'apparition, voir FIGURE 3, A) ont pris place avant la diversification des lignées principales (Tang et al. 2010) et des évènements plus récents ont été identifiés chez plusieurs espèces, dont le maïs (Swigonova et al. 2004; Schnable et al. 2009), le riz et le

blé (Yu et al. 2005; The International Wheat Genome Sequencing Consortium (IWGSC) 2014; Mckain et al. 2016).

Chez les eudicotylédones, un événement ancien de triplication de génome (nommé γ) a probablement pris place avant la diversification des plantes à fleurs (entre 90 et 130 millions d'années) (Vision et al. 2000; Blanc et al. 2003; Schranz et al. 2006; Lyons et al. 2008; Ming et al. 2008; Barker et al. 2009; Kagale et al. 2014; Edger et al. 2015). Un nombre important d'évènements de duplication de génome entier le succèdant a été largement documenté. L'événement nommé β, par exemple, a eu lieu au sein de l'ordre des Brassicales, juste après sa divergence avec la famille des Caricaceae (*Carica papaya*) (Ming et al. 2008; Barker et al. 2009), et a été suivi de l'évènement α, plus récent et spécifique à la famille des Brassicaceae (Vision et al. 2000; Barker et al. 2009; Kagale et al. 2014) (FIGURE 3, A). En effet, on ne retrouve pas la trace de cet événement chez les membres de la famille des Cleomaceae (famille sœur des Brassicaceae), qui ont subit un WGD indépendant et plus récent (FIGURE 3, B) (Schranz and Mitchell-Olds 2006; Barker et al. 2009).

A la suite de ces évènements anciens de duplication de génome (paléopolyploïdies) ont eut lieu, au cours de la diversification de plusieurs lignées, d'autres évènements de polyploïdie, plus récents, mais encore suffisamment anciens pour ne plus affecter la ploïdie des espèces actuelles (évènements de mésopolyploïdie). Le génome diploïde de *Populus trichocarpa* (famille des Salicaceae) montre par exemple des traces d'un événement de duplication de génome relativement récent et ultérieur à l'événement γ (Tuskan et al. 2006), le génome du coton diploïde (*Gossypium raimondii,* famille des Malvaceae) a récemment révélé des traces d'un événement de duplication de génome daté d'environ 60 millions d'années (Paterson et al. 2012) tandis que le séquençage du génome du soja (famille des Fabacées) a permis de révéler l'existence d'un événement de tétraploïdie daté d'environ 13 millions d'années (Walling et al. 2006). Chez la famille des Caryophyllaceae (ordre des Caryophyllales), au moins 26 événements anciens et plus récents de polyploïdie ont été révélés (Yang et al. 2015; Smith et al. 2017; Walker et al. 2017). Chez les Brassicaceae, famille végétale largement étudiée, beaucoup d'événements indépendants de mésopolyploïdie ont été mis en lumière ces dernières années (Kagale et al. 2014; Mandáková et al. 2017).

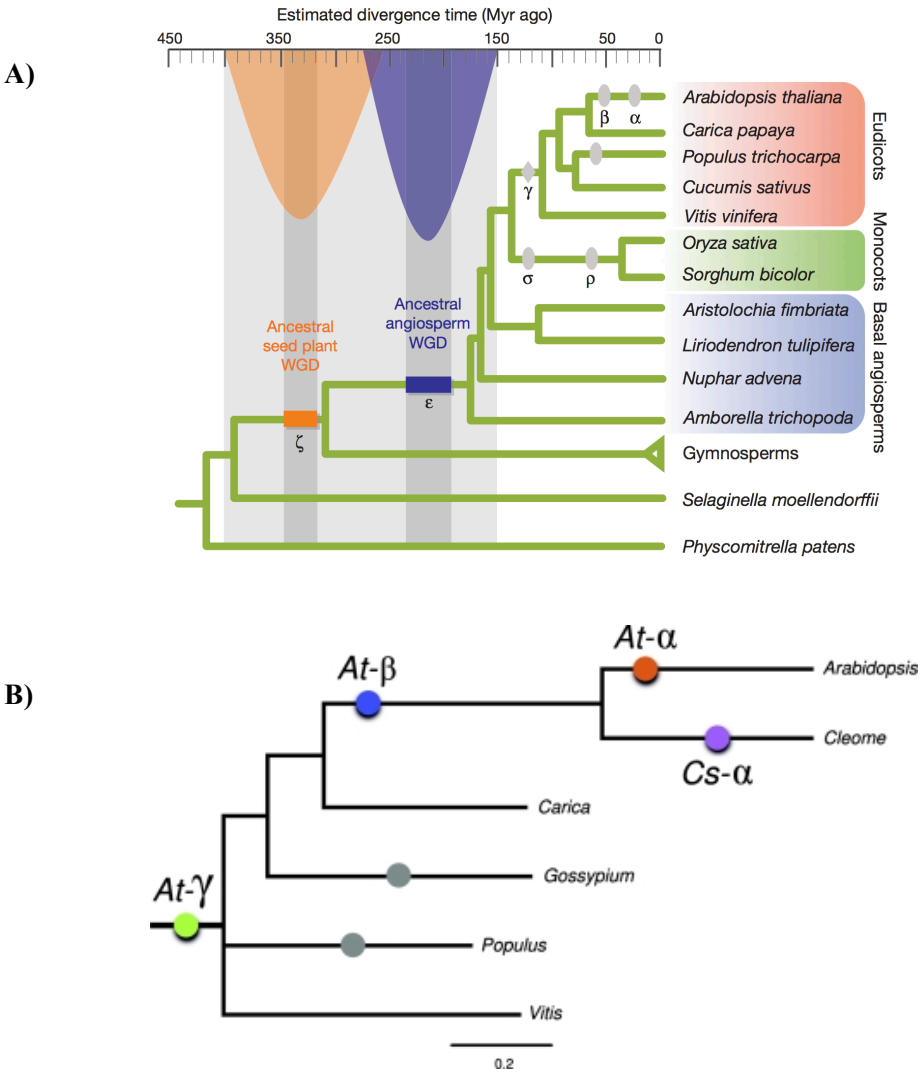


FIGURE 3 – A) Evènements ancestraux de polyploïdie chez les plantes à graines (Gymnospermes) et les Angiospermes, [Figure extraite de Jiao et al. (2011)]. Deux évènements ont été identifiés en combinant l'information apportée par les données phylogénomiques et l'horloge moléculaire de l'évolution des plantes terrestres. Les ovales représentent les évènements de duplications de génome généralement acceptés et identifiés dans les génomes séquencés (cités dans le texte de ce document). Le losange correspond à l'événement de triplication de génome probablement partagé par l'ensemble des eudicotylédones (γ). Les bars horizontales représentent l'intervalle de confiance sur les âges estimés des deux évènements ancestraux de polyploïdie (âge minimum et maximum). B) Phylogénie des eudicotylédones affichant à la fois les évènements de duplication de génome partagés par les familles des Brassicaceae et des Cleomaceae (γ et β) (Brassicales) et spécifiques à chacune des ces deux familles (At-α et C-α, respectivement), [Figure extraite de Barker et al. (2009)]. Les longueurs de branches correspondent aux valeurs moyennes de *Ks* estimées à partir de 270 gènes. Les points colorés indiquent les évènements de duplication de génome inférés dans l'étude de Barker et al. (2009) tandis que les points gris représentent ceux inférés dans des études antérieures.

2.3. Les duplications de génome entier dans la famille des Brassicaceae

Au sein du clade des Eudicotylédones, dans le groupe des Rosideae, les Brassicaceae constituent une importante famille de plantes à fleur, proche phylogénétiquement de la famille des Cleomaceae au sein de l'ordre des Brassicales. La famille des Brassicaceae (appelée jadis, la famille des Crucifères), dont l'âge est évalué entre 41 et 55 millions d'années (Beilstein et al. 2010; Franzke et al. 2011; Huang et al. 2016), est organisée en 6 lignées majeures (A – F) regroupant 51 tribus monophylétiques (FIGURE 4), 321 genres et environ 3700 espèces, distribuées dans les zones tempérées et alpines de l'ensemble des continents, en dehors de l'Arctique (Schranz et al. 2006; Beilstein et al. 2010; Warwick et al. 2010; Al-Shehbaz 2011; Franzke et al. 2011; Al-Shehbaz 2012; Huang et al. 2016).

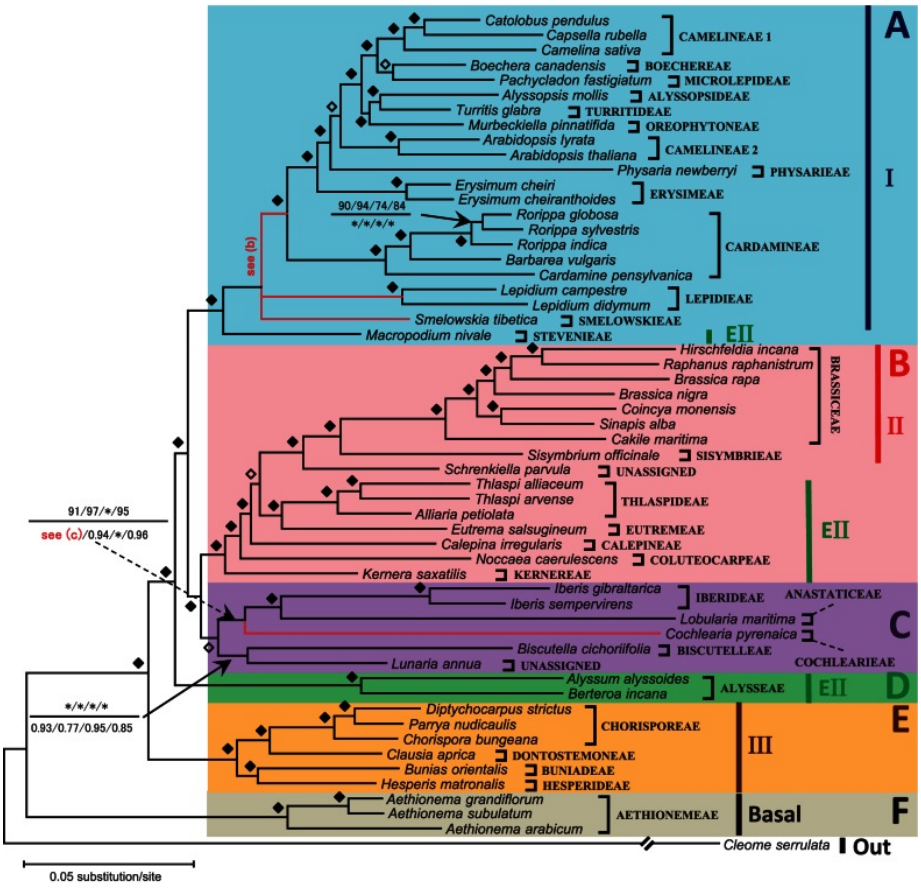


FIGURE 4 – Phylogénie des Brassicaceae obtenue par maximum de vraisemblance (ML) et inférence bayésienne (BI) [Figure extraite de Huang et al. (2016)]. La phylogénie, établie à partir d'un grand nombre de gènes nucléaires, révèle 6 clades monophylétiques (A à F) statistiquement bien soutenus, et non pas trois lignées majeures et monophylétiques comme précédemment établi (I à III). Les losanges pleins et vides indiquent les supports maximum (BP=100 et PP=1) et les supports >90 (BP) ou 0.9 (PP), respectivement. I, II, II et EII indique la position des lignées I, II et III ainsi que de la lignée II étendue (« expanded lineage II »), respectivement.

A partir du nombre de chromosomes, Warwick and Al-Shehbaz (2006) estimait que 37% des espèces de Brassicaceae sont des polyploïdes anciens ou récents (polyploïde si n≥14). A partir d'approches génomiques, transcriptomiques et de cytogénomique comparative, des évènements anciens ou récents de WGD ont été plus précisément mis en lumière dans la famille des Brassicaceae.

En plus des événements α, β et γ partagés par toutes les espèces de Brassicaceae, on recense un nombre considérable d'événements de mésopolyploïdie et de néopolyploïdie dans différentes lignées (FIGURE 4) (Haudry et al. 2013; Kagale et al. 2014; Geiser et al. 2016; Mandáková et al. 2017). Par exemple, un événement de triplication de génome (WGT pour « Whole Genome Triplication ») précède la diversification des lignées dans le genre Leavenworthia (Haudry et al. 2013) tandis qu'un événement de duplication de génome précède la diversification des lignées du genre Biscutella (Geiser et al. 2016) (FIGURE 5). L'ancêtre commun des espèces de la tribu des Brassiceae (clade B) a également subi un événement de triplication de génome (mésopolyploïdie) il y a environ 20 millions d'années (Lysak et al. 2005; Parkin et al. 2005; Ziolkowski et al. 2006; Lysak et al. 2007; Wang et al. 2011; Liu et al. 2014; Moghe et al. 2014). Le genre Orychophragmus, proche des Brassiceae, est quant a lui un ancien tétraploïde puisque l'ancêtre commun des espèces de ce genre a subi un événement de duplication de génome (mésotétraploïdie) (Lysack et al. 2007).

Ces événements de mésopolyploïdie sont suffisamment anciens pour que les espèces aient actuellement retrouvé leur état diploïde (hormis les espèces ayant par la suite subi un nouvel évènement de polyploïdie) mais suffisamment récents pour qu'on en retrouve aisément la trace dans le génome des plantes (les régions génomiques dupliquées, et donc homéologues, sont encore détectables par des approches de génétique et de cytogénomique comparatives).

Cette liste non exhaustive de l'occurrence des évènements de duplication de génome chez les Angiospermes – et particulièrement dans la famille des Brassicaceae – suffit à se faire une idée de la fréquence et de la large répartition des WGDs dans l'histoire évolutive de ce groupe.

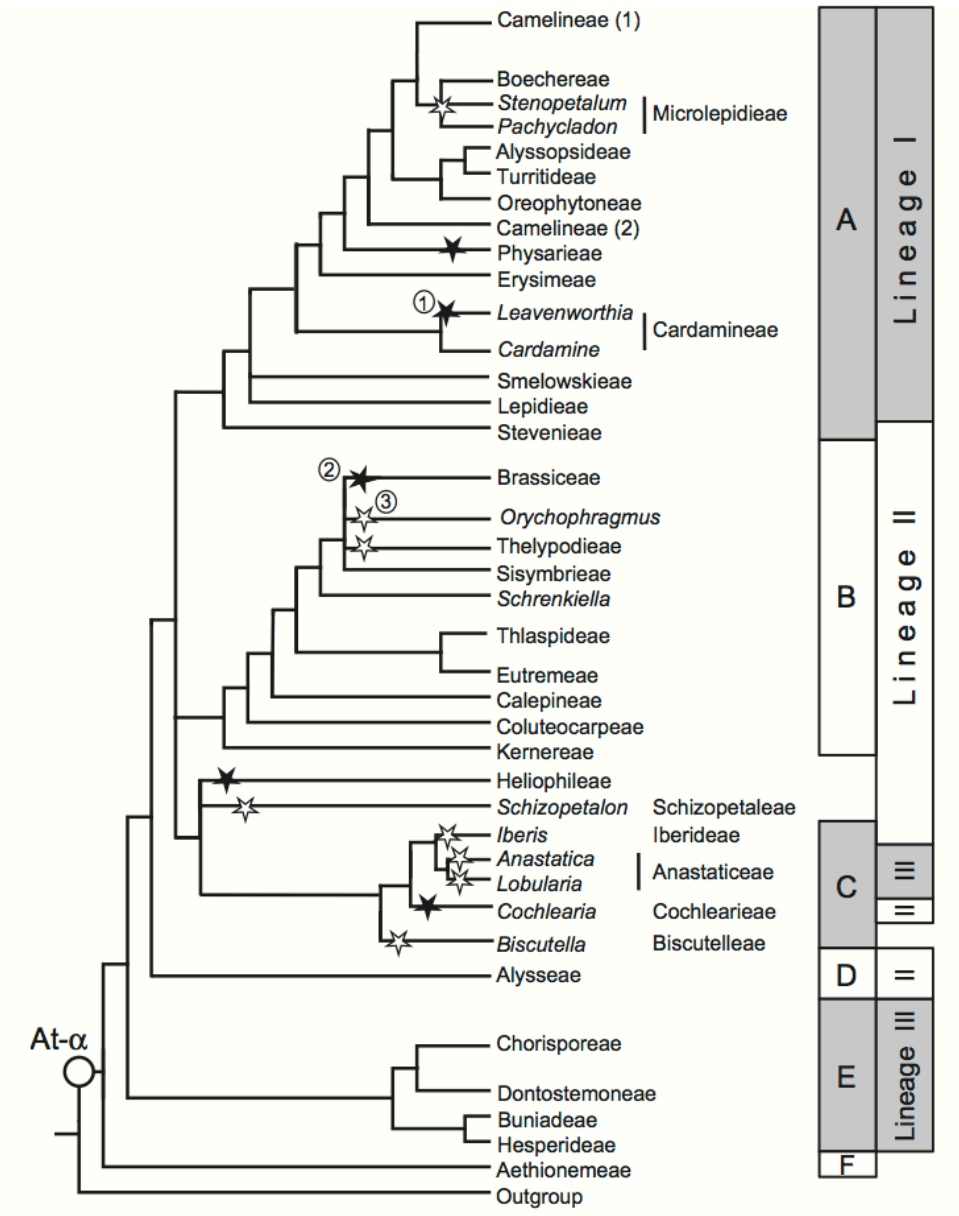


FIGURE 5 – Schéma des relations phylogénétiques de la famille des Brassicaceae et position de 13 évènements connus de mésotétraploïdie (étoiles blanches) et de mésohexaploïdie (étoiles noires) [Figure extraite de Mandáková et al. (2017)]. ① WGT spécifique au genre Leavenworthia (Haudry et al. 2013), ② WGT spécifique à la tribu des Brassiceae (Lysack et al. 2005) ③ WGD spécifique au genre Orychophragmus (Lysack et al. 2007). La topologie schématique de l'arbre est basée sur les reconstructions phylogénétiques publiées dans Franzke et al. (2011) et Huang et al. (2016) ; la classification en « clade/lineage » suit celles établies dans Huang et al. (2016), Franzke et al. (2011) et Al-Shehbaz (2012).

3. Conséquences évolutives associées aux WGD

3.1. Processus de diploïdisation

Dans la grande majorité des cas, les évènements de duplication de génome (auto- et allopolyploïdie) sont systématiquement suivis d'un processus dit de « diploïdisation » qui entraîne le retour de l'organisme à un état diploïde. Ce processus complexe implique des réarrangements génomiques (inversions, délétions, translocations, fusions de chromosomes), des pertes de gènes par pseudogénisation (mutation initiale provoquant une perte de fonction, suivie d'une accumulation de mutations sélectivement "neutres") et l'évolution de séquences codantes ou non-codantes conduisant à des néo-fonctionnalisations (un gène dupliqué acquiert une nouvelle fonction et est maintenu par la sélection naturelle tandis que l'autre copie du gène conserve sa fonction initiale) ou des sous-fonctionnalisations de gènes (un gène dupliqué n'assure plus sa fonction initiale mais seulement une partie) (Lynch and Conery 2000; Wendel 2000; Otto 2007; Tayalé and Parisod 2013; Panchy et al. 2016). De nombreuses études ont souligné le rôle central des éléments transposables dans les modifications structurelles, épigénétiques et fonctionnelles qui s'opèrent durant la polyploïdisation et le processus de diploïdisation subséquent (voir Parisod et al. (2010) et Vicient and Casacuberta (2017) pour une revue à ce sujet).

Durant le processus de diploïdisation, certains homéologues sont éliminés ou retenus en de façon aléatoire (Thomas et al. 2006; Edger and Pires 2009; Duarte et al. 2010). D'autres gènes dits sensibles à la dose (« dosage- sensitive ») ont tendance à résister au phénomène de « fractionation » (*i.e.* pertes de copies de gènes consécutives à un WGD) sous l'effet de la sélection purifiante, puisque les produits qu'ils encodent sont impliqués dans des interactions sensibles à la dose (« dosage-sensitive ») avec d'autres produits, créant un réseau de dépendance étroite (« dosage balance hypothesis », FIGURE 6, Birchler and Veitia (2007)). Les gènes qui sont préférentiellement retenus en plusieurs copies incluent ceux codant pour les composants du complexe de facteurs de transcription, du complexe de modification des protéines, des ribosomes, de la machinerie de transduction de signaux, de l'organisation cellulaire ainsi que de la réponse aux stimuli biotiques, abiotiques et hormonaux (Thomas et al. 2006; Edger and Pires 2009; Freeling 2009; Conant et al. 2014; Liu et al. 2014; McGrath et al. 2014; Moghe et al. 2014; Mandáková et al. 2017). En parallèle, certains gènes dont la fonction est très conservée parmi les Angiospermes, tels que les gènes codant pour des

fonctions chloroplastiques (photosynthèse) et mitochondriales (Duarte et al. 2010), mais aussi ceux impliqués dans la réparation de l'ADN, la recombinaison, et la réponse aux dommages de l'ADN, sont susceptibles de retourner préférentiellement à un statut de gène en simple copie (« single copy status », De Smet et al. (2013)). Les gènes présentant une redondance fonctionnelle et qui ne sont pas sensibles à la dose peuvent aussi retourner à un statut de gène en simple copie de façon purement aléatoire.

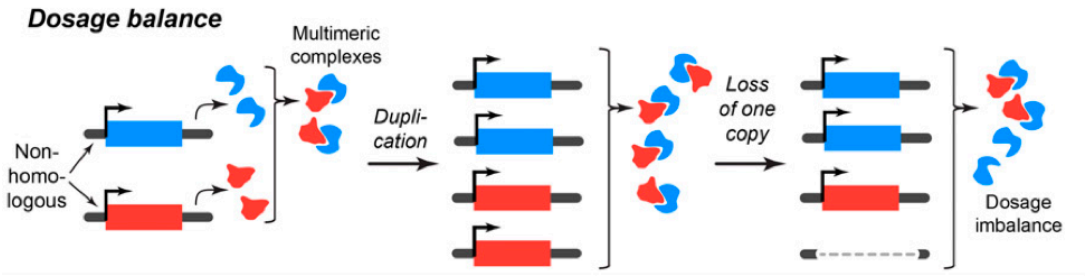


FIGURE 6 – « Dosage balance hypothesis » [Figure extraite de Panchy et al. (2016)]. Les gènes dupliqués « dosage sensitive » sont retenus dans les mêmes proportions, de façon à maintenir un équilibre stœchiométrique. Si une copie est perdue (à droite), il y a alors un déséquilibre stœchiométrique susceptible d'être contre-sélectionné (action de la sélection purifiante).

De plus, en ce qui concerne les allopolyploïdies, les pertes et rétentions de gènes qui font suite à un événement de duplication de génome dépendent non seulement de la fonction des gènes mais aussi du phénomène de « dominance » entre les différents sous-génomes qui ont fusionné (Garsmeur et al. 2014). En effet, chez certaines espèces telles qu'au sein des genres Brassica, Medicago, chez le coton, le maïs et le shorgo, il a été observé, à partir d'approches génomiques et transcriptomiques, qu'un des génomes parentaux semblait présenter un niveau d'expression et de conservation de gènes supérieur à l'autre génome parental. Ces phénomènes sont respectivement décrits sous le nom d'expression génique biaisée (« biased gene expression ») et de fractionnement génique biaisé (« biased gene fractionation ») (Chang et al. 2010; Woodhouse et al. 2010; Schnable et al. 2011; Wang et al. 2011; Cheng et al. 2012; Garsmeur et al. 2014; Renny-Byfield et al. 2015; Panchy et al. 2016; Steige and Slotte 2016). Il a été suggéré que ces biais d'expression et de perte de gènes sont spécifiques aux allopolyploïdes puisqu'ils n'ont pas été détectés chez les autopolyploïdes (Garsmeur et al. 2014). Les biais dans l'expression des gènes et dans la perte/rétention de gènes sont intimement liés car on s'attend à ce que les gènes localisés dans le sous-génome sous-exprimé et non dominant soient moins « sensibles » à la sélection purifiante agissant sur les altérations de séquences au sein des régions codantes ou non-codantes et sur les contraintes de dose (Schnable et al. 2011; Freeling et al. 2012; Steige and Slotte 2016). Ces

patrons de pertes et d'expression différentielles de gènes ont également été observés chez les amphibiens, notamment chez l'espèce allotétraploïde *Xenopus laevis* (Session et al. 2016).

Les mécanismes à l'origine de la dominance d'expression restent à ce jour très peu connus, et on ignore également si cette dernière opère immédiatement après l'événement d'allopolyploïdie ou si elle s'établit plus tardivement et progressivement dans le temps. Une étude très récente chez le genre *Mimulus* indique cependant que la dominance d'expression semble opérer instantanément après l'évènement d'hybridation allopolyploïde et qu'elle augmente significativement au cours des générations (Edger et al. 2017). Il a été démontré que l'expression des gènes peut être influencée par la proximité et le statut de méthylation des éléments transposables (TEs) voisins (Hollister and Gaut 2009). Après avoir observé une corrélation négative entre la densité en TEs méthylés et le niveau d'expression des gènes, Freeling et al. (2012) ont émis l'hypothèse que la relation entre la répression exercée par les TEs et le niveau d'expression des gènes voisins pourrait expliquer les patrons observés de dominance de génome, hypothèse qui a été tout récemment supportée par les résultats de Edger et al. (2017) chez le genre *Mimulus*. Un marquage épigénétique différentiel et héritable des sous-génomes parentaux expliquerait donc la dominance d'expression (Schnable et al. 2011). Woodhouse et al. (2014) ont également suggéré une régulation par le bas de l'expression des gènes par « silencing » lié à la présence d'éléments transposables locaux, suite à leur observation que les régions homéologues de Brassica ayant subi une perte de gène plus importante étaient enrichies en TEs et les gènes voisins présentaient des niveaux d'expression plus faibles par rapport à leur correspondant homéologue (voir aussi Vicient and Casacuberta (2017) et Edger et al. (2017)). Ainsi, selon cette hypothèse, le sous-génome présentant la charge en TE méthylés la plus faible devient dominant en terme d'expression et retient préférentiellement plus de gènes. A ce jour, les diverses études ne font état que de simples corrélations, et aucun lien de causalité n'a pu encore être précisément établi. De plus, Renny-Byfield et al. (2015) n'ont pas trouvé de corrélation entre le contenu en TEs et les biais dans la perte et l'expression des gènes chez le coton, suggérant ainsi que d'autres mécanismes peuvent contribuer aux biais d'expression et de perte de gènes. Bien que le mécanisme sous-jacent ne soit pas encore parfaitement connu, après un événement d'allopolyploïdie, les génomes parentaux se différencient à la fois en terme d'expression de gènes et de pertes de gènes, et parfois de manière particulièrement importante (Cheng et al. 2016).

Néanmoins, le phénomène de dominance génomique ne semble pas systématique chez les allopolyploïdes, comme reporté chez *Capsella bursa-pastoris* (Kasianov et al. 2017) ou encore chez le blé (Pfeifer et al. 2014). Il est possible que la dominance génomique nécessite plusieurs générations pour s'établir, puisqu'elle est très commune chez les espèces paléo et méso-allopolyploïdes tandis qu'elle est plus rare chez les allopolyploïdes synthétiques nouvellement formés (Woodhouse et al. 2014). Toutefois, Edger et al. (2017) ont trouvé que, chez le genre *Mimulus*, elle s'établissait immédiatement après la formation de l'hybride allopolyploïde. A ce jour, ces questions nécessitent de nouvelles investigations afin d'établir un lien de causalité et d'identifier les autres mécanismes à l'origine de la dominance d'expression.

3.2. Impact des WGD sur les taux de diversification des lignées végétales

Au regard des diverses modifications susmentionnées, les WGD génèrent indéniablement de la variation dans la taille, les fonctions et la structure des génomes des plantes et constituent sans aucun doute un processus non négligeable qui façonne l'évolution et la diversification des lignées végétales (Otto and Whitton 2000; Marhold and Lihová 2006; Doyle et al. 2008; Soltis et al. 2009; Soltis, Segovia-Salcedo, et al. 2014; Panchy et al. 2016; Landis et al. 2018). A ce jour, il reste difficile de savoir si la polyploïdie est généralement associée à une augmentation des taux de diversification, mais elle semble cependant associée à de nombreux évènements de spéciation (15% des évènements de spéciation chez les Angiospermes seraient accompagnés d'une augmentation de la ploïdie, Wood et al. 2009) et est largement distribuée dans l'arbre phylogénétique des plantes à fleurs, comme il l'a été mentionné plus en amont. En fait, la question se pose de savoir si les espèces polyploïdes sont abondantes à cause d'une augmentation de leurs taux de diversification provoquée par les évènements de duplication de génome (ce qui ne semble pas être le cas) ou simplement parce que la formation d'espèces polyploïdes est très fréquente. Les études indiquant des taux de diversification plus faibles chez les espèces polyploïdes par rapport aux espèces diploïdes semblent en contradiction avec l'observation de traces d'évènements de WGD dans le génome de la plupart des Angiospermes (Mayrose et al. 2011). Ces questions font à ce jour l'objet de discussions controversées (Soltis, Segovia-Salcedo, et al. 2014, Landis et al. 2018).

Une étude récente concernant cette question évoque une association possible entre les WGD et les radiations évolutives[1], avec toutefois un temps de latence entre l'événement de WGD et la radiation évolutive qui a lieu plus tardivement, après un (ou plusieurs) événement de dispersion, c'est ce qu'on appelle l'hypothèse de décalage temporel entre les WGD et les radiations (« Time-lag radiation hypothesis ») (Schranz et al. 2012). Sur les arbres phylogénétiques des espèces et à l'échelle des familles, cela se traduit, après l'événement de WGD, par la présence d'un groupe très riche en espèces et d'un groupe frère pauvre en espèces (FIGURE 7). Tank et al. (2015) ont trouvé un support statistique significatif pour une association non aléatoire entre l'augmentation des taux de diversification chez les Angiospermes et les évènements de duplication de génome (WGDs), bien que les augmentations des taux de diversification soient rarement associées parfaitement aux WGDs, mais les suivent très souvent après un petit temps de latence, ce qui corrobore tout à fait l'hypothèse du décalage temporel entre les WGDs et les radiations (« WGD radiation lag-time hypothesis »). Chez les Caryophyllaceae, Smith et al. (2017) ont trouvé que certains WGDs étaient suivis d'une augmentation dans les taux de diversification, bien qu'aucune corrélation statistique n'ait pu être effectuée. Chez les téléostéens, en revanche, cette hypothèse n'a pas été validée (Laurent et al. 2017). Toutefois, l'hypothèse que les évènements de WGDs seraient à l'origine de radiations évolutives est encore très préliminaire et des analyses supplémentaires sont nécessaires.

Chez les Brassicales, il semblerait que les WGDs soient à l'origine d'innovations majeures dans la chimie des glucosinolates, permettant d'échapper à l'herbivorie. De façon très intéressante, cette évolution chimique semble être étroitement associée aux radiations chez les Brassicales (Edger et al. 2015; Tank et al. 2015; Soltis and Soltis 2016). Finalement, on peut se demander si les WGDs ne sont pas plutôt étroitement associés à des adaptations et/ou des processus évolutifs plutôt qu'à une augmentation du taux de diversification, qui ne serait alors qu'une conséquence indirecte des WGDs (Edger et al. 2015, Smith et al. 2017).

[1] Evènements de spéciation successifs et rapides, à partir d'un ancêtre commun, aboutissant à un ensemble d'espèces caractérisées par une grande diversité écologique et morphologique.

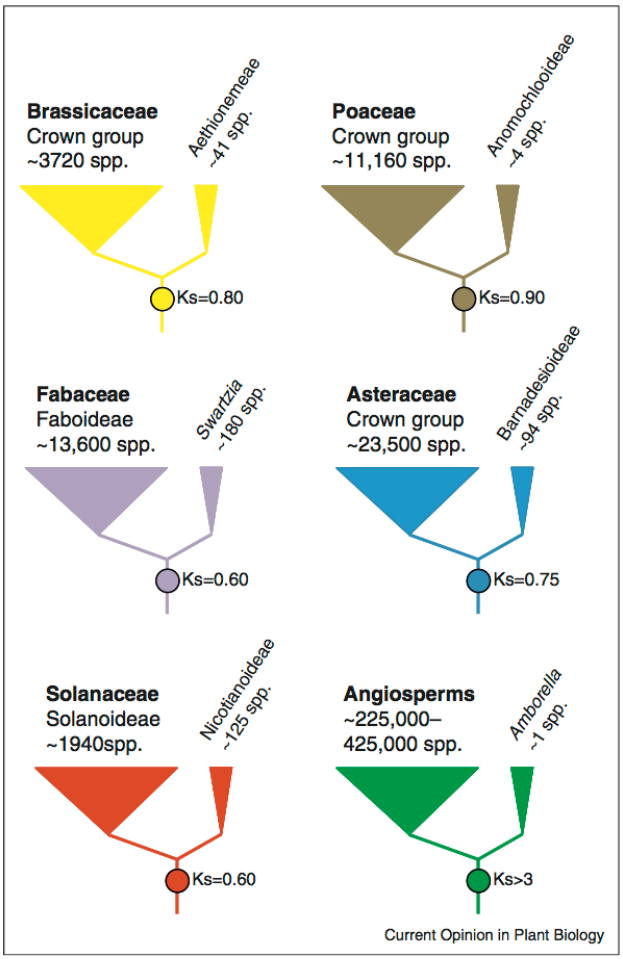


FIGURE 7 – WGD et arbres phylogénétiques asymétriques : la figure décrit, pour six lignées végétales ayant subit un WGD, un groupe riche en espèces suite à une radiation adaptative et son groupe frère pauvre en espèces [Figure extraite de Schranz et al. (2012)]. La phylogénie de chaque groupe au sein de chaque famille est une simplification de la phylogénie réelle. Le nombre d'espèces des groupes riches en espèces et des groupes frères pauvres en espèces est indiqué. Les cercles correspondent à la position potentielle des WGD et la divergence est présentée comme la valeur moyenne de *Ks* des paires de gènes dupliqués. Dans le modèle « WGD radiation lag-time », (1) un WGD se produit, (2) le WGD contribue à l'évolution de certains traits, (3) la divergence des lignées se produit et des évènements de diversification se produisent et persistent dans la région qui est supposée être le centre d'origine, (4) après plusieurs millions d'années, un ou plusieurs évènements majeurs de dispersion impliquant une des deux lignées se produisent, (5) la dispersion est à l'origine de la radiation à travers le globe actuellement observée chez le groupe riche en espèces, tandis que le groupe frère ne disperse pas et peut éventuellement subir des niveaux mineurs de radiation plus tardivement.

Plusieurs études ont suggéré que les patrons de distribution des WGD dans le temps ne sont pas aléatoires, et que la plupart ont lieu au cours de périodes prolongées de stress environnementaux tels que des conditions climatiques et écologiques instables (Fawcett et al. 2009; Kagale et al. 2014; Vanneste et al. 2014; Lohaus and Van de Peer 2016). En particulier, de nombreux WGDs indépendants tels que ceux identifiés chez le riz, *Medicago truncatula*, la tomate, la laitue (*Lactuca sativa*), le coton (*Gossypium hirsutum*), le peuplier et la banane semblent tous avoir eu lieu il y a environ 60 et 70 millions d'années. Il a ainsi été suggéré que ces évènements de duplication de génome sont associés à l'événement d'extinction de masse du Crétacé-Paléogène (nommé K-Pg) (Fawcett et al. 2009; Vanneste et al. 2014; Lohaus and Van de Peer 2016) qui est l'événement d'extinction de masse le plus récent (voir FIGURE 1. 2), entraînant l'extinction de 60 à 70% des espèces de plantes et d'animaux – dont les dinosaures – ainsi que des modifications et un réchauffement climatiques majeurs.

La polyploïdie présente à la fois des avantages et des inconvénients, mais il est probable que dans des conditions environnementales stables, elle soit plutôt désavantageuse à cause de la fitness réduite des individus polyploïdes liée à leur isolement reproducteur et leur plus faible fertilité (Comai 2005). Cependant, dans des conditions environnementales extrêmes et instables, les conséquences génétiques et fonctionnelles de la polyploïdie (augmentation de l'hétérozygotie, effet tampon des gènes redondants sur les mutations, néo-fonctionnalisation, autofécondation ou reproduction asexuée, etc.) pourraient conférer un avantage compétitif important aux polyploïdes (Comai 2005). C'est probablement en partie pour cette raison que la polyploïdie n'est pas toujours associée à des taux supérieurs de diversification des lignées, notamment lorsqu'on compare ce taux entre les espèces néopolyploides et les espèces diploïdes (Mayrose et al. 2011) et qu'en dépit de leur forte occurrence, seulement quelques rares polyploïdes – dans des conditions très spécifiques – survivent et se diversifient sur le long terme tandis que la majorité pourrait constituer des culs-de-sac évolutifs (Arrigo and Barker 2012; Van de Peer et al. 2017).

3.3. Conséquences évolutives des WGD sur les systèmes de reproduction

3.3.1. Introduction sur les systèmes d'auto-incompatibilité

Chez les Angiospermes, les plantes sont en grande majorité hermaphrodites, c'est-à-dire qu'elles produisent des fleurs dites bisexuelles ou hermaphrodites, chacune comportant à la fois les organes mâles et les organes femelles (Barrett 2002). Chez les végétaux hermaphrodites, il existe toute une gamme de régimes de reproduction avec deux extrêmes: l'autofécondation stricte et l'allofécondation (ou fécondation croisée) stricte. Entre ces deux extrêmes on retrouve un grand nombre d'espèces pratiquant un mode de reproduction dit « mixte », c'est-à-dire qu'elles pratiquent à la fois l'autofécondation et l'allofécondation en proportions variables (Goodwillie et al. 2005; Igic and Busch 2013). L'autofécondation présente de nombreux avantages sur le plan éco-évolutif tels que l'assurance reproductive (par exemple en l'absence de pollinisateurs animaux) ou l'avantage de transmission (l'individu parental unique transmet ses gènes à la fois par les voies mâles et femelles à chaque descendant), mais elle est associée à la dépression de consanguinité, reconnue comme la principale force sélective qui façonne l'évolution des stratégies de reproduction des plantes (Charlesworth and Charlesworth 1987; Goodwillie et al. 2005). Les systèmes d'auto-incompatibilité (SI pour « self-incompatibility »), recensés chez 100 familles d'Angiospermes et environ 40% des espèces, constituent les mécanismes génétiques les plus répandus et les plus diversifiés pour favoriser l'allofécondation et éviter l'autogamie (Igic et al. 2008; Ferrer and Good 2012; Nasrallah 2017). Selon la présence ou l'absence d'une variation dans la morphologie des fleurs associée au système SI, on distingue l'auto-incompatibilité dite hétéromorphe et l'auto-incompatibilité homomorphe (Barrett 2002). Parmi les systèmes SI homomorphes, au moins 3 modes d'action moléculaire complètement distincts ont été mis en évidence, ce qui démontre leur apparition multiple de manière indépendante dans l'histoire évolutive des Angiospermes (Takayama and Isogai 2005; Nasrallah 2017). Parmi ces systèmes SI homomorphes, la réponse auto-incompatible correspond soit à un mécanisme physiologique de reconnaissance du soi ou de reconnaissance du "non soi" entre le pistil et le pollen, qui aboutit dans les deux cas à l'inhibition du développement du tube pollinique en cas de pollinisation par de l'auto-pollen.

Chez la plupart des espèces auto-incompatibles, la reconnaissance du soi ou du non soi est gouvernée par un seul locus ou "supergène" multi-allélique, le locus S, et l'inhibition du développement du tube pollinique est initiée lorsque la même spécificité

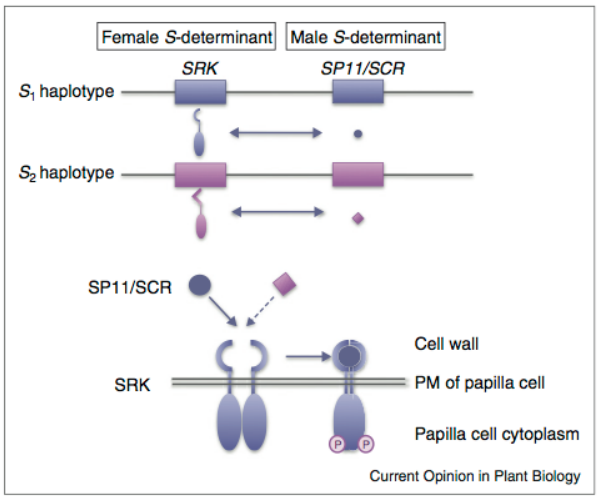
allélique est exprimée par le pollen et le pistil (Takayama and Isogai 2005; Nasrallah 2017). Malgré la diversité et la complexité des systèmes SI, tous contiennent au moins un déterminant génétique mâle et un déterminant génétique femelle représentés par des gènes distincts situés au locus S et fortement liés génétiquement. Lorsque les protéines du pollen et du pistil possèdent la même spécificité (lorsqu'un même allèle S est exprimé à la surface du pollen et du stigmate), une réaction d'inhibition du développement du tube pollinique est initiée par voie de signalisation et la fécondation devient alors impossible (Takayama and Isogai 2005; Nasrallah 2017). Classiquement, on distingue deux catégories de systèmes d'auto-incompatibilité homomorphe, les systèmes gamétophytiques (GSI) et sporophytiques (SSI), sur la base du déterminisme génétique du phénotype d'auto-incompatibilité du pollen. Le phénotype SI du pollen dans les systèmes GSI est déterminé par le génotype haploïde du pollen, tandis que dans les systèmes SSI il est déterminé par le génotype diploïde du sporophyte (tissus de l'anthère) qui l'a produit. En général les systèmes SI homomorphes sont caractérisés par une diversité allélique très élevée au locus S (Castric and Vekemans 2004) en raison de la sélection fréquence-dépendante négative qui agit en favorisant les nouveaux allèles mutants (Wright 1939). En effet, un nouvel allèle étant rare, les grains de pollen porteurs d'un tel allèle auront accès à un plus grand nombre de pistils compatibles dans la population que ceux portant des allèles déjà répandus: un individu portant un allèle rare au locus S aura donc accès via la fonction mâle à plus de partenaires compatibles. Dans les systèmes SSI, cela dépend néanmoins des relations de dominance et de récessivité entre les allèles.

3.3.2. Le système d'auto-incompatibilité sporophytique des Brassicaceae

Chez les Brassicaceae, présentant un système SSI, la compatibilité ou l'incompatibilité de deux partenaires sexuels dépendent des protéines de reconnaissance présentes à la surface du pollen et du stigmate et qui sont déterminées respectivement par les gènes multiallèliques *SP11*/*SCR* (pour « *S-locus protein 11* » et « *S-locus cysteine-rich protein* »; dans le reste de ce manuscrit j'utiliserai l'appellation *SCR*) et *SRK* (pour « *S-locus receptor kinase* »), localisés dans la même région génomique, appelée le locus S (FIGURE 8) (Takayama and Isogai 2005; Iwano and Takayama 2012; Nasrallah 2017). La protéine SRK des Brassicaceae est ancrée dans la membrane des cellules de la papille stigmatique de telle sorte que le domaine S extracellulaire pointe vers l'extérieur lui permettant d'interagir avec les molécules de ligand présentes à la surface du pollen (Nasrallah 2002). La protéine SCR, constituant cette molécule ligand produite par les tapetum de l'anthère, est quant à elle

déposée à la surface du pollen, et la diversité de sa séquence est très élevée (Shiba et al. 2002). Ces deux molécules fonctionnent comme un système clé-serrure largement diversifié. Les voies cellulaires qui conduisent à la réjection de l'auto-pollen impliquent un influx en ions calcium (Ca^{2+}) dans les cellules des papilles stigmatiques (Iwano et al. 2015). Deux molécules candidates ont été identifiées à ce jour en tant que médiateurs positifs de la voie de signalisation de la réaction d'auto-incompatibilité chez Brassica: ARC1 (« Arm-repeat-containing 1 ») et MLPK (« M locus protein kinase »), toutes deux localisées au niveau des papilles stigmatiques (Takayama and Isogai 2005; Ivanov et al. 2010). Le lien génétique étroit entre les composants pollen et stigmate du locus S implique que le complexe multigénique du locus S est hérité comme une seule et unique unité de ségrégation, et donc les variants de ce complexe génique sont appelés « haplotypes S » (Takayama and Isogai 2005). Chez Brassica et chez Arabidopsis, les analyses phylogénétiques des séquences des gènes *SRK* et *SCR* produisent des topologies presque identiques, ce qui démontre leur co-évolution sur plusieurs dizaines de millions d'années (Watanabe et al. 2000; Sato et al. 2002; Shiba et al. 2002; Goubet et al. 2012) et suggère l'absence de recombinaison au locus S. Une analyse comparative de l'organisation du locus S pour une dizaine d'haplotypes S chez *Arabidopsis halleri* et *A. lyrata* a montré qu'il s'agit bien d'une région non recombinante présentant une absence quasi-totale d'homologie de séquence entre haplotypes S, mis à part les séquences codantes des deux seuls gènes présents *SRK* et *SCR*, et couvrant une région dont la taille varie entre 30 et 110 kb (Goubet et al. 2012). Chez Brassica, on observe un troisième gène au sein du locus S, le gène *SLG*, qui est un paralogue hautement polymorphe de *SRK* dont il manque les domaines transmembranaire et kinase (Nishio and Kusaba 2000; Takasaki et al. 2000; Sato et al. 2002; Fobis-Loisy and Gaude 2004). Il semblerait qu'il n'ait pas de rôle indispensable dans la réaction d'auto-incompatibilité (Nishio and Kusaba 2000; Suzuki et al. 2000), bien qu'il semble l'améliorer (Takasaki et al. 2000). Les phylogénies des allèles *SRK* et *SLG* chez Brassica suggèrent qu'il existe des évènements de conversion génique fréquents entre ces deux gènes causant leur évolution concertée (Sato et al. 2002; Takuno et al. 2008). Plus de 30 et 50 haplotypes S ont été identifiés chez *B. rapa* (syn. *campestris*) et *B. oleracea*, respectivement (Nou et al. 1993; Ockendon 2000), et une soixantaine d'haplotypes S ont été recensés chez *A. halleri* (Vincent Castric, communication personnelle).

(A)



(B)

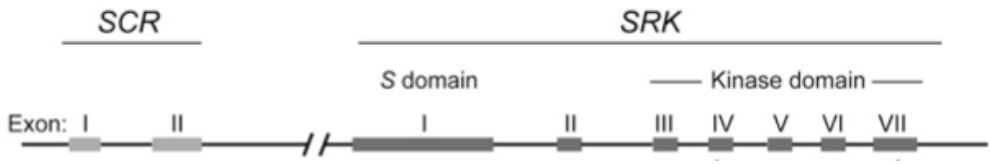


FIGURE 8 – A) Le système d'auto-incompatibilité des Brassicaceae [Figure extraite de Iwano and Takayama (2012)] et B) organisation schématique des gènes *SCR* et *SRK* du locus S chez les Brassicaceae [Figure extraite de Edh et al. (2009)]. A) Le locus S contient les déterminants femelle et mâle, *SRK* et *SCR*, respectivement. En cas d'autopollinisation, la liaison de la protéine ligand SCR à SRK (issus tous deux du même haplotype S) stabilise la protéine SRK en une forme dimérique active dans la membrane plastique, ce qui active la réponse d'auto-incompatibilité dans les cellules des papilles stigmatiques. B) Les deux exons du gène SCR et les 7 exons du gène SRK sont indiqués. Le gène SRK est constitué du domaine S (exon I), d'un domaine transmembranaire (exon II) et du domaine kinase (exon III à exon VII).

Il semblerait que le locus S soit apparu très tôt dans l'histoire évolutive des Brassicaceae, avant la divergence des différents clades majeurs, et que la diversification allélique à ce locus ait été très rapide. En effet, plusieurs allèles fonctionnels des gènes *SCR* et *SRK* ont été retrouvés dans différents genres de Brassicaceae, pour certains très divergents, tels que Brassica (Sato et al. 2002), Raphanus (Lim et al. 2002), Arabidopsis (Castric and Vekemans 2007), Capsella (Paetsch et al. 2006; Guo et al. 2009), Arabis (Tedder et al. 2011) ainsi que Biscutella (Leducq et al. 2014). Le polymorphisme ancestral (polymorphisme trans-spécifique) est partagé non seulement entre espèces mais aussi entre genres lointainement apparentés (Castric and Vekemans 2004). Autrement dit, certaines paires d'allèles échantillonnées dans des espèces différentes de Brassicaceae sont plus étroitement apparentées que des paires échantillonnées au sein d'une même espèce. Cependant, les niveaux de divergence moléculaire entre allèles peuvent fortement varier selon les clades de

Brassicaceae, avec par exemple une très large diversité phylogénétique entre allèles (à noter que le locus S est non-recombinant ce qui valide la réalisation de phylogénie d'allèles) pour les allèles *SRK* des genres Arabidopsis et Capsella, alors que les allèles *SRK* des genres Brassica et Raphanus se regroupent en seulement deux groupes phylogénétiques distincts (dénommés les allèles de classe I et allèles de classe II; Uyenoyama 1995; Edh et al. 2009) (FIGURE 1.9). De même, dans le genre Biscutella, les allèles *SRK* (voir FIGURE 9) se regroupent en seulement trois clades distincts, qui par ailleurs sont différents des deux clades identifiés chez Brassica et Raphanus (Leducq et al. 2014).

Comme évoqué précédemment, dans les systèmes SSI, il existe des relations de dominance entre les haplotypes S, avec généralement un seul des deux allèles exprimés pour le gène *SCR* dans les génotypes hétérozygotes, et des relations de dominance ou de co-dominance au gène *SRK* (Llaurens et al. 2008). On s'attend à ce que les relations de dominance affectent la fréquence populationnelle des haplotypes S puisque la force de sélection fréquence-dépendante négative typiquement observée dans les systèmes d'auto-incompatibilité est plus intense sur les haplotypes dominants que sur les haplotypes récessifs (Schierup et al. 1997). En effet, les allèles récessifs peuvent être présents dans plus de classes phénotypiques que les dominants, or la sélection fréquence-dépendante négative tend à équilibrer les fréquences phénotypiques et non pas génotypique; ainsi les allèles récessifs peuvent atteindre des fréquences plus élevées en population (voir Billiard et al. (2007). Chez Arabidopsis, les relations de dominance entre allèles S du gène *SCR* semblent très complexes et impliquent un très grand nombre de niveaux de dominance (Prigoda et al. 2005; Llaurens et al. 2008; Durand et al. 2014). En revanche, chez Brassica, deux classes majoritaires de dominance ont été trouvées dans le pollen : les allèles dominants dits de classe I et les allèles récessifs dits de classe II, chaque classe d'allèles correspondant à un des deux clusters phylogénétiques d'haplotypes S (voir FIGURE 9) (Nasrallah et al. 1991; Nasrallah and Nasrallah 1993; Hatakeyama et al. 1998; Shiba et al. 2002; Kakizaki et al. 2003). Dans le pistil, chez Brassica, les allèles sont presque toujours codominants (Hatakeyama et al. 1998). En accord avec les attendus théoriques, les allèles récessifs chez Brassica sont effectivement présents en fréquence plus élevée que les allèles dominants dans les populations naturelles (Glémin et al. 2005).

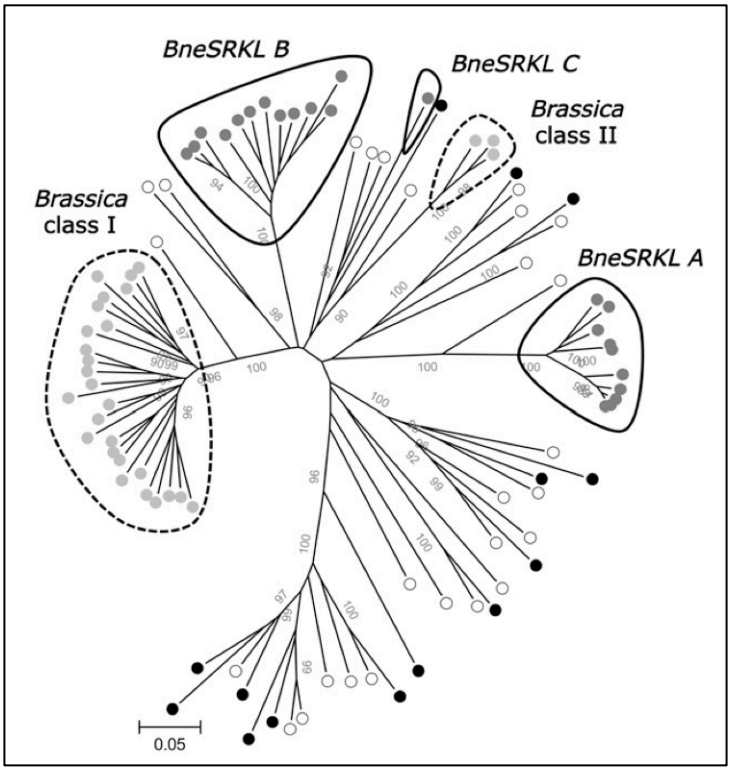


FIGURE 9 – Relations phylogénétiques entre les séquences SRK et SRK-like (SRKL) chez quatre espèces de Brassicaceae [Figure extraite de Leducq et al. (2014)] : *Capsella grandiflora* (noir), *Brassica oleracera* (gris), *Arabidopsis lyrata* (blanc), and *Biscutella neustriaca* (gris foncé). Les trois clades de séquences identifiés chez Biscutella (ligne continue, BneSRKL A, B et C) sont différents des deux clades connus de Brassica (ligne en pointillée, Brassica class I et II).

De façon surprenante, il a été montré que les relations de dominance entre allèles du gène *SCR* sont en partie régies par des petits ARN (small RNA, sRNA) qui régulent l'expression de manière allèle-spécifique. Chez des individus hétérozygotes classe I/classe II au gène *SCR* de Brassica, il a été trouvé que les régions promotrices des allèles récessifs (classe II) sont méthylées dans le tapetum des anthères, ce qui empêche leur transcription (Shiba et al. 2006). Les déterminants génétiques de ce processus de méthylation allèle-spécifique, et donc de l'inhibition de l'expression des allèles *SCR* récessifs, sont constitués par des petits ARN (appelés SMI pour SCR METHYLATION INDUCER) dont les précurseurs sont localisés dans la région flanquante des allèles *SCR* dominants (FIGURE 10), et qui sont exprimés dans le tapetum des anthères juste avant l'initiation de la transcription de *SCR* (Tarutani et al. 2010). Yasuda et al. (2016) ont identifié un autre petit ARN (SMI2) localisé en aval du gène *SRK* dans les haplotypes de classe II (absent des haplotypes de classe I), cette fois impliqué dans la régulation des relations de dominance entre les différents allèles de classe II, qui présentent des relations de dominance linéaire (FIGURE 10) (Kakizaki et al. 2003).

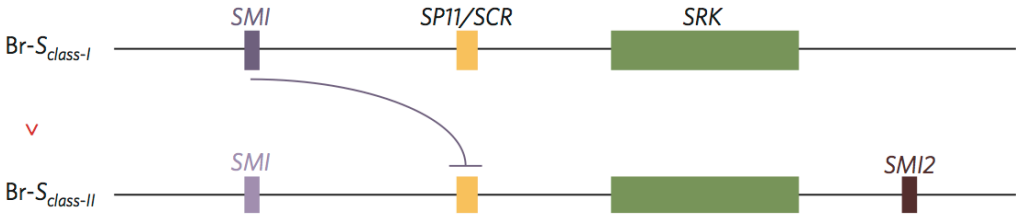


FIGURE 10 – Chez Brassica, le petit ARN (SMI) présent dans la région flanquante du gène *SCR* des haplotypes dominants de classe I (Br-$S_{class-I}$) cible la région promotrice du gène *SCR* de l'haplotype récessif de classe II (Br-$S_{class-II}$) et provoque sa méthylation, ce qui débouche sur une inhibition de l'expression du gène. SMI2 est localisé en aval du gène *SRK* seulement chez les haplotypes de classe II et cible la région promotrice du gène *SCR* des autres haplotypes de classe II qui lui sont récessifs, ce qui provoque l'inhibition de l'expression de ces derniers [Figure extraite de Goring (2016)].

3.3.3. Pertes des systèmes d'auto-incompatibilité

Comme mentionné précédemment, les systèmes d'auto-incompatibilité homomorphe sont largement distribués parmi les Angiospermes et sont apparus plusieurs fois indépendamment. Malgré leur très large distribution, il apparaît que les évènements de perte du système d'auto-incompatibilité (et la transition progressive vers l'autogamie qui s'ensuit) sont bien plus fréquents que les gains et sont considérés comme la transition de système de reproduction la plus fréquente dans l'évolution des plantes à fleurs (Igic et al. 2008). La prépondérance des évènements de perte du SI par rapport au gain de SI est considérée comme liée à deux causes principales. Premièrement, lorsque qu'un système SI devient non fonctionnel et que la population évolue vers un régime de reproduction autogame, on s'attend à ce que le polymorphisme au locus S s'effondre au cours du temps. Cette perte de polymorphisme peut être due à l'effet de la dérive génétique suite au relâchement important de la sélection fréquence-dépendante négative dans le cas où la mutation affecte un gène non lié au locus S mais agissant dans sa régulation ou dans la voie de signalisation du SI (modificateur [2]), ou à l'effet de la sélection positive qui peut s'exercer sur un allèle S mutant non-fonctionnel (conférant l'auto-compatibilité) qui se retrouverait favorisé dans certaines conditions (Vekemans et al. 2014). La perte du polymorphisme allélique au locus S constituerait un frein pour un retour vers un système fonctionnel, même si les conditions écologiques favorisant temporairement l'autofécondation sont interrompues. Deuxièmement, tous les SI homomorphes documentés à ce jour reposent sur le fonctionnement de plusieurs

[2] Gène non lié au locus S, mais impliqué dans les processus biochimiques sous-jacents de la réponse auto-incompatible.

gènes, dont certains sont non directement liés au locus S et à la reconnaissance pollen/pistil (notamment des gènes de signalisation). Des mutations supplémentaires induisant une perte de fonction sont susceptibles de s'accumuler dans ces gènes après la perte du SI, ce qui rend le retour du SI encore plus difficile et improbable. Chez les Solanaceae, le taux de perte du SI a été estimé comme environ 70 fois plus important que le taux de gain du SI, suggérant que les pertes du SI peuvent être considérées comme étant irréversibles (Igic et al. 2008). Chez les Brassicaceae, un second système SSI non homologue au système *SCR*/*SRK* a été identifié dans le genre Leavenworthia (Chantha et al. 2013). Les gènes impliqués dans la reconnaissance pollen/pistil dans ce genre (appelés respectivement *SCRL* et *LaLal2*) appartiennent aux mêmes familles fonctionnelles que *SCR* et *SRK*, respectivement, mais constituent néanmoins une évolution indépendante du SI. Ceci indique qu'au sein d'une famille, plusieurs évolutions indépendantes du SI sont possibles, mais néanmoins les évènements de perte d'un SI fonctionnel sont nettement plus fréquents que les évènements de gain (Vekemans et al. 2014).

De nombreuses études chez des espèces auto-incompatibles ont trouvé des individus ou des populations devenus récemment auto-compatibles, et il apparaît même que la perte du SI a eu lieu plusieurs fois indépendamment au sein de certaines espèces (Igic et al. 2008, Vekemans et al. 2014). Chez les Solanaceae (2600 espèces), 40% des espèces sont auto-incompatibles et présentent un polymorphisme trans-spécifique au locus S, suggérant une apparition unique du SI chez l'ancêtre commun des espèces analysées. Parmi ces espèces, le SI a été perdu au moins 60 fois indépendamment (FIGURE 11) (Igic et al. 2006; Goldberg et al. 2010).

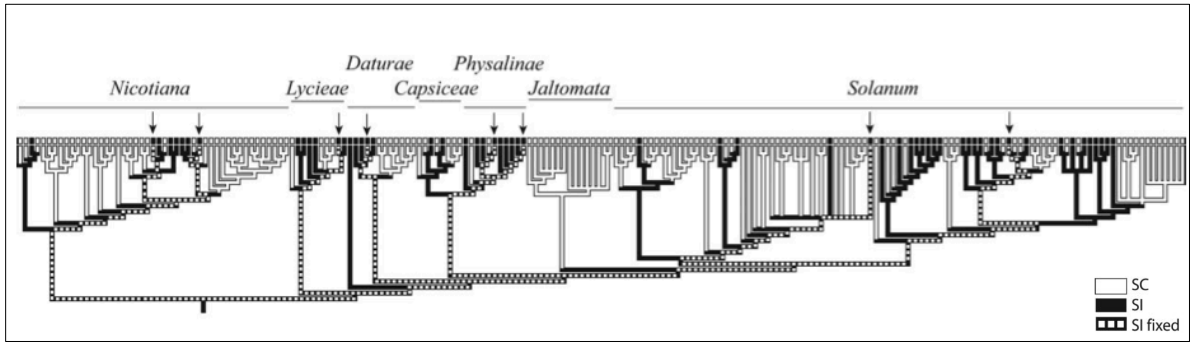


FIGURE 11 – Relations phylogénétiques entre 202 taxa de Solanaceae et système de reproduction (SI versus SC) [Figure extraite de Igic et al. (2006)]. L'auto-incompatibilité (carrés pleins) et l'auto-compatibilité (carrés vides) sont indiqués à l'extrémité de chaque branche. On remarque un entremêlement des deux états le long de la phylogénie. Huit taxa auto-incompatibles (indiqués par une flèche) présentent un polymorphisme ancestral partagé au locus S. Le chemin évolutif depuis l'ancêtre commun jusqu'à chacun de ces huit taxa est assigné à l'état SI (en pointillé).

Une façon de déterminer si les espèces ou les populations sont auto-compatibles consiste à analyser la diversité allélique au locus S. En effet, à cause de la sélection fréquence-dépendante négative, on s'attend à retrouver un grand nombre d'allèles différents chez les espèces ou les populations SI et un seul ou peu d'allèles chez les espèces ou les populations SC (pour "Self-Compatible", FIGURE 12) (Vekemans et al. 2014; Mable et al. 2017). Selon le nombre et l'identité des allèles retrouvés chez les espèces ou les populations SC, on peut estimer l'âge relatif de la perte du SI, s'il y a eu plusieurs événements de perte indépendants, et même dans certains cas, identifier la nature de la mutation causale associée à la perte du SI (Vekemans et al. 2014). Lorsqu'un seul et unique allèle S non fonctionnel est retrouvé dans une espèce ou une population SC (comme chez *Capsella rubella* et *Leavenworthia alabamica-race a4*), cela suggère que la mutation est apparue au locus S et qu'elle s'est rapidement fixée dans la population, ou bien que la mutation a eu lieu ailleurs (dans un modificateur) mais il y a très longtemps, laissant suffisamment de temps à la dérive génétique pour éliminer tout le polymorphisme ancestral au locus S (Vekemans et al. 2014). Au contraire, lorsque plusieurs allèles S sont retrouvés chez une espèce ou une population SC (comme chez *Arabidopsis thaliana*), cela suggère que plusieurs mutations indépendantes causant une transition vers SC ont eu lieu, ou bien qu'un modificateur a causé une transition vers SC et que la dérive génétique a érodé partiellement le polymorphisme ancestral au locus S. Pour choisir entre l'une ou l'autre des deux explications possibles, il faut déterminer si la mutation initiale causant la transition vers SC est liée au locus S ou non.

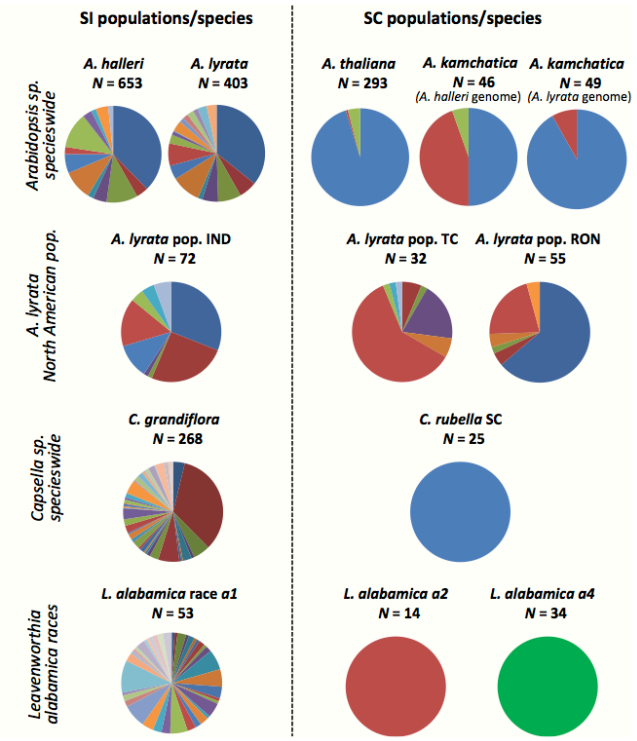


FIGURE 12 – Distribution des fréquences alléliques au locus S chez des espèces ou des populations auto-incompatibles (SI) et auto-compatibles (SC) chez les Brassicaceae [Figure extraite de Vekemans et al. (2014)]. N : nombre d'individus génotypés dans chaque échantillon. Pour l'espèce allotétraploïde *A. kamchatica*, les fréquences alléliques sont reportées séparément pour les deux génomes parentaux, *A. lyrata* et *A. halleri*.

En effet, chez les Brassicaceae, des mutants auto-compatibles ont été identifiés comme étant causés par une mutation non fonctionnelle dans *SCR*, dans *SRK*, ou dans un modificateur (Tsuchimatsu and Shimizu 2013). Néanmoins, il est difficile d'identifier a posteriori la ou les mutations initialement responsables de la perte du SI puisque dès lors que le SI est perdu, des mutations peuvent s'accumuler dans une partie ou l'ensemble des gènes associés au phénotype d'auto-incompatibilité. Chez *Arabidopsis thaliana* (Brassicaceae), à 95-99% autogame, plusieurs haplotypes S non-fonctionnels présentant une forte similarité de séquence avec des allèles fonctionnels de *A. lyrata* et *A. halleri* ont été identifiés (Shimizu et al. 2004; Bechsgaard et al. 2006). Une étude menée à l'échelle de l'espèce (1088 accessions) sur des données de re-séquençage génomique a montré que ces haplotypes S présentent des mutations différentes dans *SCR* et/ou *SRK* générant la perte de fonction des gènes, des délétions partielles ou complètes de ces gènes dans certains haplotypes, et même dans certains cas, des réorganisations de la structure du locus S et peuvent même résulter d'évènements de recombinaison entre haplotypes S différents (Nasrallah 2017; Tsuchimatsu et al. 2017). En outre, la découverte d'une délétion fixée dans un modificateur de l'auto-incompatibilité (gène

ARC1) chez cette espèce pourrait suggérer que la perte de fonction du locus S n'est pas nécessairement la première étape de la transition vers l'autogamie (Liu et al. 2007). De même, chez *A. lyrata*, une espèce proche auto-incompatible présentant quelques populations auto-compatibles, il a été suggéré suite à des croisements SI × SC que la transition vers l'autogamie dans certaines populations est associée à un modificateur de l'expression de l'haplotype S1, récessif et non lié au locus S, plutôt qu'à des mutations dans les séquences *SCR* ou *SRK* (Mable et al. 2017).

L'observation de cas fréquents de perte du SI pose question quant aux conditions dans lesquelles elle s'opère et se maintient. La limitation en disponibilité de partenaires sexuels compatibles constitue un des principaux moteurs des transitions SI vers SC (Goodwillie et al. 2005; Busch and Schoen 2008; Vekemans et al. 2014). Plusieurs facteurs peuvent être responsables de la limitation en partenaires sexuels compatibles dans les taxa SI, comme une réduction du nombre d'haplotypes S (sous l'effet par exemple de goulots d'étranglement liés à la fragmentation des populations en limite d'aire de répartition, ou d'effets de fondation induits par des évènements d' allopolyploïdie, voir section 3.3.4.) ou encore une raréfaction des vecteurs de pollinisation. Quand le nombre de partenaires compatibles devient limité, la proportion d'ovules fécondés diminue d'autant plus que les allèles au locus S portés par un individu sont fréquents dans la population, et la production de graines par plante sera plus importante pour les individus présentant les génotypes les plus rares, c'est la composante "voie femelle" de la sélection fréquence-dépendante négative agissant sur le locus S (Vekemans et al. 1998). En condition de limitation en partenaires compatibles, la sélection pour l'assurance reproductive pourrait favoriser les mutations conférant un phénotype SC et celles-ci pourraient atteindre de fortes fréquences, notamment si la dépression de consanguinité est modérée (Goodwillie et al. 2005; Busch and Schoen 2008; Castric et al. 2014). Le succès des génotypes SC dépend donc de plusieurs paramètres déterminant si les avantages reproductifs associés à l'auto-fécondation peuvent ou non contrebalancer les effets négatifs associés à la dépression de consanguinité (Igic and Busch 2013). A terme et sous certaines conditions, les mutations SC peuvent totalement envahir la population ou se maintenir à une fréquence intermédiaire sous la forme d'un système de reproduction mixte (Stone 2002).

3.3.4. Perte des systèmes d'auto-incompatibilité et allopolyploïdie

Comme évoqué précédemment, la limitation en partenaires sexuels compatibles constitue un terrain favorable pour la transition SI vers SC. Or, chez les espèces allopolyploïdes nouvellement formées, un fort isolement reproducteur est attendu, ce qui peut favoriser une transition rapide vers un SC par sélection pour l'assurance reproductive (Soltis, Visger, et al. 2014; Vekemans et al. 2014). L'existence d'une association phylogénétique entre la perte du SI et l'évolution de la polyploïdie a longtemps été discutée et a fait l'objet de conclusions contradictoires (voir Stone 2002; Mable 2004b; Barringer 2007). Une revue de littérature sur l'association entre la polyploïdie et l'auto-incompatibilité (Miller and Venable 2000) incluant des polyploïdes synthétiques et naturels a trouvé que la transition entre l'état diploïde et l'état polyploïde était associée à une transition de SI à SC chez 70% des espèces étudiées, et que cette association est plus forte pour les espèces avec un SI gamétophytique que pour celles avec un système sporophytique. Il a également été montré que les polyploïdes tendent à avoir des taux d'autofécondation plus élevés que leurs confrères diploïdes (Barringer 2007). Cela peut facilement s'expliquer par le fait que les individus auto-compatibles des espèces polyploïdes ne devraient pas souffrir de la limitation en partenaires compatibles (ici il faut considérer les partenaires de même niveau de ploïdie) puisqu'ils peuvent s'autoféconder (Mable et al. 2004b), permettant ainsi aux populations polyploïdes nouvellement formées de s'établir en conditions de faible densité. Les niveaux de dépression de consanguinité réduits attendus chez les polyploïdes, dus à la redondance fonctionnelle assurée par les paralogues (Comai 2005), peuvent expliquer pourquoi les polyploïdes tendent à avoir des taux d'auto-fécondation plus élevés que leurs congénères diploïdes. Toutefois, le lien entre la polyploïdie et des niveaux réduits de dépression de consanguinité n'est pas encore parfaitement établi (Barringer 2007).

Il semblerait que la polyploïdie conduise souvent à l'auto-compatibilité dans les lignées SI, surtout dans les systèmes GSI (Ramsey and Schemske 1998; Stone 2002; Robertson et al. 2010), mais cela ne signifie évidemment pas pour autant que les transitions SI vers SC s'expliquent toutes par la polyploïdie. Beaucoup de cas de perte du SI ont été identifiés sans qu'aucun événement de polyploïdie n'ait été rapporté (voir section 3.3.3, Robertson et al. 2010). En revanche, la polyploïdie, en exposant le ou les nouveaux individus polyploïdes formés à pas ou peu de partenaires sexuels compatibles (à la fois à cause d'un niveau de ploïdie différent de celui des individus parentaux, d'un nombre d'allèles S très réduits par effet de fondation, et de changements physiologiques et/ou morphologiques qui

suivent la polyploïdisation, voir Soltis, Visger, et al. 2014), devrait contribuer à une évolution rapide de l'auto-compatibilité. La moindre mutation cassant le SI devrait tout naturellement être favorisée par la sélection naturelle en permettant aux polyploïdes nouvellement formés de s'établir. La perte du SI peut être partielle ou complète, temporaire ou irréversible. Toutefois, les modalités de recouvrement du SI après sa perte restent à ce jour mal comprises.

A l'heure actuelle, chez les Brassicaceae, quelques cas de polyploïdie récente associée à une perte du SI ont été documentés. L'espèce allotétraploïde et auto-compatible *Arabidopsis kamchatica* résulte de l'hybridation allopolyploïde entre deux espèces diploïdes auto-incompatibles, *A. halleri* et *A. lyrata,* associée à une perte du SI (Tsuchimatsu et al. 2012). Il y a eu vraisemblablement plusieurs évènements indépendants d'hybridation allopolyploïde puisque cinq allèles S distincts et non fonctionnels – existants chez les espèces parentales – ont été retrouvés chez l'allotétraploîde *A. kamchatica.* Chez l'espèce allotétraploïde et auto-compatible *Arabidopsis suecica* résultant de l'hybridation allopolyploïde entre *A. arenosa* (SI) et *A. thaliana* (SC), le locus S hérité du génome de *A. thaliana* est fixé pour une inversion dans *SCR* également rencontrée chez *A. thaliana* et supposée être à l'origine de la perte du SI chez cette dernière, tandis que le locus S hérité du génome de *A. arenosa* contient un allèle proche phylogénétiquement d'un allèle rencontré chez *A. halleri*, récessif par rapport à l'allèle hérité de *A. thaliana* (Novikova et al. 2017). Il a été suggéré que l'allèle initialement fonctionnel issu du parent *A. arenosa* a été rendu silencieux au niveau transcriptionnel par l'allèle du locus S hérité d'*A. thaliana* (Novikova et al. 2017). Le petit ARN issu du parent *A. thaliana* et sa cible sur l'allèle issu du parent *A arenosa* impliqués dans ce mécanisme de dominance ont en effet été supposés fonctionnels chez *A. suecica* (Novikova et al. 2017). Ainsi, il est fort probable que *A. suecica* soit devenue immédiatement SC (au moins partiellement SC) puisque l'allèle S non fonctionnel hérité de *A. thaliana* est dominant sur l'allèle S fonctionnel hérité de *A. arenosa*, et qu'une mutation induisant une perte de fonction s'est fixée plus tard dans l'allèle de *A. arenosa* (Novikova et al. 2017). Les espèces allotétraploïdes SC *Capsella bursa-pastoris* et *Cardamine flexuosa* sont également issues d'une hybridation allopolyploïde entre une espèce diploïde SI et une espèce diploïde SC (*Capsella grandiflora* × *C. orientalis* et *Cardamine amara* × *C. hirsuta*, respectivement) (Mandáková et al. 2014; Douglas et al. 2015). Enfin, chez les Brassiceae, les espèces parentales de l'allotétraploïde auto-compatible *Diplotaxis muralis* sont *a priori D. tenuifolia* qui est auto-incompatible et *D. vimenea* qui est auto-compatible (voir Marhold and Lihová 2006 et les références associées), suggérant là encore qu'une perte du SI fait suite à

une hybridation allopolyploïde de type SC × SI. Cependant, tous ces exemples concernent des espèces polyploïdes relativement récentes (<100 000 ans) dont les espèces parentales sont encore actuelles, et il est bien plus difficile d'analyser le cas des espèces mésopolyploïdes dont les espèces parentales sont bien souvent éteintes et dont on peut difficilement déterminer le système de reproduction.

Dans le genre Leavenworthia, l'analyse du polymorphisme au locus S chez des espèces SI et des espèces SC suggère que l'ancêtre commun des espèces de ce genre était auto-incompatible et a subi un fort goulot d'étranglement génétique qui a été suivi d'une apparente re-diversification allélique au locus S (Herman et al. 2012). Plus tard, il a été montré que le système d'auto-incompatibilité chez Leavenworthia n'est pas homologue au système SI des autres Brassicaceae (Chantha et al. 2013), et qu'il a donc évolué *de novo* (à partir de gènes paralogues non polymorphes) après la perte du SI impliquant *SRK*/*SCR*. Il est connu que l'ancêtre commun de ce genre a subi un événement de mésopolyploïdie (WGT) (Haudry et al. 2013), et il est tentant de supposer, par analogie aux cas récents d'allopolyploïdie décrits ci-dessus, que la perte du SI homologue au système SI des Brassicaceae pourrait être associée à cet évènement. Chez les genres Brassica et Raphanus, le regroupement des allèles *SRK* et *SCR* en seulement deux classes alléliques (voir section 3.3.2) suggère un évènement de goulot d'étranglement majeur, qui pourrait être associé avec l'évènement ancien d'allohexaploïde (WGT) partagé par les membres de la tribu des Brassiceae (suite à deux événements successifs d'hybridation allopolyploïde). Il est possible que le SI ait été partiellement ou totalement perdu de façon temporaire avant d'être regagné et donc avant la re-diversification allélique au locus S. Cette re-diversification allélique récente, sous l'effet de la sélection fréquence-dépendante négative associée au système SI, est démontrée par les reconstructions phylogénétiques et également par l'observation de signatures très fortes de sélection positive dans le domaine extracellulaire du gène SRK chez Brassica (Castric and Vekemans 2007). De manière analogue, le goulot d'étranglement au locus S observé dans le genre Biscutella (Leducq et al. 2014, FIGURE 9) pourrait également avoir été causé par un évènement de mésopolyploïdie (WGD) car un tel évènement dans l'histoire du genre Biscutella a été mis en évidence récemment (Geiser et al. 2016). Les évènements d'allopolyploïdie provoqueraient donc la perte de fonctionnalité du système SI et la production de lignées SC évoluant vers l'autogamie, mais dans certains cas, une nouvelle mise en place du système SI (ou l'évolution d'un système SI non homologue) accompagnée néanmoins d'une forte réduction de la diversité allélique au locus S semble possible.

4. Modèle d'étude : la « tribu du chou », les Brassiceae

4.1. Caractéristiques générales

Les Brassiceae forment une tribu monophylétique du clade B (anciennement lignée II) de la famille des Brassicaceae, proche des tribus Thlaspideae, Eutremeae, Sisymbrieae, et Isatideae (Huang et al. 2016; Guo et al. 2017). Cette tribu comporte 47 genres et 227 espèces (Al-Shehbaz 2012). Les membres de cette tribu se distinguent des autres Brassicaceae par le fait qu'ils possèdent des cotylédons pliés longitudinalement le long de la radicule et/ou des fruits segmentés transversalement (fruits hétéroarthrocarpiques, FIGURE 13), caractéristiques absentes chez les autres Brassicaceae (Gómez-Campo 1980; Al-Shehbaz 1985; Warwick and Sauder 2005; Hall et al. 2011). La majorité des genres se distinguent seulement par des différences mineures dans la structure des fruits (Al-Shehbaz 2012).

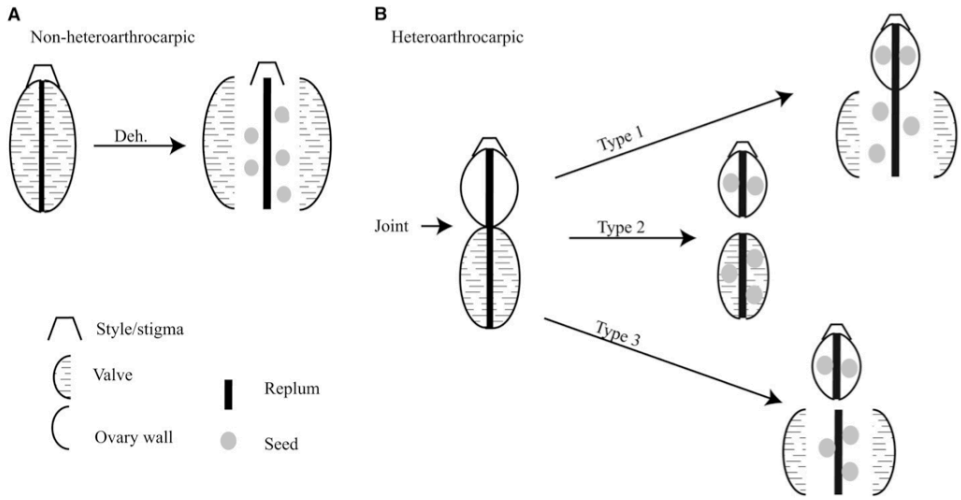


FIGURE 13 – Maturation des fruits et dispersion des graines chez les Brassiceae [Figure extraite de Hall et al. (2011)]. (A) Les taxa avec des fruits non-hétéroarthrocarpiques dispersent leurs graines librement dans l'environnement tandis que (B) les taxa avec des fruits hétéroarthrocarpiques libèrent leurs graines de trois manières différentes. Le fruit est segmenté en deux parties (partie distale et partie proximale) par le joint. Lorsque la partie distale se sépare et se disperse indépendamment de la partie proximale qui elle reste attachée au pied maternel, on parle de désarticulation ou encore d'abscission. Type 1 = le segment proximal est déhiscent, il n'y a pas de désarticulation. Type 2 = fruit complètement indéhiscent, avec une désarticulation. Type 3 = le segment proximal est déhiscent, avec une désarticulation. Replum : tissus du placenta qui persiste. Les fruits hétéroarthrocarpiques sont retrouvés chez 27 des 47 genres de Brassiceae (Warwick and Sauder 2005).

Les membres de la tribu des Brassiceae sont principalement distribués le long du pourtour Méditerranéen, avec quelques représentants en Europe, en Asie du Sud Ouest, dans le centre et le Sud de l'Afrique ainsi que sur la côte Est de l'Amérique du Nord (Al-Shehbaz 2012). Le centre d'origine de cette tribu ne correspond pas à son centre de diversité, puisqu'il

a été trouvé qu'elle a émergé il y a environ 24 millions d'années depuis la région saharo-sindienne (une ceinture de désert qui s'étend du Nord Ouest de l'Afrique au Nord Ouest de l'Inde) (Arias et al. 2014). Cette tribu est particulièrement connue pour ses représentants consommés ou utilisés à des fins ornementales. En effet, c'est au sein de la tribu des Brassiceae que l'on rencontre la centaine de variétés de choux cultivés parmi lesquelles figurent en tête de liste le chou pommé, le chou de Bruxelles, le chou brocoli, le chou rouge, le chou-fleur, etc.. Le chou sauvage, encore appelé « chou des falaises » (*Brassica oleracea*), est l'ancêtre commun de presque tous les choux cultivés aujourd'hui. *Brassica rapa* est l'ancêtre de plusieurs variétés potagères telles que le pak-choï, le pe-tsaï (choux chinois) et le navet et est proche phylogénétiquement de plusieurs autres espèces fourragères, oléagineuses ou servant à la préparation de condiments (les navettes, la moutarde, etc.). Le colza (*Brassica napus*) est un hybride allotétraploïde naturel entre les espèces *B. oleracea* et *B. rapa*. Le chou maritime (*Crambe maritima*) est une espèce plus éloignée du genre Brassica mais également consommée pour ses tiges.

La tribu des Brassiceae comporte les membres de la famille des Brassicaceae les plus importants économiquement (Al-Shehbaz 1985; Al-Shehbaz 2011). Brassica et Raphanus fournissent beaucoup de légumes qui sont cultivés pour leurs racines charnues, leurs tiges gonflées, leurs feuilles, leurs bourgeons, leurs fleurs et leurs jeunes fruits. Des huiles comestibles et industrielles sont extraites des graines de Brassica, de Crambe et de Eruca, tandis que des condiments sont obtenus à partir des graines de Sinapis et de Brassica. En Asie, beaucoup de propriétés médicinales sont attribuées à Brassica et Raphanus, telles que des propriétés laxatives, stimulantes, vomitives, toniques et antiseptiques et des préparations sont même utilisées pour traiter la dysenterie, la toux, l'asthme et le diabète (Al-Shehbaz 2011).

4.2. Systématique

Les relations phylogénétiques au sein des Brassiceae ont été examinées à partir de caractères morphologiques (Gómez-Campo 1980, Al-Shehbaz 1985) et de données moléculaires (Pradhan et al. 1992; Warwick et al. 1992; Warwick and Black 1993; Warwick and Black 1994; Warwick and Black 1997; Crespo et al. 2000; Yang et al. 2002; Warwick and Sauder 2005; Arias and Pires 2012; Willis et al. 2014). Ces études ont permis d'identifier huit clades monophylétiques : Cakile, Crambe, Nigra, Savignya, Zilla, Vella, Henophyton et Oleracea (FIGURE 14). La plupart des genres ne sont pas monophylétiques et se trouvent répartis dans deux ou plus de ces huit clades, ce qui nécessiterait une large révision

taxonomique. La résolution phylogénétique au sein des clades est plutôt bonne mais les relations phylogénétiques entre les clades est encore très incertaine, étant donné la forte discordance entre les résultats d'une étude à l'autre et parfois au sein d'une même étude, selon les marqueurs utilisés (Willis et al. 2014). De plus, la plupart des études récentes sont basées sur des gènes ribosomaux, des ITS ou des marqueurs chloroplastiques, tous présentant certaines limitations pour les inférences phylogénétiques. A ce jour, aucune phylogénie robuste à partir de plusieurs gènes nucléaires permettant de résoudre les relations phylogénétiques entre les huit clades de Brassiceae n'a été publiée. Comme mentionné par Al-Shehbaz (2012), suite entre autres à la mise en évidence de la polyphylie de nombreux genres que l'on pensait monophylétiques, il est urgent de fournir pour cette tribu une analyse phylogénétique incluant un grand nombre de marqueurs nucléaires et chloroplastiques.

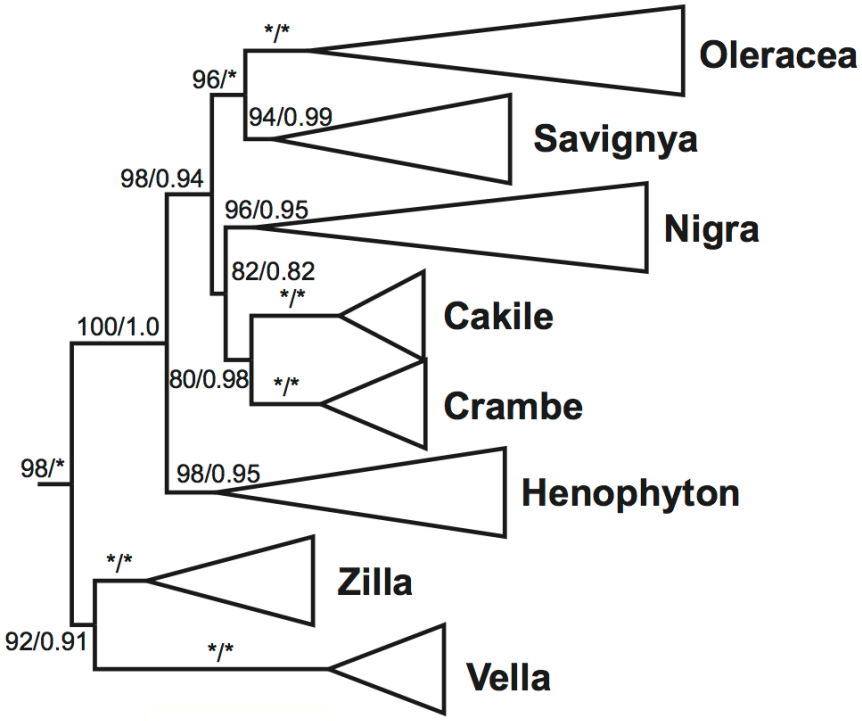


FIGURE 14 – Phylogénie des huit clades de Brassiceae obtenue à partir de quatre marqueurs chloroplastiques hypervariables et d'une analyse bayésienne [Figure extraite de Arias and Pires (2012)]. Les valeurs sur les branches correspondent aux valeurs de bootstrap (BP, pour l'analyse en maximum de vraisemblance) et de probabilités postérieures (PP, pour l'analyse bayésienne). Les astérisques correspondent à des valeurs de support maximales (BP = 100, PP = 1).

4.3. Polyploïdie chez les Brassiceae

Le nombre de chromosomes a été reporté pour environ 180 espèces de Brassiceae (presque 70% de la tribu) réparties dans 44 genres (Al-Shehbaz 1985). Le nombre de chromosomes le plus élevé a été trouvé chez l'espèce *Crambe gordjaginii* Sprygrin & Popov (n=75) tandis que le nombre le plus faible a été trouvé chez l'espèce *Eurucaria cakiloidea* (DC.) O. E. Schulz (n=6) (voir Al-Shehbaz 1985). Chez les Brassiceae, le nombre de chromosomes des génomes haploïdes varie souvent énormément (par exemple, chez Brassica avec n = 7, 8, 9, 10, 11, 16, 17, 18, 19) (Warwick and Al-Shehbaz 2006; Lysak et al. 2007) et les plus communs (7, 8, 9 et 15) se retrouvent chez 14 à 18% des espèces (Al-Shehbaz 1985). Lorsque définie par $n \geq 14$, la polyploïdie apparaît chez environ 36% des espèces de Brassiceae et semble ainsi être exclusive chez les membres des genres Crambe, Moricandia, Vella, Zilla, Schouwia, Henophyton, Fortuynia, Savignya et Succowia (Al-Shehbaz 1985, Warwick and Al-Shehbaz 2006). Toutefois, et comme nous allons le voir, l'estimation de la polyploïdie sur la base du nombre de chromosome s'est avérée ne pas être fiable.

De la polyploïdie intraspécifique a été recensé chez plusieurs espèces dont des espèces du genre Vella et Crambe, constituant *a priori* de l'autopolyploïdie (Warwick and Al-Shehbaz 2006, Lysak et al. 2005, 2007). L'allopolyploïdie est également un mode fréquent de spéciation hybride chez les Brassiceae et beaucoup d'exemples sont largement documentés : *Brassica carinata*, *Brassica juncea*, *Brassica napus* (Yang et al. 2016), *Brassica balerica*, *Diplotaxis muralis* et *Erucastrum gallicum* (Warwick and Al-Shehbaz 2006, Lysak et al. 2005). Les études ayant démontré l'existence d'un événement de mésopolyploïdie (WGT) chez l'ancêtre des Brassiceae (Lysak et al. 2005; Lysak et al. 2007; Kagale et al. 2014) suggèrent cependant que, dans cette tribu, i) les nombres élevés de chromosomes pourraient représenter le caractère le plus ancestral tandis que les nombres faibles de chromosomes pourraient résulter de fusions de chromosomes et ii) que le concept traditionnel utilisé pour estimer le nombre de chromosomes de base et les niveaux de polyploïdie est désuet et non valide (Al-Shehbaz 2011, Lysak et al. 2007). Des processus de réduction du nombre de chromosomes expliquent probablement la variation dans le nombre de chromosomes observée chez les espèces de Brassiceae actuelles, et des nombres élevés de chromosomes ne sont donc pas forcément associés à des évènements récents de polyploïdie (Lysak et al. 2007). La polyploïdie chez les Brassiceae a donc été largement surestimée. Depuis, et à ma connaissance, aucune estimation de la fréquence de la polyploïdie chez les Brassiceae n'a été établie en tenant compte de ces conclusions.

4.4. Auto-incompatibilité chez les Brassiceae

Comme vu précédemment, le système d'auto-incompatibilité chez les Brassiceae est homologue au système rencontré chez toutes les Brassicaceae (excepté le genre Leavenworthia), et implique donc notamment les deux gènes de reconnaissance *SCR* et *SRK*. Cependant, seules deux classes alléliques distinctes ont été retrouvées au locus S chez Brassica et Raphanus, suggérant un fort goulot d'étranglement chez l'ancêtre de ces deux genres, suivi d'une re-diversification allélique au locus S (voir section 3.3.2).

Des cartographies génétiques comparatives du locus S chez Brassica et *Arabidopsis lyrata* ont montré que le locus S occupe des régions chromosomiques différentes chez *Brassica* spp. et chez *Arabidopsis* spp. Chez Brassica, le locus S se trouve dans une région qui est synténique à une région chromosomique localisée sur le chromosome I de *A. thaliana*, tandis que chez *A. lyrata*, le locus S se trouve dans la région chromosomique contenant ARK3, localisée sur le chromosome IV de *A. thaliana* (voir Fobis-Loisy and Gaude 2004 et les références associées), qui s'avère être la position ancestrale du locus S chez les Brassicaceae (Chantha et al. 2013, Vekemans et al. 2014). Cela suggère que le locus S aurait subi une translocation chez l'ancêtre des Brassica (FIGURE 15). Des rétro-élements, que l'on pense être associés au phénomène de réorganisation génomique, sont localisés dans les haplotypes S chez ces deux genres, et auraient donc pu contribuer à la translocation génomique du locus S (Fobis-Loisy and Gaude 2004). Dans ce travail de thèse, nous étudierons l'hypothèse selon laquelle cette translocation génomique du locus S serait associée à l'événement de triplication de génome qui a touché l'ancêtre commun des Brassiceae (Lysak et al. 2005; Lysak et al. 2007; Wang et al. 2011; Vekemans et al. 2014), d'autant que l'on sait que les WGD sont associés à des réorganisations génomiques majeures, dont des évènements de translocation (Tayalé and Parisod 2013). Cependant, la description de la localisation génomique précise du locus S à partir de la séquence complète du génome de *Brassica rapa* n'a pas été étudiée en lien avec les trois sous-génomes parentaux issus du WGT, et nous ignorons également si cette localisation est spécifique à *Brassica rapa* ou concerne l'ensemble des Brassiceae.

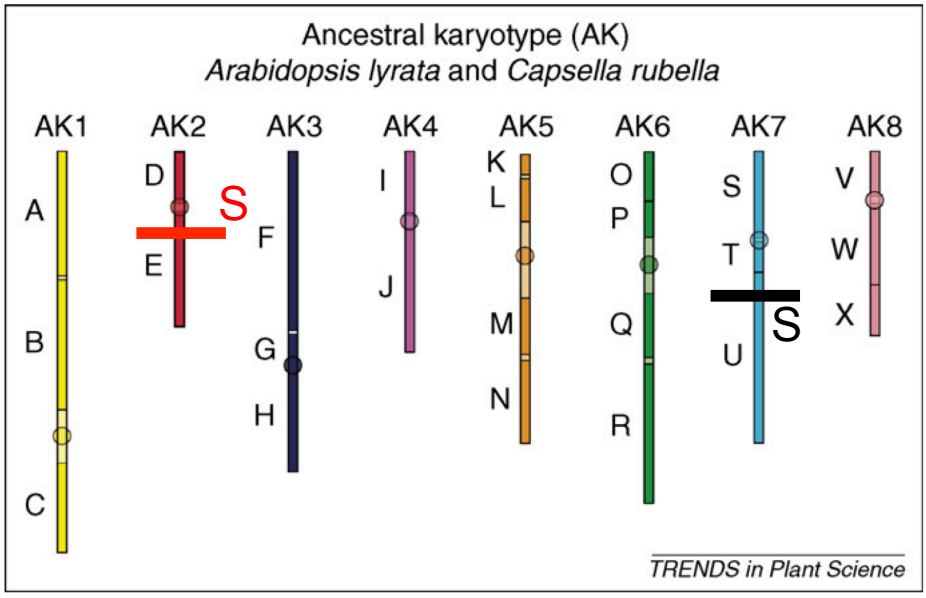


FIGURE 15 – Blocs génomiques du caryotype ancestral des Brassicaceae (n=8) reconstruit à partir d'analyses cytologiques et de cartographies génétiques de *Arabidopsis lyrata* et *Capsella rubella* [Figure extraite et adaptée de Schranz et al. (2006)] et deux des trois localisations génomiques du locus S connues chez les Brassicaceae (voir texte principal pour les références associées). AK1 à AK8 : chromosomes 1 à 8 du caryotype ancestral. En rouge : localisation génomique du locus S chez Brassica. En noir : localisation ancestrale du locus S chez les Brassicaceae (Arabidopsis, Capsella, Eutrema, Sisymbrium, Noccaea).

5. Contexte et objectifs de la thèse

Récemment, une perte drastique de diversité phylogénétique au locus S chez plusieurs espèces des clades « Cakile », « Nigra » et « Oleracea » au sein de la tribu des Brassiceae a été observée (Céline Poux, com. pers.). La richesse allélique à ce locus reste forte mais tous les allèles se regroupent dans les deux clades déjà identifiés dans le genre Brassica et Raphanus (classes I & II, voir section 3.3.2). Par opposition, la diversité phylogénétique au locus S chez Arabidopsis est très élevée, ce qui suggère qu'un goulot d'étranglement majeur aurait eu lieu à l'origine de la tribu des Brassiceae, suivi d'une re-diversification allélique sous l'effet de la sélection fréquence-dépendante négative agissant sur le système SI. De plus, la localisation génomique du locus S chez *Brassica rapa* (Brassiceae, clade Oleracea) est distincte de la position génomique ancestrale observée chez *Arabidopsis sp.* (voir section 3.4.4 et FIGURE 15). Ces observations faites au locus S suggèrent un scénario évolutif original : cette perte de diversité phylogénétique au locus d'auto-incompatibilité et le changement de localisation génomique du locus S pourraient être liés à l'évènement de triplication du génome qui pourrait avoir touché l'ancêtre commun de la tribu des Brassiceae. Ce scénario, par ailleurs, pourrait s'étendre à d'autres exemples au sein de la famille des Brassicaceae et dans d'autres familles possédant un système SI multiallélique. Pour évaluer un tel scénario, nous nous proposons d'étudier les différents clades de Brassiceae, de façon à vérifier d'une part s'ils partagent bien le même évènement unique d'allohexaploïdie, et d'autre part s'ils partagent les mêmes signatures de goulot d'étranglement et de changement de localisation génomique du locus-S. Nous souhaitons également déterminer si le genre mésotétraploïde Orychophragmus, phylogénétiquement proche des Brassiceae, partage ou non la première étape du processus d'allohexaploïdie (c'est-à-dire, le premier événement de WGD) avec la tribu des Brassiceae, et le cas échéant si sa diversité allélique au locus S présente certaines ressemblances avec celle des Brassiceae. Les données moléculaires utilisées dans ce travail sont constituées d'une part de données publiées (données d'assemblage de génomes de Brassiceae et d'autres Brassicaceae; séquences d'allèles au gène *SRK* chez plusieurs espèces de Brassicaceae), et d'autre part de données originales obtenues par l'approche RNAseq à partir d'inflorescences de 6 espèces de la tribu des Brassiceae (TABLEAU 1), espèces choisies pour représenter les différents clades décrits chez les Brassiceae qui ne seraient pas représentés dans les données publiées.

Pour évaluer notre scénario évolutif, nous posons les objectifs opérationnels suivants:

(1) établir une phylogénie robuste des lignées de Brassiceae à partir de gènes nucléaires d'une part et de gènes chloroplastiques d'autre part;

(2) trouver la trace des évènements de WGD chez les espèces de Brassiceae étudiées et chez *Orychophragmus violaceus* et déterminer s'il y a partage d'un de ces évènements;

(3) caractériser le régime de reproduction des espèces étudiées (SI versus SC);

(4) préciser la localisation génomique du locus-S dans les génomes publiés de *Brassica rapa*, *B. oleracea* et *Raphanus sativus*, en lien avec la localisation génomique des sous-génomes parentaux issus de l'évènement WGT;

(5) caractériser la diversité allélique du gène *SRK* chez les Brassiceae et les relations phylogénétiques entre allèles S, de manière à tester l'hypothèse d'un partage du goulot d'étranglement identifié précédemment chez Brassica.

Le premier objectif est essentiel pour pouvoir ensuite positionner sur l'arbre phylogénétique des espèces les évènements successifs de duplication de génome entier, de translocation et de goulot d'étranglement au locus-S. Au cours de cette thèse, l'objectif premier a donc été de développer une méthodologie pour inférer de façon fiable les relations d'orthologie entre séquences codantes obtenues chez les espèces appartenant à la tribu des Brassiceae. Cette méthodologie originale nous a permis de séparer les séquences appartenant à chacun des trois sous-génomes parentaux impliqués dans l'évènement WGT. Après avoir identifié chaque groupe d'orthologues, des méthodes classiques de reconstruction phylogénétique ont été utilisées afin de retracer l'histoire évolutive de la tribu des Brassiceae et de ses lignées parentales. Ces recherches font l'objet du chapitre 1. L'objectif (2) fait l'objet du second chapitre, tandis que les objectifs (4) et (5) sont développés dans le troisième et dernier chapitre de résultat de cette thèse. Le point (3) est présenté en annexe de cette thèse.

TABLEAU 1 – Liste des espèces sélectionnées pour représenter les différents clades de Brassiceae (selon Arias and Pires 2012) qui ne sont pas représentés dans les données publiées disponibles.

Espèces étudiées	Clades de Brassiceae
Carrichtera annua	Vella
Zilla spinosa subsp macroptera	Zilla
Schouwia purpurea	Zilla
Psychine stylosa	Savignya
*Crambe maritima**	Crambe
Cakile maritima	Cakile

*données RNAseq obtenues en fin de thèse donc l'espèce est intégrée dans une partie seulement des différentes analyses

RÉFÉRENCES

Adamowicz SJ, Gregory RT, Marinone MC, Hebert PDN. 2002. New insights into the distribution of polyploid Daphnia : the Holarctic revisited and Argentina explored. Mol. Ecol. 11:1209–1217.

Al-Shehbaz IA. 1985. The genera of Brassiceae (Cruciferae; Brassicaceae) in the southeastern united states. J. Arnold Arbor. 66:279–301.

Al-Shehbaz IA. 2011. Brassicaceae (Mustard Family). eLS. John Wiley Sons, Ltd Chichester.

Al-Shehbaz IA. 2012. A generic and tribal synopsis of the Brassicaceae (Cruciferae). Taxon 61:931–954.

Arias T, Beilstein MA, Tang M, McKain MR, Pires JC. 2014. Diversification times among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. Am. J. Bot. 101:86–91.

Arias T, Pires JC. 2012. A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae : Brassiceae): Novel clades and potential taxonomic implications. Taxon 61:980–988.

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. Curr. Opin. Plant Biol. 15:140–146.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2016. On the relative abundance of autopolyploids and allopolyploids. New Phytol. 210:391–398.

Barker MS, Vogel H, Schranz ME. 2009. Paleopolyploidy in the Brassicales : analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. Genome Biol. Evol. 1:391–399.

Barrett SCH. 2002. The evolution of plant sexual diversity. Nat. Rev. Genet. 3:274–284.

Barringer BC. 2007. Polyploidy and self-fertilization in flowering plants. Am. J. Bot. 94:1527–1533.

Bechsgaard JS, Castric V, Charlesworth D, Vekemans X, Schierup MH. 2006. The transition to self-compatibility in Arabidopsis thaliana and evolution within S-haplotypes over 10 Myr. Mol. Biol. Evol. 23:1741–1750.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. PNAS 107:18724–18718.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Silva C Da, Labadie K, Alberti A, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. 5:1–10.

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

Birchler JA, Veitia RA. 2007. The gene balance hypothesis : from classical genetics to modern genomics. Plant Cell 19:395–402.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. 13:137–144.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678.

Brownfield L, Köhler C. 2011. Unreduced gamete formation in plants : mechanisms and prospects. J. Exp. Bot. 62:1659–1668.

Busch JW, Schoen DJ. 2008. The evolution of self-incompatibility when mates are limiting. Trends Plant Sci. 13:128–136.

Castric V, Billiard S, Vekemans X. 2014. Trait transitions in explicit ecological and genomic contexts: plant mating systems as case studies. In: Landry CR, Aubin-Horth N, editors. Ecological Genomics: Ecology and the Evolution of Genes and Genomes, Advances in Experimental Medicine and Biology. p. 781.

Castric V, Vekemans X. 2004. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Mol. Ecol. 13:2873–2889.

Castric V, Vekemans X. 2007. Evolution under strong balancing selection: how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae? BMC Evol. Biol. 7.

Chang PL, Dilkes BP, Mcmahon M, Comai L, Nuzhdin S V. 2010. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. 11:1–17.

Chantha S-C, Herman AC, Platts AE, Vekemans X, Schoen DJ. 2013. Secondary evolution of a self-incompatibility locus in the Brassicaceae genus Leavenworthia. PLoS Biol. 11:e1001560.

Charlesworth D, Charlesworth B. 1987. Inbreeding depression and its evolutionary consequences. Annu. Rev. Ecol. Syst. 18:237–268.

Cheng F, Sun C, Wu J, Schnable J, Woodhouse MR, Liang J, Cai C, Freeling M, Wang X. 2016. Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. New Phytol. 211:288–299.

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One 7:1–9.

Comai L. 2005. The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6:836–846.

Le Comber SC, Smith C. 2004. Polyploidy in fishes : patterns and processes. Biol. J. Linn. Soc. 82:431–442.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization : clarifying the interplay of multiple models for duplicate gene evolution over time. Curr. Opin. Plant Biol. 19:91–98.

Crespo M, Lledó DM, Fay MF, Chase MW. 2000. Subtribe Vellinae (Brassiceae, Brassicaceae): a combined analysis of ITS nrDNA sequences and morphological data. Ann. Bot. 86:53–62.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:738–749.

Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Hazzouri KM, Wang W, Platts AE, et al. 2015. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. PNAS 112:2806–2811.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary genetics of genome merger and doubling in plants. Annu. Rev. Genet. 42:443–461.

Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-mack J, dePamphilis CW. 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10:1–18.

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

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proc. Natl. Acad. Sci. U. S. A. 112:8362–8366.

Edger PP, Pires JC. 2009. Gene and genome duplications : the impact of dosage-sensitivity on the fate of nuclear genes. Chromosom. Res. 17:699–717.

Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-W, Bewick AJ, Ji L, Platts AE, Bowman MJ, et al. 2017. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell 29:2150–2167.

Edh K, Widen B, Ceplitis A. 2009. Molecular population genetics of the SRK and SCR Self-Incompatibility genes in the wild plant species Brassica cretica (Brassicaceae). Genetics 181:985–995.

Evans BJ, Upham NS, Golding GB, Ojeda RA, Ojeda AA. 2017. Evolution of the largest mammalian genome. Genome Biol. Evol. 9:1711–1724.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous – Tertiary. PNAS 106:5737–5742.

Ferrer MM, Good S V. 2012. Self-sterility in flowering plants : preventing self-fertilization increases family diversification rates. Ann. Bot. 110:535–553.

Fobis-Loisy I, Gaude T. 2004. Molecular evolution of the S Locus controlling mating in the Brassicaceae. Plant Biol. 6:109–118.

Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K. 2011. Cabbage family affairs: The evolutionary history of Brassicaceae. Trends Plant Sci. 16:108–116.

Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr. Opin. Plant Biol. 15:131–139.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication : tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. 60:433–453.

Gallardo MH, Kausel G, Jiménez A, Bacquet C, González C, Figueroa J, Köhler N, Ojeda R. 2004. Whole-genome duplications in South American desert rodents (Octodontidae). Biol. J. Linn. Soc. 82:443–451.

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. 2014. Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. 31:448–454.

Geiser C, Mandáková T, Arrigo N, Lysak MA, Parisod C. 2016. Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in buckler mustard. Plant Cell 28:17–27.

Glasauer SMK, Neuhauss SCF. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol. Genet. Genomics 289:1045–1060.

Glémin S, Gaude T, Guillemin M-L, Lourmas M, Olivieri I, Mignot A. 2005. Balancing selection in the wild: testing population genetics theory of self-incompatibility in the rare species Brassica insularis ´. Genetics 171:279–289.

Goldberg EE, Kohn JR, Lande R, Robertson K a, Smith S a, Igić B. 2010. Species selection maintains self-incompatibility. Science 330:493–495.

Gómez-Campo C. 1980. Morphology and morpho-taxonomy of the tribe Brassiceae. In: Tsunoda, S., Hinata, K. & Gómez-Campo C, editors. Brassica crops and the wild allies: Biology and breeding. Tokyo: Japan Scientific Societies Press. p. 3–31.

Goodwillie C, Kalisz S, Eckert CG. 2005. The evolutionary enigma of mixed mating systems in plants: occurrence, theorical explanations, and empirical evidence. Annu. Rev. Ecol. Evol. Syst. 36:47–79.

Goring DR. 2016. Dominance modifier: expanding mate options. Nat. Plants 3.

Goubet P, Bergès H, Bellec A, Prat E, Helmstetter N, Mangenot S, Gallina S, Holl A-C, Fobis-Loisy I, Vekemans X, et al. 2012. Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. PLoS Genet. 8:e1002495.

Guo X, Liu J, Hao G, Zhang L, Mao K, Wang X, Zhang D, Ma T, Hu Q, Al-shehbaz IA, et al. 2017. Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics 18:1–9.

Guo Y, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D, Schierup MH. 2009. Recent speciation of Capsella rubella from Capsella grandiflora, associated with loss of self-incompatibility and an extreme bottleneck. PNAS 106:5246–5251.

Hall JC, Tisdale TE, Donohue K, Wheeler A, Al-yahya MA, Kramer EM. 2011. Convergent evolution of a complex fruit structure in the tribe Brassiceae (Brassicaceae). Am. J. Bot. 98:1989–2003.

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

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90, 000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. 45:891–898.

Herman AC, Busch JW, Schoen DJ. 2012. Phylogeny of Leavenworthia S-alleles suggests unidirectionnal mating system evolution and enhanced positive selection following an ancient population bottleneck. Evolution 66:1849–1861.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements : A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19:1419–1428.

Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-shehbaz I, Edger PP, et al. 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol. Biol. Evol. 33:394–412.

Igic B, Bohs L, Kohn JR. 2006. Ancient polymorphism reveals unidirectional breeding system shifts. PNAS 103:1359–1363.

Igic B, Busch JW. 2013. Is self-fertilization an evolutionary dead end ? New Phytol. 198:386–397.

Igic B, Lande R, Kohn JR. 2008. Loss of self-incompatibility and its evolutionary consequences. Int. J. Plant Sci. 169:93–104.

Ivanov R, Fobis-Loisy I, Gaude T. 2010. When no means no: guide to Brassicaceae self-incompatibility. Trends Plant Sci. 15:387–394.

Iwano M, Ito K, Fujii S, Kakita M, Asano-shimosato H, Igarashi M, Kaothien-nakayama P, Entani T, Kanatani A, Takehisa M, et al. 2015. Calcium signalling mediates self-incompatibility response in the Brassicaceae. Nat. Plants 1:1–8.

Iwano M, Takayama S. 2012. Self/non-self discrimination in angiosperm self-incompatibility. Curr. Opin. Plant Biol. 15:78–83.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali S, Landherr L, Ralph PE, Jiao Y, Wickett NJ, Ayyampalayam S. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–102.

Kagale S, Robinson SJ, Nixon J, Xiao R, Huebert T, Condie J, Kessler D, Clarke WE, Edger PP, Links MG, et al. 2014. Polyploid evolution of the Brassicaceae during the Cenozoic Era. Plant Cell 26:2777–2791.

Kakizaki T, Takada Y, Ito A. 2003. Linear dominance relationship among four class-II S haplotypes in pollen is determined by the expression of SP11 in Brassica self-incompatibility. Plant Cell Physiol. 44:70–75.

Kasianov AS, Klepikova A V, Kulakovskiy I V, Gerasimov ES, Fedotova A V, Besedina EG, Kondrashov AS, Logacheva MD, Penin AA. 2017. High-quality genome assembly of Capsella bursa-pastoris reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. Plant J. 91:278–291.

Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, Soltis PS. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. Am. J. Bot. 105:1–16.

Laurent S, Salamin N, Robinson-rechavi M. 2017. No evidence for the radiation time lag model after whole genome duplications in Teleostei. PLoS One 12:e0176384.

Leducq J, Gosset CC, Gries R, Calin K, Schmitt É, Castric V, Vekemans X. 2014. Self-Incompatibility in Brassicaceae : identification and characterization of SRK-Like Sequences linked to the S-Locus in the tribe Biscutelleae. G3(Bethesda) 4:983–992.

Lim S, Cho H, Lee S, Cho Y, Kim B. 2002. Identification and classification of S haplotypes in Raphanus sativus by PCR-RFLP of the S locus glycoprotein (SLG) gene and the S locus receptor kinase (SRK) gene. Theor 104:1253–1262.

Liu P, Sherman-Broyles S, Nasrallah ME, Nasrallah JB. 2007. A cryptic modifier causing transient self-incompatibility in Arabidopsis thaliana. Curr. Biol. 17:734–740.

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5:1–11.

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

Lohaus R, Van de Peer Y. 2016. Of dups and dinos : evolution at the K/Pg boundary. Curr. Opin. Plant Biol. 30:62–69.

Lynch M, Conery S. 2000. The evolutionary fate and consequences of duplicate genes.

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, et al. 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups Papaya, Poplar, and Grape : CoGe with Rosids. Plant Physiol. 148:1772–1781.

Lysak MA, Cheung K, Kitschke M, Bures P. 2007. Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. Plant Physiol. 145:402–410.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. Genome Res. 15:516–525.

Mable BK. 2004a. “Why polyploidy is rarer in animals than in plants” : myths and mechanisms. Biol. J. Linn. Soc. 82:453–466.

Mable BK. 2004b. Polyploidy and self-compatibility : is there an association ? New Phytol. 162:803–811.

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

Mandáková T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA. 2010. Fast diploidization in close mesopolyploid relatives of Arabidopsis. Plant Cell 22:2277–2290.

Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. 91:3–21.

Mandáková T, Marhold K, Lysak MA. 2014. The widespread crucifer species Cardamine flexuosa is an allotetraploid with a conserved subgenomic structure. New Phytol. 201:982–992.

Marcet-Houben M, Gabaldón T. 2015. Beyond the whole-genome duplication : phylogenetic evidence for an ancient interspecies hybridization in the Baker’s Yeast lineage. PLoS Biol. 7:1–26.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science 345.

Marhold K, Lihová J. 2006. Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Syst. Evol. 259:143–174.

Mason AS, Nelson MN, Yan G, Cowling WA. 2011. Production of viable male unreduced gametes in Brassica interspecific hybrids is genotype specific and stimulated by cold temperatures. BMC Plant Biol. 11.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. Science 333:1257.

McGrath CL, Gout J-F, Johri P, Doak TG, Lynch M. 2014. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. Genome Res. 24:1665–1675.

Mckain MR, Tang H, Mcneal JR, Ayyampalayam S, Davis JI, Claude W, Givnish TJ, Pires JC, Stevenson DW, Leebens-Mack JH. 2016. A Phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. Genome Biol. Evol. 8:1150–1164.

Miller JS, Venable DL. 2000. Polyploidy and the evolution of gender dimorphism in plants. Science (80-.). 289:2335–2338.

Ming R, Hou S, Feng Y, Yu Q, Dionne-laporte A, Saw JH, Senin P, Wang W, Ly B V, Lewis KLT, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452:991–997.

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S. 2014. Consequences of whole-genome triplication as revealed by comparative genomic analyses of the

wild radish Raphanus raphanistrum and three other Brassicaceae species. Plant Cell 26:1925–1937.

Nasrallah JB, Nasrallah ME. 1993. Pollen-stigma signaling in the sporophytic self-incompatibility response. Plant Cell 5:1325–1335.

Nasrallah JB, Nishio T, Nasrallah ME. 1991. The self-incompatibility genes of Brassia: expression and use in genetic ablation of floral tissues. Annu. Rev. Plant Physiol. Plant Mol. Biol. 42:393–422.

Nasrallah JB. 2002. Recognition and rejection of self in plant reproduction. Science (80-.). 296:305–309.

Nasrallah JB. 2017. Plant mating systems : self-incompatibility and evolutionary transitions to self-fertility in the mustard family. Curr. Opin. Genet. Dev. 47:54–60.

Nishio T, Kusaba M. 2000. Sequence diversity of SLG and SRK in Brassica oleracea L. Ann. Bot. 85:141–146.

Nou IS, Watanabe M, Isogai A, Hinata K. 1993. Sexual plant reproduction comparison of S-alleles and S-glycoproteins between two wild populations of Brassica campestris in Turkey and Japan. Sex. Plant Reprod. 6:79–86.

Novikova PY, Tsuchimatsu T, Simon S, Nizhynska V, Voronin V, Burns R, Fedorenko OM, Holm S, Prat E, Marande W, et al. 2017. Genome sequencing reveals the origin of the allotetraploid Arabidopsis suecica. Mol. Biol. Evol. 34:957–968.

Ockendon D. 2000. The s-allele collection of Brassica oleracea. Acta Hortic. 539:25–30.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annu. Rev. Genet. 34:401–437.

Otto SP. 2007. The evolutionary consequences of polyploidy. Cell 131:452–462.

Paetsch M, Mayland-Quellhorst S, Neuffer B. 2006. Evolution of the self-incompatibility system in the Brassicaceae : identification of S-locus receptor kinase (SRK) in self-incompatible Capsella grandiflora. Heredity 97:283–290.

Panchy N, Lehti-Shiu M, Shiu S. 2016. Evolution of Gene Duplication in Plants. Plant Physiol. 171:2294–2316.

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien M-A. 2010. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol. 186:37–45.

Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ. 2005. Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. Genetics 171:765–781.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. 2012. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492:423–428.

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K. 2009. The flowering world : a tale of duplications. Trends Plant Sci. 14:680–688.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nat. Rev. Genet. 18:411–424.

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Mayer KFX, Olsen O-A. 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. Science (80-.). 345.

Pradhan AK, Prakash S, Mukhopadhyay A, Pental D. 1992. Phylogeny of Brassica and allied genera based on variation in chloroplast and mitochondrial DNA patterns : molecular and taxonomic classifications are incongruous. Theor. Appl. Genet. 85:331–340.

Prigoda NL, Nassuth A, Mable BK. 2005. Phenotypic and genotypic expression of self-incompatibility haplotypes in Arabidopsis lyrata suggests unique origin of alleles in different dominance classes. Mol. Biol. Evol. 22:1609–1620.

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-rechavi M, Shoguchi E, Terry A, Yu J, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1072.

Ramsey J, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst. 29:467–501.

Ramsey J, Schemske DW. 2002. Neopolyploidy in flowering plants. Annu. Rev. Ecol. Syst. 33:589–639.

Renny-Byfield S, Gong L, Gallagher JP, Wendel JF. 2015. Persistence of subgenomes in paleopolyploid cotton after 60 My of evolution. Mol. Biol. Evol. 32:1063–1071.

Robertson K, Goldberg EE, Igi B. 2010. Comparative evidence for the correlated evolution of polyploidy and self-compatibility in Solanaceae. Evolution 65:139–155.

Sato K, Nishio T, Kimura R, Kusaba M, Suzuki T, Hatakeyama K, Ockendon DJ, Satta Y. 2002. Coevolution of the S-Locus genes SRK, SLG and SP11/SCR in Brassica oleracea and B. rapa. Genetics 162:931–940.

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

Schmid M, Evans J, Bogart JP. 2015. Polyploidy in Amphibia. Cytogenet. Genome Res. 145:315–330.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. PNAS 108:4069–4074.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.

Schranz ME, Lysak MA, Mitchell-Olds T. 2006. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. 11:535–542.

Schranz ME, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell 18:1152–1165.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification : the WGD Radiation Lag-Time Model. Curr. Opin. Plant Biol. 15:147–153.

Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, et al. 2016. Genome evolution in the allotetraploid frog Xenopus laevis. Nature 538:336–343.

Shiba H, Iwano M, Entani T, Ishimoto K, Shimosato H, Che F, Satta Y, Ito A, Takada Y, Watanabe M, et al. 2002. The dominance of alleles controlling self-incompatibility in Brassica pollen is regulated at the RNA level. Plant Cell 14:491–504.

Shiba H, Kakizaki T, Iwano M, Tarutani Y, Watanabe M, Isogai A, Takayama S. 2006. Dominance relationships between self-incompatibility alleles controlled by DNA methylation. Nat. Genet. 38:297–300.

Shimizu KK, Cork JM, Caicedo AL, Mays CA, Moore RC, Olsen KM, Ruzsa S, Coop G, Bustamante CD, Awadalla P, et al. 2004. Darwinian selection on a selfing locus. Science 306:2081–2084.

De Smet R, Adams KL, Vandepoele K, Montagu MCE Van, Maere S. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. PNAS 110:2898–2903.

Smith SA, Brown JW, Yang Y, Bruenn R, Drummond CP, Brockington SF, Walker JF, Last N, Douglas NA, Moore MJ. 2017. Disparity, diversity, and duplications in the Caryophyllales. New Phytol. in press.

Soltis DE, Albert VA, Leebens-mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Claude W, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. Am. J. Bot. 96:336–348.

Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, Mavrodiev E V., Mei W, Cortez MB, Soltis PS, Gitzendanner MA. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). New Phytol. 202:1105–1117.

Soltis DE, Soltis PS. 1993. Molecular data and the dynamic nature of polyploidy. CRC. Crit. Rev. Plant Sci. 12:243–273.

Soltis DE, Visger CJ, Soltis PS. 2014. The polyploidy revolution then...and now: Stebbins revisited. Am. J. Bot. 101:1057–1078.

Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. Curr. Opin. Plant Biol. 30:159–165.

Steige KA, Slotte T. 2016. Genomic legacies of the progenitors and the evolutionary consequences of allopolyploidy. Annu. Rev. Genet. 30:88–93.

Stone JL. 2002. Molecular mechanisms underlying the breakdown of gametophytic self-incompatibility. Q. Rev. Biol. 77:17–32.

Suzuki T, Kusaba M, Matsushita M, Okazaki K, Nishio T. 2000. Characterization of Brassica S-haplotypes lacking S-locus glycoprotein. FEBS Lett. 482:102–108.

Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. On the tetraploid origin of the maize genome. Comp. Funct. Genomics 5:281–284.

Takasaki T, Hatakeyama K, Suzuki G, Watanabe M, Isegai A, Hinata K. 2000. The S receptor kinase determines self-incompatibility in Brassica stigma. Nature 403:913–916.

Takayama S, Isogai A. 2005. Self-incompatibility in plants. Annu. Rev. Plant Biol. 56:467–489.

Takuno S, Nishio T, Satta Y, Innan H. 2008. Preservation of a pseudogene by gene conversion and diversifying selection. Genetics 180:517–531.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ. 2015. Nested radiations and the pulse of angiosperm diversification : increased diversification rates often follow whole genome duplications. New Phytol. 207:454–467.

Tarutani Y, Shiba H, Iwano M, Kakizaki T, Suzuki G, Watanabe M, Isogai A, Takayama S. 2010. Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. Nature 466:983–987.

Tayalé A, Parisod C. 2013. Natural pathways to polyploidy in plants and consequences for genome reorganization. Cytogenet. Genome Res. 140:79–96.

Tedder A, Ansell SW, Lao X, Vogel JC, Mable BK. 2011. Sporophytic self-incompatibility genes and mating system variation in Arabis alpina. Ann. Bot. 108:699–713.

The International Wheat Genome Sequencing Consortium (IWGSC). 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345:1251788.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16:934–946.

Tsuchimatsu T, Goubet PM, Gallina S, Holl A-C, Fobis-loisy I, Bergès H, Marande W, Prat E, Meng D, Long Q, et al. 2017. Patterns of polymorphism at the self-incompatibility locus in 1,083 Arabidopsis thaliana genomes. Mol. Biol. Evol. 34:1878–1889.

Tsuchimatsu T, Kaiser P, Yew C, Bachelier JB, Shimizu KK. 2012. Recent loss of self-incompatibility by degradation of the male component in allotetraploid Arabidopsis kamchatica. PLoS Genet. 8:e1002838.

Tsuchimatsu T, Shimizu KK. 2013. Effects of pollen availability and the mutation bias on the fixation of mutations disabling the male specificity of self-incompatibility. J. Evol. Biol. 26:2221–2232.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596–1604.

Uyenoyama MK. 1995. A Generalized Least-Squares Estimate for. Genetics 139:975–992.

Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous − Paleogene boundary. Genome Res. 24:1334–1347.

Vekemans X, Poux C, Goubet PM, Castric V. 2014. The evolution of selfing from outcrossing ancestors in Brassicaceae: what have we learned from variation at the S-locus? J. Evol. Biol. 27:1372–1385.

Vekemans X, Schierup MH, Christiansen FB. 1998. Mate availability and fecundity selection in multi-allelic self-incompatibility systems in plants. Evolution (N. Y). 52:19–29.

Vergilino R, Belzile C, Dufresne F. 2009. Genome size evolution and polyploidy in the Daphnia pulex complex (Cladocera : Daphniidae). Biol. J. Linn. Soc. 97:68–79.

Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. Ann. Bot. 120:195–207.

Vision TJ, Brown DG, Tanksley SD. 2000. The Origins of Genomic Duplications in Arabidopsis. Science 290:2114–2117.

Walker JF, Yang Y, Moore MJ, Mikenas J, Timoneda A, Brockington SF, Smith SA. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. Am. J. Bot. 104:858–867.

Walling JG, Shoemaker R, Young N, Mudge J, Jackson S. 2006. Chromosome-level homeology in paleopolyploid soybean (Glycine max) revealed through integration of genetic and chromosome maps. Genetics 172:1893–1900.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43:1035–1040.

Warwick SI, Al-Shehbaz IA. 2006. Brassicaceae: Chromosome number index and database on CD-Rom. Plant Syst. Evol. 259:237–248.

Warwick SI, Black LD, Aguinagalde I. 1992. Molecular systematics of Brassica and allied genera (Subtribe Brassicinae , Brassiceae) - chloroplast DNA variation in the genus Diplotaxis. Theor. Appl. Genet. 83:839–850.

Warwick SI, Black LD. 1993. Molecular relationships in subtribe Brassicinae (Cruciferae, tribe Brassiceae). Can. J. Bot. 71:906–918.

Warwick SI, Black LD. 1994. Evaluation of the subtribes Moricandiinae, Savignyinae, Vellinae, and Zillinae (Brassicaceae, tribe Brassiceae) using chloroplast DNA restriction site variation. Can. J. Bot. 72:1692–1701.

Warwick SI, Black LD. 1997. Phylogenetic implications of chloroplast DNA restriction site variation in subtribes Raphaninae and Cakilinae (Brassicaceae, tribe Brassiceae). Can. J. Bot. 75:960–973.

Warwick SI, Mummenhoff K, Sauder CA, Koch MA, Al-Shehbaz IA. 2010. Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. Plant Syst. Evol. 285:209–232.

Warwick SI, Sauder CA. 2005. Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trn L intron sequences. Can. J. Bot. 83:467–483.

Watanabe M, Ito A, Takada Y, Ninomiya C, Kakizaki T, Takahata Y, Hatakeyama K, Hinata K, Suzuki G, Takasaki T, 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 Lett. 473:139–144.

Wendel JF. 2000. Genome evolution in polyploids. Plant Mol. Biol. 42:225–249.

Willis CG, Hall JC, Casas RR De, Wang TY, Donohue K. 2014. Diversification and the evolution of dispersal ability in the tribe Brassiceae (Brassicaceae). Ann. Bot. 114:1675–1686.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. PNAS 106:13875–13879.

Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. PNAS 111:5283–5288.

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M. 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biol. 8:e1000409.

Wright S. 1939. The distribution of self-sterility alleles in populations. Genetics 24:538–552.

Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, Xu J, Zheng X, Ren L, Wang G, et al. 2014. Genome sequence and genetic diversity of the common carp, Cyprinus carpio. Nat. Genet. 46:1212–1219.

Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, Hu Z, Chen S, Pental D, Ju Y, et al. 2016. The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. Nat. Genet. 48:1225–1234.

Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK, Carpenter EJ, Zhang Y, Chen L, Yan Z, Xie Y, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Mol. Biol. Evol. 32:2001–2014.

Yang Y, Tai P, Chen Y, Li W. 2002. A study of the phylogeny of Brassica rapa, B. nigra, Raphanus sativus, and their related genera using noncoding regions of chloroplast DNA. Mol. Phylogenet. Evol. 23:268–275.

Yasuda S, Wada Y, Kakizaki T, Tarutani Y, Miura-uno E, Murase K, Fujii S, Hioki T, Shimoda T, Takada Y, et al. 2016. A complex dominance hierarchy is controlled by polymorphism of small RNAs and their targets. Nat. Plants 3:1–5.

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, et al. 2005. The genomes of Oryza sativa : a history of duplications. PLoS Biol. 3:266–281.

Ziolkowski PA, Kaczmarek M, Babula D, Sadowski J. 2006. Genome evolution in Arabidopsis/Brassica: conservation and divergence of ancient rearranged segments and their breakpoints. Plant J. 47:63–74.

CHAPITRE I

Développement d'une approche méthodologique pour résoudre la phylogénie nucléaire de la tribu allo-hexapolyploïde des Brassiceae

A ce jour, bien que toutes les études ayant établi une phylogénie des Brassiceae s'accordent sur la présence de 8 clades monophylétiques dont la composition en espèces est plutôt bien résolue, la topologie des clades varie d'une étude à l'autre selon les marqueurs utilisés (voir section 3.4.2). Une des difficultés majeures rencontrées pour établir une phylogénie des Brassiceae est liée à la présence d'un grand nombre de gènes en multiples copies dans le génome nucléaire, conséquence directe du WGT unique qui aurait touché l'ancêtre commun des Brassiceae (Lysack et al. 2005, 2007). De plus, bien qu'une partie des pertes et rétention de gènes suivant les évènements de polyploïdie semble en général être rapide et précéder la divergence des lignées, on ne peut exclure une perte différentielle des gènes entre les lignées, ce qui complexifie encore davantage la recherche des orthologues. Les méthodes classiques de clustering tels que les algorithmes ORTHO-MCL ou encore ORTHO-FINDER ne permettent pas d'inférer les orthologues de façon fiable dans un contexte de duplication de génome et à partir de données lacunaires telles que des assemblages de transcriptomes (séquences incomplètes, faible couverture ou expression de certains gènes). C'est pourquoi, nous avons choisi d'utiliser une approche basée sur les arbres de gènes (tree-based approach) de façon à identifier sur chaque arbre de gènes chacun des trois groupes d'homéologues attendus. Nous avons utilisé pour cela l'annotation des sous-génomes parentaux effectuée sur les génomes assemblés de *Brassica rapa* et *B. oleracea* (Liu et al. 2014) de façon à pouvoir assigner les homéologues des espèces étudiées à leur sous-génome parental selon leur position phylogénétique dans l'arbre de gènes. Cette méthodologie et les résultats qui en découlent pour la résolution de la phylogénie de la tribu des Brassiceae sont présentés dans l'article suivant (en préparation pour publication).

Références

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5:1–11.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. Genome Res. 15:516–525.

Lysak MA, Cheung K, Kitschke M, Bures P. 2007. Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. Plant Physiol. 145:402–410.

Title: A new methodological framework for inferring the evolutionary history of mesopolyploid lineages: an application to the Brassiceae tribe (Brassicaceae).

Authors

Hénocq L.[1], Gallina S.[1], Schmitt E.[1], Genete M.[1], Castric V.[1], Vekemans X.[1], Poux C.[1]

Affiliations

[1]Univ. Lille, CNRS, UMR 8198 – Evo-Eco-Paleo, Lille, France;

Abstract

Whole genome duplication (WGD) events are notably widespread in plants and increasing evidence shows that since the origin of the group, nearly all extant angiosperm lineages have experienced single or multiple ancient polyploidization events. After going through an allopolyploidization event, a genome contains homologous gene sequences from different evolutionary origins (homoeologs). This brings the difficulty of using sequences from the nuclear genome for phylogenetic inferences, due to the risk of orthology and paralogy conflation. Thus, reconstructing the evolutionary history of allopolyploid lineages from nuclear genome sequences requires separating the homoeologs gene copies (*i.e.* gene duplicates originating from whole genome duplication) into ortholog groups from a common parental origin. In this study, we propose a methodological approach to resolve the phylogenetic network of allo-mesopolyploid clades. This method requires a well-annotated reference genome of a member of the allo-mesopolyploid clade for which the identification of the parental subgenome fragments has been performed. We used transcriptomic data for the other taxa. Focusing on fully retained genes (i.e. genes for which all homoeologous gene copies inherited from the parental lineages are still present in the reference genome), we constructed gene trees that were multi-labelled (homolog trees) for assigning each homoeolog gene sequence to its diploid parental lineage, *i.e.* for separating orthologs from paralogs. Once the identification of homoeologs was performed for each studied species, tree-building methods were used to reconstruct the species tree based on the concatenation of the orthologs' groups (orthologs' group *sensu* Yang et al. (2014)). This method allows resolving the phylogenetic relationships (i) among all extant investigated species within a mesopolyploid clade, (ii) among the parental lineages of a mesopolyploid lineage and, (iii) between the parental lineages and closely related extant species. To illustrate our methodological framework we applied it to a clade belonging to the Brassicaceae plant family, the Brassiceae tribe that experienced a mesohexaploidy event. By applying our methodological approach, we provide the first nuclear-based phylogeny of the Brassiceae tribe as well as a fully resolved chloroplast phylogeny based on extended genomic datasets from both nuclear and chloroplast genomes.

Key words: WGD, mesopolyploid, phylogenetic inference, homoeologs, orthologs' group**,** homolog trees, tree-based orthology inference

Introduction

Whole genome duplication (WGD) events correspond to large-scale gene duplication processes resulting in the formation of a polyploid organism either by duplication of the whole genome within a given species (autopolyploidy) or by merging of two or more diverged genomes during an interspecific hybridization event (allopolyploidy). WGD events are notably widespread in plants and increasing evidence shows that since the origin of the group, nearly all extant angiosperm lineages have experienced single or multiple ancient polyploidization events (Jiao et al. 2011; Arrigo and Barker 2012; Panchy et al. 2016; Van de Peer et al. 2017). Moreover, nearly 35% of extant flowering plant species are neopolyploids (i.e. recently formed polyploids) and it is estimated that at least 15% of recent speciation events are associated with ploidy increase (Wood et al. 2009). WGDs generate high diversity in size, function and structure of genomes and are undoubtedly a fundamental process shaping the evolution and diversification of plant lineages (Otto and Whitton 2000; Marhold and Lihová 2006; Doyle et al. 2008; Soltis et al. 2009; Soltis, Segovia-Salcedo, et al. 2014; Panchy et al. 2016; Landis et al. 2018).

Polyploid taxa can be classified as neo-, meso-, or paleo-polyploids according to the age of the last WGD event in their history, and to the degree of genomic rearrangement (Mandáková et al. 2010). Neopolyploids are the most recently formed polyploids: they display an increase in chromosome numbers and genome size, a highly redundant gene content, and their diploid parents are often still present in the extant flora or fauna. Mesopolyploids have experienced some level of genome reshuffling and gene fractionation (loss of homoeologous copies) across the parental subgenomes, but individual genomic blocks corresponding to each of the parental subgenomes can usually be identified through comparative genomic approaches. Moreover, the level of gene redundancy in mesopolyploids is generally highly variable among gene ontologies due to differential constraints on functional redundancy and gene dosage flexibility within functional gene networks (Lou et al. 2012; Geiser et al. 2016; Mandáková et al. 2017). Paleopolyploids are the most ancient polyploids, whose parental nuclear subgenomes are more difficult to identify as a consequence of strong gene fractionation and severe genomic rearrangements, and the level of gene redundancy is more strongly reduced.

After going through an allopolyploidization event, a genome contains homologous gene sequences from different parental origins, called homoeologs. This brings the difficulty of using the nuclear genome for phylogenetic inferences, due to the risk of orthology and paralogy conflation (Vanderpoorten et al. 2004; Van Der Niet and Linder 2008). This problem is especially prominent in mesopolyploid taxa deriving from allopolyploid events for the following reasons: (1) assigning an homoeologous copy to a given parental subgenome is not trivial because in many cases the parental lineages or their diploid descendants are extinct, while this is generally not true for neopolyploids and (2) the divergence among the parental lineages preceding the WGD event may be non-negligible as compared to the divergence among extant taxa sharing the same WGD event, while this is not the case for autopolyploids and is minimized in paleopolyploids. The consequence of wrong homoeologs assignation will then be the reconstruction of wrong phylogenetic trees. Given the high frequency of mesopolyploid taxa in flowering plants (Mandáková et al. 2017; Van de Peer et al. 2017), the availability of homoeologs' assignation procedures would be useful to investigate the evolutionary history of an important proportion of plant species.

Consequently, chloroplast genic and intergenic regions have been widely used to investigate the phylogenetic relationships among mesopolyploid plants, as the chloroplast genome contains a low proportion of duplicated regions (Warwick et al. 2010; Arias and Pires 2012). Moreover, the increasing availability of complete chloroplast genomes has fostered their use for obtaining accurate phylogenetic inferences in seed plants (Bortiri et al. 2008; Parks et al. 2009; Hohmann et al. 2015; Guo et al. 2017), but because the chloroplast genome is usually maternally inherited in flowering plants it will only recover a single parental lineage, when applied to allopolyploids. Therefore, the information carried by nuclear genes is not only necessary to detect hybridization and ancient allopolyploidization events (Sang 2002) but as well to encompass the whole evolutionary history of a lineage. Nuclear genes in angiosperms vary in copy number due to local duplication and/or large scale or whole genome duplications, which complicates the identification of orthologous sequences. Thus, low-copy nuclear genes, and especially single-copy genes, have been widely used to infer phylogenetic relationships among plants (Duarte et al. 2010; Zhang et al. 2012; Yang et al. 2015; Huang et al. 2016). However, in paleo- or mesopolyploids, some low or single-copy genes shared among species can actually be paralogs instead of orthologs, due to alternate homoeologous gene copy losses in different lineages since the shared polyploidization event.

Reconstructing the evolutionary history of allopolyploid lineages from nuclear genome sequences requires separating the homoeologous gene copies (*i.e.* gene duplicates originating from whole genome duplication) into orthologs groups sharing a common origin. Several methods have been proposed to reconstruct phylogenies of polyploid lineages however they are mainly dedicated to neopolyploids, in other words, taxa for which at least some of the parental lineages of lower ploidy levels are not extinct, and for which ploidy information is clearly shown by variation in chromosome count (Oxelman et al. 2017). For example multi-labelled trees (trees displaying several leaves for a single species) have been proposed to model the putative evolutionary history of a polyploid species. The multi-labelled gene tree topologies can be summarized into a consensus tree under the implicit assumption that the majority of gene trees reflect the underlying genome tree. The multi-labelled consensus is then folded into a uni-labelled network by minimizing the number of hybridation events (Huber et al. 2006; Lott et al. 2009; Albrecht et al. 2012). Next-generation sequencing methods motivated the development of species network inference methods allowing assigning homoeologs to their respective diploid parental lineage (Jones et al. 2013, Oberprieler 2017). Methodologies allowing reconstructing phylogenies for homoploid hybrid taxa have also been proposed (Glémin et al. 2018). In this case the new hybrid taxon does not undergo any increase in ploidy level and recombination proceeds between the parental sub-genomes. However, none of the aforementioned methods can be used reliably for reconstructing phylogenies of mesopolyploid lineages because (i) in these lineages gene copies are often lost differentially among species, which increases drastically the confusion between orthologs and paralogs, (ii) mesopolyploïd species went through a diploidization process therefore their chromosome counts are uninformative and, (iii) mesopolyploid parental species or lineages are often extinct and therefore cannot be included in the phylogenetic analysis.

Recently, Yang & Smith (2014) proposed a tree-based orthology inference approach for non-model organisms capable of accommodating genome duplication, incompleteness of transcriptomes and low-coverage genome data. This method allows reconstructing the phylogeny of mesopolyploid clades, but the assignation of the homoeologous genes to specific parental lineages cannot be achieved, so the history of the parental lineages cannot be inferred. To the best of our knowledge, a methodological framework using a large dataset and specifically adapted to the reconstruction of phylogenies of allo-mesopolyploid taxa– whose sub-genomes have been partially reshuffled and fractionated since the allopolyploidy event – is still lacking. A number of ancient allopolyploid genomes have been sequenced and

annotated (Schnable et al. 2011; Wang et al. 2011; Renny-Byfield et al. 2015). Using these genomes as reference, it is possible to take advantage of the homoeologous genes assignation to reconstruct the evolutionary history of entire allopolyploid lineages using next generation sequencing data and performing tree-based orthology inferences.

In this study, we propose a methodological approach to resolve the phylogeny of allo-mesopolyploid clades. This method requires at least one well-annotated reference genome of a member of the allo-mesopolyploid clade for which the identification of the parental (progenitors) subgenome fragments has been performed. Focusing on fully retained genes (*i.e.* genes for which one copy is retained in each of the merged parental subgenomes), we constructed gene trees that were multi-labelled (homolog trees) for assigning each homoeologous gene sequence to its diploid parental lineage, *i.e.* for separating orthologs from paralogs. Once the identification of homoeologs was performed for each studied species, tree-building methods were used to reconstruct the species tree based on the concatenation of the ortholog groups (ortholog group *sensu* Yang et *al.* (2014)). This method allows us to resolve the phylogenetic relationships *(i)* among the parental lineages, which are likely to be extinct lineages and *(ii)* among all extant investigated species within the mesopolyploid clade.

In order to test our methodological framework, we applied it to a clade belonging to the Brassicaceae plant family, the Brassiceae tribe. All Brassicaceae species share a unique WGD event (called α event) that took place during the early history of the family, before its diversification (Vision et al. 2000; Blanc et al. 2003; Schranz and Mitchell-olds 2006; Barker et al. 2009; Kagale et al. 2014; Edger et al. 2015). In addition to this event, multiple independent mesopolyploid WGD events postdating the α-WGD have occurred in several Brassicaceae lineages and are likely associated with important species radiation (Mandáková et al. 2017); as for example within the Biscutelleae tribe (Geiser et al. 2016), the Leavenworthia genus (Haudry et al., 2013), and the Brassiceae tribe (Lagercrantz 1998; Lysak et al. 2005; Parkin et al. 2005; Ziolkowski et al. 2006; Lysak et al. 2007; Wang et al. 2011a; Cheng et al. 2013; Cheng et al. 2014). Broadly, 11 (22%) of the 49 recognized tribes of the Brassicaceae family (Al-shehbaz 2012) have a mesopolyploid ancestry (Lysak et al. 2005; Lysak et al. 2007; Kagale et al. 2014; Mandáková et al. 2017). In the Brassiceae tribe, species with diploid-like genomes analysed to date contain either three or six (in neotetraploid species) copies of orthologous genomic regions of *A. thaliana*. This suggests that the Brassiceae tribe has experienced two successive WGD events, generating a whole genome triplication (WGT) and that all present-day diploid species in the Brassiceae tribe derived

from a mesohexaploid ancestor. Comparative subgenome analyses in *Brassica rapa*, focusing on patterns of gene fractionation (i.e. post-WGD gene losses), suggested a two-step origin for the allohexaploid lineage involving, firstly, an allotetraploidization event from two diploid ancestral genomes (named MF1 and MF2, for "Medium fractionated" subgenome and "Most Fractionated" subgenome, respectively), followed by genomic reshuffling and gene fractionation, and secondly, a subsequent hybridization with a third diploid parental genome (named LF, for "Least Fractionated" subgenome), followed by a second round of genomic reshuffling and gene fractionation (Cheng et al. 2012; Tang et al. 2012; Cheng et al. 2013; Cheng et al. 2014; Murat et al. 2015). Consequently, inferring phylogenetic relationships among Brassiceae using nuclear genes is difficult given the high number of homoeologs present in Brassiceae genomes, and differential gene loss/retention following the WGT. The Brassiceae tribe is firmly placed within the clade B of the Brassicaceae family together with the tribes Thlaspideae, Eutremeae, Sisymbrieae, and Isatideae (Huang et al. 2016; Guo et al. 2017). It contains 47 genera and 227 ssp. (Al-shehbaz 2012) and eight monophyletic sub-tribes: “Vella”, “Zilla”, “Cakile”, “Crambe”, “Henophyton”, “Nigra”, “Oleracea” and “Savignya” (Arias and Pires 2012). To date, the reconstruction of phylogenetic relationships within and among clades of the tribe Brassiceae, as well as the tribe circumscription, were performed by using chloroplast markers, mitochondrial DNA restriction profiles, or nuclear markers limited to ITS and restriction sites polymorphisms (Pradhan et al. 1992; Warwick and Black 1991; Warwick and Black 1993; Warwick and Black 1994; Warwick and Black 1997; Warwick and Sauder 2005). At last, Arias and Pires (2012) provided a fully resolved phylogeny of the tribe based on few rapidly evolving non-coding chloroplast regions, but other studies suggested that nuclear and chloroplast markers reconstruct markedly different topologies (Couvreur et al. 2010; Hall et al. 2011). This result may be explained by a lack of orthology between the aligned sequences but as well by hybridization events after diversification of the tribe (Yang et al. 2002). Consequently, it appears that a robust Brassiceae phylogeny should be assessed using nuclear genes associated with an appropriate methodology.

Methods

Obtaining and assembling transcriptomic data

Plant material

The following species have been sequenced for the present study: *Orychophragmus violaceus* (Brassicaceae, used as outgroup), *Carrichtera annua* (Brassiceae, Vella clade), *Zilla spinosa subsp. macroptera* (Brassiceae, Zilla clade), *Schouwia purpurea* (Brassiceae, Zilla clade), *Psychine stylosa* (Brassiceae, Savignya clade) and *Cakile maritima* (Brassiceae, Cakile clade). Flower buds of *Cakile maritima* were collected in natural populations and seeds of all other species were samples from the Canadian Genbank (http://pgrc3.agr.gc.ca/acc/search-recherche_e.html) and originated from various botanical gardens (see Table S1). Seeds from each species were sown in potting soil and germinated at 20°C in a greenhouse providing a 12-h photoperiod for several weeks and controlled conditions until germination. After germination, all plants were grown in controlled greenhouse conditions until flowering. Flower buds were collected for RNA sequencing, except for *Zilla spinosa subsp. macroptera* for which the RNA extraction was performed from leaves due to the lack of flowering. We sequenced a variable number of individuals per species (2 to 12, Table S1).

cDNA library preparation and transcriptome sequencing

Total RNA was extracted from flower buds or leaves (Zilla) through the Spectrum Plant Total RNA kit (Sigma, Inc., USA), following the manufacturer's protocol, and treated with a DNAse. We performed a paired-end RNA sequencing using the TruSeq RNA sample Preparation v2 kit (Illumina Inc., USA) and the Illumina technology. For this purpose, cDNA libraries were prepared, bar-coded, mixed and paired-end sequenced for each individual by using Illumina HiSeq2000 (*Orychophragmus violaceus*), Illumina HiSeq2500 (*Cakile maritima*) and Illumina HiSeq3000 (other sequenced species) (Table S1). Demultiplexing was performed using CASAVA 1.8.1 (Illumina Inc., USA) to produce paired sequences files containing reads for each sample in the Illumina FASTQ format. Then we used FastQC (Andrews, 2010), a quality control tool for high throughput sequence data. RNA extraction and sequencing were done by the sequencing platform MGX (Montpellier, France).

De novo transcriptome assembly

After filtering out low quality reads, adaptor sequences and poly-A tails were trimmed and reads showing GC content bias, low complexity or size as well as exact duplicates were removed using PRINSEQ (Schmieder and Edwards 2011) and Cutadapt (Martin 2011). We controlled the quality of clean reads by using FastQC (Table S1). Transcriptome assembly was performed using TRINITY with default parameters (Grabherr et al. 2013). To minimize redundancy in each transcriptome assembly (mainly due to the presence of numerous isoforms) we used CAP3, a multiple sequence alignment method used to generate consensus sequences (Huang and Madan 1999). An overlap of 120bp and 98% of identity between two or more isoforms induced the construction of consensus sequences. Then, we used QUAST (Gurevich et al. 2013) to evaluate the assembly quality (Table S2).

Chloroplast phylogeny

Annotation of chloroplast contigs

We used published chloroplast genomes of 20 species including 8 non-Brassiceae species and 22 transcriptome assemblies representing 5 species (Table 1) to build the chloroplast Brassiceae species tree. By using BLASTN (Altschul et al. 1990) with an e-value cutoff of $1e^{-6}$, a percentage of identity of 80% and a minimum alignment length of 100bp, coding chloroplast DNA sequences of *A. thaliana* (78 coding genes) were mapped onto the DNA sequences of the 20 whole chloroplast genomes. *A. thaliana* was selected as reference in order to confirm the annotation of the published chloroplast genomes and thereby the orthology of the sequences. By using the same procedure with the annotated coding chloroplast sequences of *Brassica nigra* (Brassiceae), we extracted the orthologous sequences in all Brassiceae and Orychophragmus transcriptome assemblies. We only extracted the portion of each best-hit contig that aligned with the reference sequence.

Due to the lack of available chloroplast genome of *Sisymbrium irio* and because we wanted to assess its phylogenetic position compared to the tribe Brassiceae and *O. violaceus*, we extracted chloroplast gene sequences from the raw reads obtained in the genome sequencing project of *S. irio* (Haudry et al. 2013) (Table 1). For this purpose, we mapped these raw reads onto each coding DNA sequence of the chloroplast genome of *B. nigra* using Bowtie2 (Langmead and Salzberg 2012). Mapping statistics can be found in Supplementary Table S3. SAM files were compressed and sorted using SAMtools (Li and Durbin 2009). Then, SPAdes was used to perform the *de novo* assembly of mapped reads (Bankevich et al.

2012), separately for each corresponding sequence. Finally, we used the genomic similarity search tool YASS (Noe and Kucherov 2005) to check the conformity of each resulting contig with its reference, which is reported in Table S3. When several contigs of *S. irio* were constructed from mapped reads for a given *B. nigra* reference sequence, the longest contig was extracted for further analysis. Two contigs with an error rate above 4.5% were discarded for the analyses (Table S3).

Based on this procedure, we retrieved from 48 to 78 chloroplast gene sequences per species. Only one sequence per species and per gene, the longest, was selected before further analysis.

Table 1. Whole chloroplast genomes, transcriptome assemblies and paired-end sequencing data used for the chloroplast phylogenetic analysis. The number of genes used for each species is indicated.

Data	Species (number of individuals)	Brassicaceae clade[a].lineage[b]	Brassiceae clade[c]	No. accession	Number of genes
Whole chloroplast genome	*Alliaria grandifolia*	B – II	-	KX342847	73
	Arabidopsis thaliana	A – I	-	NC_00932	74
	Barbarea verna	A – I	-	NC_009269.1	74
	Brassica napus	B – II	Oleracea	NC_016734	74
	Brassica nigra	B – II	Nigra	KT878383	74
	Brassica oleracea	B – II	Oleracea	KR233156	74
	Brassica rapa	B – II	Oleracea	NC_015139	66
	Cakile arabica	B – II	Cakile	NC_030775.1	73
	Cardamine limprichtiana	B – II	-	KX342848	62
	Eutrema salsugineum	B – II	-	NC_028170	74
	Isatis tinctoria	B – II	-	NC_028415.1	73
	Lobularia maritima	C – III	-	NC_009274.1	73
	Orychophragmus diffusus	B – II	-	NC_033498.1	74
	Orychophragmus hupehensis	B – II	-	NC_033500.1	74
	Orychophragmus longisiliqus	B – II	-	KX756549.1	74
	Orychophragmus taibaiensis	B – II	-	NC_033499.1	74
	Orychophragmus violaceus	B – II	-	KX364399	74
	Orychophragmus zhongtiaoshanus	B – II	-	KX756547	74
	Raphanus sativus	B – II	*Undetermined*	NC_024469	74
	Schrenkiella parvula	B – II	-	NC_028726	74
Transcriptome assemblies	*Cakile maritima* (12)	B – II	Cakile	-	61
	Carrichtera annua (3)	B – II	Vella	-	58
	Pychine stylosa (2)	B – II	Savignya	-	57
	Schouwia purpurea (3)	B – II	Zilla	-	59
	Zilla macroptera (2)	B – II	Zilla	-	48
Paired-end reads	*Sisymbrium irio*	B – II	-	PRJNA202979[d]	61

[a] according to Huang *et al.* (2016).
[b] according to Franzke *et al.* (2011).
[c] according to Arias and Pires (2012).
[d] Haudry *et al.* (2013) (http://mustang.biol.mcgill.ca:8885/cgi-bin/hgGateway?hgsid=11412&clade=Plants&org=S.+irio&db=0)

Phylogenetic reconstruction

After performing alignments with MACSE (Ranwez et al. 2011), a visual inspection of each alignment was performed in order to discard bad alignments and manually correct or trim ambiguous alignment regions. One alignment was discarded due to a suspicion of paralogous sequences (rpl22) and three others due to a length smaller than 100bp (psbT, petN and petL). Then, the 74 remaining chloroplast genes were concatenated in one supermatrix of 66 906 bp. Starting from a partitioning scheme containing 222 partitions (a separate partition for each codon positions of each gene), the best partitioning schemes and evolutionary models were selected using the rcluster search mode implemented in PartitionFinder 2 (Lanfear et al. 2017), under the corrected Akaike Information Criterion (AICc). The resulting partitioning scheme contained 67 partitions and was used in the following analyses. Multiple phylogenetic analyses were performed, including maximum likelihood (ML) and Bayesian inference (BI). The GTR + G model was used for all partitions in ML analyses with RAxML version 8.2.10 (Stamatakis 2014) under automatically determined numbers of bootstrap replicates to assess node supports (Stamatakis 2006; Stamatakis et al. 2008; Stamatakis 2014). BI analyses were performed using MrBayes v3.2.6 (Ronquist et al. 2012) using a specific model for each DNA partition, according to the best-partition scheme selected by PartitionFinder 2 (GTR, GTR + Γ or GTR + I + Γ). Two runs of 4 MC^3 chains with 10, 000, 000 generations were completed. Trees and parameters were sampled every 20 generations. Plots of the likelihood-by-generation evolution were drawn to check chains convergence. It was also supported by an average standard deviation of split frequencies between runs smaller than 0.01. The first 25% of trees from both runs were discarded as burn-in. A majority-rule consensus of the remaining trees was performed to obtain the final tree with posterior probabilities (PP).

Following the recommendation of Roure *et al.* (2012), an approximation of the percentage of missing data is reported for each investigated species in Table 3. The potential impact of missing data on the topology was assessed by performing all aforementioned phylogenetic analyses after removing sites at which there was an alignment gap for at least two of the four following species, which exhibited a high proportion of alignment gaps (>30%): *C. annua*, *P. stylosa*, *Z. spinosa subsp. macroptera* and *S. purpurea* (Table 3.). The resulting matrix contained 37,431 bp. In order to keep the reading frame, we remove from the alignment the entire codon for each missing site.

Assignment of nuclear homoeologous gene copies to parental subgenomes within the Brassiceae

We focused on the specific genes that have been retained in three homoeologous copies in *B. rapa* and *B. oleracea*, representing the LF, MF1 and MF2 subgenomes (triplets), according to the analysis of Liu et al. (2014) (1,344 genes). *A. thaliana* orthologous sequences to the Brassica's 1,344 genes were extracted from the TAIR database if they were longer than 500 bp (1,163 genes). Only *B. rapa* and *B. oleracea* gene sequences with a minimum length of 500bp and a length coverage of the orthologous *A. thaliana* sequence higher than 60% were selected (1,085 remaining genes).

1,085 homolog groups (called DS1 for DataSet1), containing three copies of *B. rapa* and *B. oleracea* as well as a single copy of *A. thaliana*, was aligned with MACSE (Ranwez et al. 2011). Alignments were trimmed on both sides following the *A. thaliana* sequence and at poorly aligned sites using trimAl v1.2 with default settings (Capella-Gutiérrez et al. 2009). When the final alignment was greater or equal to 500bp, phylogenetic trees were built with RAxML (Stamatakis 2006; Stamatakis 2014) with a GTR+Γ model of sequence evolution (Stamatakis 2014). From this first set of homolog trees, we discarded all trees (and the corresponding alignments) in which *(i)* the *B. rapa* and *B. oleracea* copies annotated as originating from the same subgenome (LF, MF1 or MF2) were not monophyletic – which may arise due to annotation errors or to ectopic gene conversion events between homoeologs (Soltis et al. 2014; Scienski et al. 2015) – and *(ii)* one or more subgenomes were not represented in the tree due to sequence elimination during the alignment trimming. For each of the 1,077 remaining homolog groups, single orthologous sequences from the outgroups and up to three homoeologous sequences from each of the Brassiceae species were added as explained below (this correspond to the dataset DS2 for DataSet2).

Using the selected TAIR *A. thaliana* gene sequences as references, we extracted orthologous sequences from the following published genomes of Brassicaceae species that do not share any of the WGD events experienced by the Brassiceae tribe with BLASTN (Altschul et al. 1990) (minimum percentage of identity: 80% ; minimum length : 60% of the reference sequence) : *Eutrema salsugineum* (Yang et al. 2013) , *Schrenkiella parvula* (Dassanayake et al. 2011) and *Sisymbrium irio* (Haudry et al. 2013).

The homoeologous genes present in three copies in *B. rapa* and *B. oleracea* were mapped onto (1) each of our original transcriptomes, (2) the genome of *Raphanus sativus* (Brassiceae, (Kitashiba et al. 2014)), (3) the genome of *Raphanus raphanistrum* (Brassiceae,

(Moghe et al. 2014)) and (4) the genome of *Brassica nigra* (Brassiceae, Nigra clade, (Yang et al. 2016), NCBI: ASM168289v1) using BLASTN to extract orthologous sequences (minimum percentage of identity: 80% ; minimum length : 60% of the reference sequence) (Fig. 1, step 1). In order to avoid the introduction of chimeric contigs – particularly from transcriptomes – we only extracted the portion of each best-hit contig that aligned with the reference sequence. Because of alternative splicing, several isoforms can be the best-match sequence of the same Brassica reference sequence. We only extracted the longest isoform. Table 2 summarizes the data and the species included in the present nuclear phylogenetic analysis and their source.

Table 2. Genomic data and transcriptome assemblies used for the phylogenetic analysis based on nuclear genes.

Data	Species (number of individuals)	Brassicaceae clade[a] – lineage[b]	Brassiceae clade[c]	No. Accession/source
Genomic data (coding sequences)	*Arabidopsis thaliana*	A – I	-	TAIR 10 peptide database
	Eutrema salsugineum	B – II	-	PRJNA73205
	Schrenkiella parvula	B – II	-	AFAN00000000.1.
	Sisymbrium irio	B – II	-	PRJNA202979
	Brassica nigra	B – II	Nigra	PRJNA285130
	Brassica rapa	B – II	Oleracea	Brassicadb.org
	Raphanus raphanistrum	B – II	*Undetermined*	PRJNA209513
	Raphanus sativus	B – II	*Undetermined*	http://radish.kazusa.or.jp
	Brassica oleracea	B – II	Oleracea	Brassicadb.org
Transcriptome assemblies	*Cakile maritima* (12)	B – II	Cakile	-
	Carrichtera annua (3)	B – II	Vella	-
	Pychine stylosa (2)	B – II	Savignya	-
	Schouwia purpurea (3)	B – II	Zilla	-
	Zilla macroptera (2)	B – II	Zilla	-
	Orychophragmus violaceus (2)	B – II	-	-

[a] according to Huang *et al.* (2016).
[b] according to Franzke *et al.* (2011).
[c] according to Arias and Pires (2012).

Homolog trees for all the 1,077 homolog groups from the DS2 were built and filtered according to the two aforementioned rules (same as those applied on DS1 trees). This led to 1,046 remaining homolog trees (Fig. 1, step 2). For each homolog tree, we identified each of the three expected orthologs' sub-trees by performing node annotations (Fig. 1, step 3) with the ETE v3 toolkit (Huerta-Cepas et al. 2016) as described below:

(i) The two reference sequences of a given subgenome were localized on the tree (*e.g.* for the subgenome MF1, we localized the *B. rapa* and *B. oleracea* MF1 sequences).

(ii) From the node representing the common ancestor of the two reference sequences, we climbed backward through the tree until reaching the last node

defining a clade from which sequences of any outgroup species (outside of the Brassiceae), of *O. violaceus* and of reference sequences from other subgenomes (*e.g.* for the MF1 subgenome, *B. rapa* / *B. oleracea* MF2 or LF sequences) were excluded.

(iii) We annotated the corresponding node with the subgenome label of the two *B. rapa* / *B. oleracea* reference sequences.

(iv) The process is repeated in order to annotate the 3 sub-trees corresponding to the three sub-genomes (LF, MF1 and MF2) for each homolog tree.

Homolog trees with more than 10% of Brassiceae transcriptome sequences localized outside the ortholog sub-trees were discarded (183 homolog trees, 17.5%). A total of 863 remaining homolog trees were kept.

The phylogenetic position of the most basal node of each of the three subgenomes LF, MF1 and MF2 in the homolog trees should indicate the position of each (presumably extinct) parental lineage that gave birth to the extant Brassiceae tribe through allopolyploid hybridization followed by lineage diversification. Indeed, orthologs' sub-trees contain all homologous genes that evolved from a single ancestral gene after a given speciation event, thus include orthologous genes and paralogues that evolved by lineage-specific duplication after the relevant speciation event (in-paralogs) (Gabaldón and Koonin 2013).

Step 1: Sequence acquisition and construction of homolog groups

BLASTN: For each of the 1,077 homolog groups of the DS1 (see main text), the three copies of *B. rapa* and *B. oleracea* (LF, MF1 and MF2) were mapped onto the transcriptome assemblies (our original data) and the genomic coding sequences of *Raphanus sativus*, *R raphanistrum*, and *B. nigra*. Each homolog group was reported in a file as shown here.

→ 1077 homolog groups

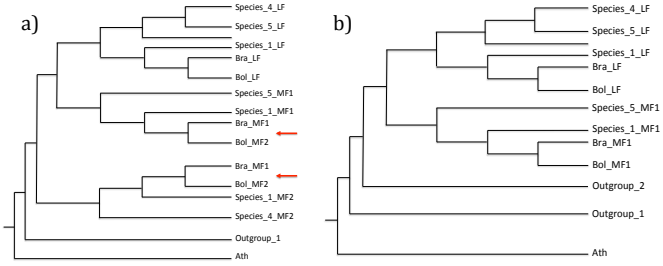
Reference sequences

Gene *i*, *i* ∈ [1; 1077]
Ath ----------
Bra_LF ----------
Bra_MF1 ----------
Bra_MF2 ----------
Bol_LF ----------
Bol_MF1 ----------
Bol_MF2 ----------
Outgroup_1 ----------
Outgroup_2 ----------
Species_1_LF ----------
Species_1_MF2 ----------
Species_2_LF ----------
...

Step 2: Construction of the homolog trees

For each homolog group:

i) Sequence alignment and trimming
ii) Construction of homolog trees
iii) Tree filtering as shown below:



We discarded homolog trees in which:

a) Copies of the same subgenome of *B. rapa* and *B. oleracea* (LF, MF1 or MF2) were not monophyletic
b) One or more subgenomes were not represented in the tree (*e.g.* MF2 is absent)

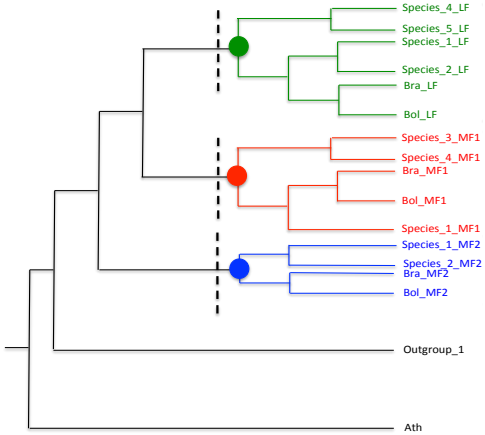
→ 1046 remaining homolog trees

Step 3: Identifying ortholog sub-trees (nodes annotation)

We located the common ancestor of the two reference sequences (Bra and Bol) for each subgenome and we climbed backward through the tree until reaching the sub-genome ancestral node (*i. e.* single ancestral gene) presenting all the conditions detailed in the main text. Three ortholog subtrees (LF, in green, MF1, in red and MF2, in blue) were thus identified in each homolog tree.

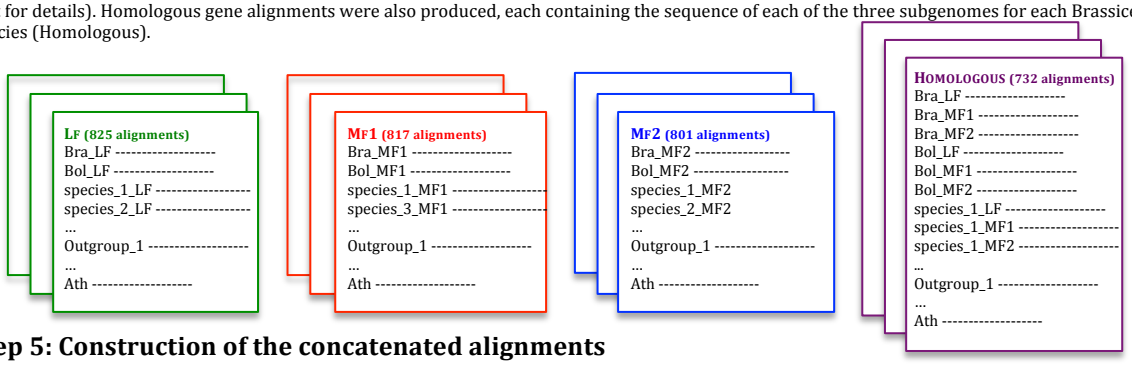
Homolog trees with more than 10% of transcriptome sequences not included in an ortholog sub-tree were discarded.

→ 863 remaining homolog trees



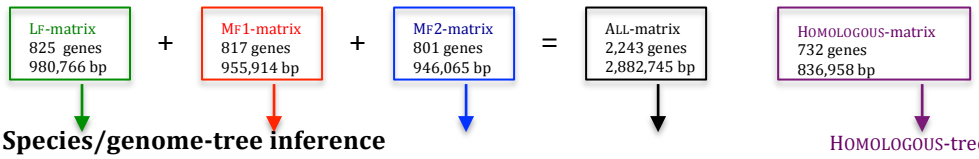
Step 4: Construction of orthologous gene alignments and homologous gene alignment

Extraction of the ortholog sub-trees (LF, MF1 and MF2) in each of the 863 remaining homolog tree to produce orthologous gene alignments (see main text for details). Homologous gene alignments were also produced, each containing the sequence of each of the three subgenomes for each Brassiceae species (Homologous).



Step 5: Construction of the concatenated alignments

LF, MF1 and MF2 genes were concatenated i) separately in the LF-matrix (in green), in the MF1-matrix (in red) and in the MF2-matrix (in blue) and *ii)* all together in the ALL-matrix (in black). All homologous alignments were concatenated in the HOMOLOGOUS matrix (in purple).



Step 6: Species/genome-tree inference

Tree-building methods were used for reconstructing the phylogenetic species trees from the five matrices.

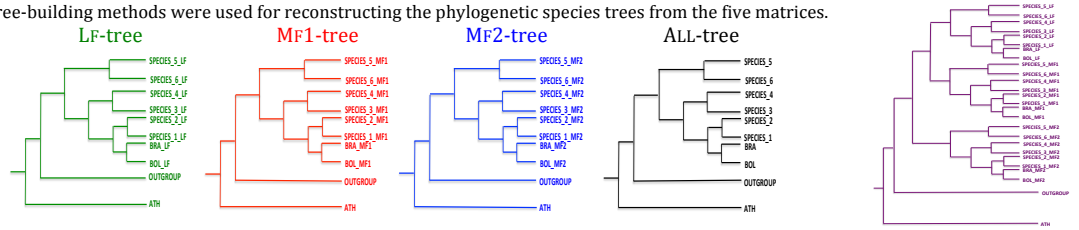


Fig. 1. Methodological framework used in this study for reconstructing the evolutionary history of a mesopolyploid lineage.

Inference of the Brassiceae species tree using nuclear genes

Following the defined orthologs' sub-trees, for each homolog group, sequences of each Brassiceae species were extracted from the "homologous" alignments and written in three distinct fasta files representing each orthologous group (LF.fasta, MF1.fasta and MF2.fasta, Fig. 1, step 4). For constructing these orthologous gene alignments, we checked that all sequences belonging to the same species were monophyletic in a given ortholog sub-tree before extracting the longest one. If not, none sequence was extracted for the concerned species in the concerned ortholog sub-tree (177 cases). Thus, in-paralogs were ignored. Even though *O. violaceus* underwent a WGD event, we will show in the next chapter that this event is not shared with the Brassiceae. Therefore we extracted only one sequence of *O. violaceus*, the longest, from each homolog tree. Sequences of outgroup species were reported in each orthologous gene alignment. Three orthologous gene alignments were thus produced from each homolog tree (Fig. 1, step 4).

Ortholog sub-trees in which there were less than 2 transcriptome sequences belonging to two different species were filtered out (140 discarded ortholog sub-trees). After this filtering, LF, MF1 and MF2 genes were concatenated (1) separately in a LF matrix (825 genes; 980,766 bp), a MF1 matrix (817 genes; 955,914 bp) and a MF2 matrix (801 genes; 946,065 bp), respectively and (2) all together in the ALL matrix (2, 443 genes; 2, 882, 745 bp) (Fig. 1, step 5). The proportion of alignment gaps for each species and each matrix is reported in Table 3 and the number of LF, MF1 and MF2 gene sequences for each Brassiceae species is indicated in Table 4. Alignment gaps can be due to indels, gene copy deletion and missing data (due in part to low level of expression of a given homoeologous gene copy), and the occurrence of a high proportion of alignment gaps might indicate a high proportion of missing information, which could impair the accuracy of the phylogenetic inference (Lemmon et al. 2009). We therefore produced four additional filtered matrices (LF, MF1, MF2 and ALL filtered-matrix) in which sites displaying missing data for at least one of the four following species *C. annua*, *P. stylosa*, *Z. spinosa subsp. macroptera* and *S. purpurea*, were removed. We focused on these four species because they display a large amount of missing data (45 to 73% of alignment gaps, depending on the matrix, Table 3). The lengths of the resulting matrices were 160,245, 65,355, 60,449 and 286,049 bp for the LF, MF1, MF2 and the ALL filtered-matrices, respectively. In order to resolve the evolutionary history of the mesopolyploid group, we built a ninth matrix (called HOMOLOGOUS) as follows: for each of the 863 homolog trees, the sequences of each Brassiceae species were extracted from each ortholog sub-tree as previously explained and written in a unique fasta file (Fig. 1, step 5). For

this purpose, we gave a distinct species label for the LF, MF1 and MF2 sequences belonging to the same species depending on its sub-genome of origin (*e.g.* species_1_LF, species_1_MF2 and species_1_MF2, see Fig. 1, step 5). 732 alignments were thus obtained and concatenated for the construction of the HOMOLOGOUS matrix (836,958 bp).

Table 3. Percentage of alignment gaps for each species in both chloroplast and nuclear alignments, calculated as the total number of alignment gaps relative to the total alignment length (bp). For the concatenated alignments of nuclear genes, the percentage of gaps is given for each of the four matrices (ALL, LF, MF1 and MF2). For each species, the reported values mirror together sites with missing information and indels, except for *A. thaliana* where the reported values mirror only the indels as we assume that there is no missing information in the TAIR database.

Alignment gaps present in chloroplast genes for each Brassicacea species		
Species	Data type	Alignment gaps (%)
Alliaria grandifolia	chloroplast cds	0.40
Arabidopsis thaliana	chloroplast cds	0.06
Barbarea verna	chloroplast cds	0.12
Brassica napus	chloroplast cds	0.10
Brassica nigra	chloroplast cds	0.04
Brassica oleracea	chloroplast cds	6.41
Brassica rapa	chloroplast cds	7.06
Cakile arabica	chloroplast cds	0.23
Cardamine limprichtiana	chloroplast cds	31.95
Eutrema salsugineum	chloroplast cds	0.07
Isatis tinctoria	chloroplast cds	11.09
Lobularia maritima	chloroplast cds	1.09
Orychophragmus diffusus	chloroplast cds	0
Orychophragmus hupehensis	chloroplast cds	0
Orychophragmus longisiliqus	chloroplast cds	0
Orychophragmus taibaiensis	chloroplast cds	0.01
Orychophragmus violaceus	chloroplast cds	0
Orychophragmus zhongtiaoshanus	chloroplast cds	0
Raphanus sativus	chloroplast cds	0.03
Schrenkiella parvula	chloroplast cds	0.07
Sisymbrium irio	assembly from raw genomic reads	24.62
Cakile maritima	transcriptome assembly	28.77
Carrichtera annua	transcriptome assembly	31.97
Pychine stylosa	transcriptome assembly	38.99
Schouwia purpurea	transcriptome assembly	36.04
Zilla macroptera	transcriptome assembly	57.05

Alignment gaps present in nuclear genes for each Brassicaceae species					
		Alignment gaps (%)			
Species	Data type	ALL	LF	MF1	MF2
Arabidopsis thaliana	genomic cds	0.76	0.73	0.77	0.78
Brassica nigra	genomic cds	19.14	18.73	18.25	20.47
Brassica oleracea	genomic cds	9.10	9.30	8.90	9.10
Brassica rapa	genomic cds	4.40	3.12	5.33	4.78

Table 3. *(continued)*

Species	Data type	Alignment gaps (%) ALL	LF	MF1	MF2
Raphanus sativus	genomic cds	29.36	24.33	29.65	34.29
Raphanus raphanistrum	genomic cds	43.32	40.50	47.13	42.40
Eutrema salsugineum	genomic cds	10.72	10.56	10.71	10.89
Schrenkiella parvula	genomic cds	12.65	12.68	12.84	12.43
Sisymbrium irio	genomic cds	23.50	23.67	23.07	23.76
Orychophragmus violaceus	transcriptome assembly	19.98	20.15	19.61	20.17
Cakile maritima	transcriptome assembly	47.46	37.31	52.87	52.52
Carrichtera annua	transcriptome assembly	65.19	55.86	69.37	70.65
Pychine stylosa	transcriptome assembly	60.75	50.95	64.88	66.75
Schouwia purpurea	transcriptome assembly	55.70	45.07	58.41	63.98
Zilla macroptera	transcriptome assembly	66.92	58.10	70.17	72.79

Table 4. Number of LF, MF1 and MF2 gene sequences collected in each investigated Brassiceae species in the final set of the 863 homolog groups used for phylogenetic reconstruction.

Species	Number of gene sequences in each subgenome		
	LF	**MF1**	**MF2**
B. rapa	845	834	837
B. oleracea	813	814	804
B. nigra	719	709	692
R. sativus	671	618	584
R. raphanistrum	544	488	521
C. maritima	546	425	433
S. purpurea	493	379	334
P. stylosa	432	313	304
Z. spinosa subsp macroptera	368	276	237
C. annua	383	272	258

The best partitioning scheme was assessed by PartitionFinder 2 with the rcluster search mode, under the corrected Akaike Information Criterion (AICc), following the recommendation of the authors (Lanfear et al. 2017). Specific codon partitions could not be set as the gene sequences lost their reading frame after the TrimAl trimming. Then, a phylogenetic species tree was constructed from each matrix (Fig. 1, step 6) using RAxML v8.2.10 as maximum likelihood estimation method with a partitioned analysis in order to estimate a distinct model of nucleotide substitution for each DNA partition, under the GAMMA model of rate heterogeneity (GTR+G) and automatically determined numbers of bootstrap replicates (Stamatakis 2006; Stamatakis et al. 2008; Stamatakis 2014). Bayesian inference analyses were performed using MrBayes v3.2.6 (Ronquist and Huelsenbeck 2003;

Ronquist et al. 2012) with a different model for each DNA partition, according to the best-partition scheme selected by PartitionFinder 2. Two runs of 75,000,000 generations were completed with four chains each and trees were sampled every 2000 generations for the LF, MF1, MF2, ALL and HOMOLOGOUS matrices. For the LF, MF1, MF2 and ALL filtered matrices, 10,000,000 generations were completed and trees were sampled every 400 generations. Plots of the likelihood-by-generation were drawn to check chain convergence, indicated as well by an average standard deviation of split frequencies smaller than 0.01, an Potential Scale Reduction Factor (PSRF) at 1.0 and effective sample size (ESS) values above 100. The first 25% of trees from all runs were discarded as burn-in. A majority-rule consensus of the remaining trees from both runs was used to obtain the posterior probability tree.

Results

Transcriptome assemblies

The transcriptome of 22 individuals from five different species of Brassiceae and the two transcriptomes of the outgroup Orychophragmus (see Table S1) were sequenced and paired-end Illumina reads were assembled after cleaning. Depending on the species and the tissue sampled, the transcriptome assemblies yielded between 28,710 and 57,725 contigs (Table S2). Total length (in bp) varies between 38,803,080 and 79,652,025, the N50 (in bp) varies between 1,596 and 1,857 and the percentage in GC varies between 42.11 and 43.84 % (Table S2). The two individuals of *Zilla spinosa subsp macroptera*, represented by leaf tissues, display the lowest total length and the lowest number of contigs. This is not surprising as reproductive organs have a much broader range of expression than leaves (Schmid et al. 2005).

Gene orthology assignment

Starting with a selection of 1,085 genes based on their occurrence in three homoeologous copies in *B. oleracea* and *B. rapa*, a length greater than 500 bp, and a sequence coverage higher than 60% with the orthologous sequence of *Arabidopsis thaliana*, we obtained a dataset of 863 genes for which each homoeologous copy sequenced in a Brassiceae species could be assigned to a specific subgenome based on the *B. oleracea* and *B. rapa* reference sequences. The 222 discarded genes were removed because of loss of any of the *B. oleracea* or *B. rapa* homoeologous gene copies during the phase of trimming or due to phylogenetic reconstruction problems (Fig.1).

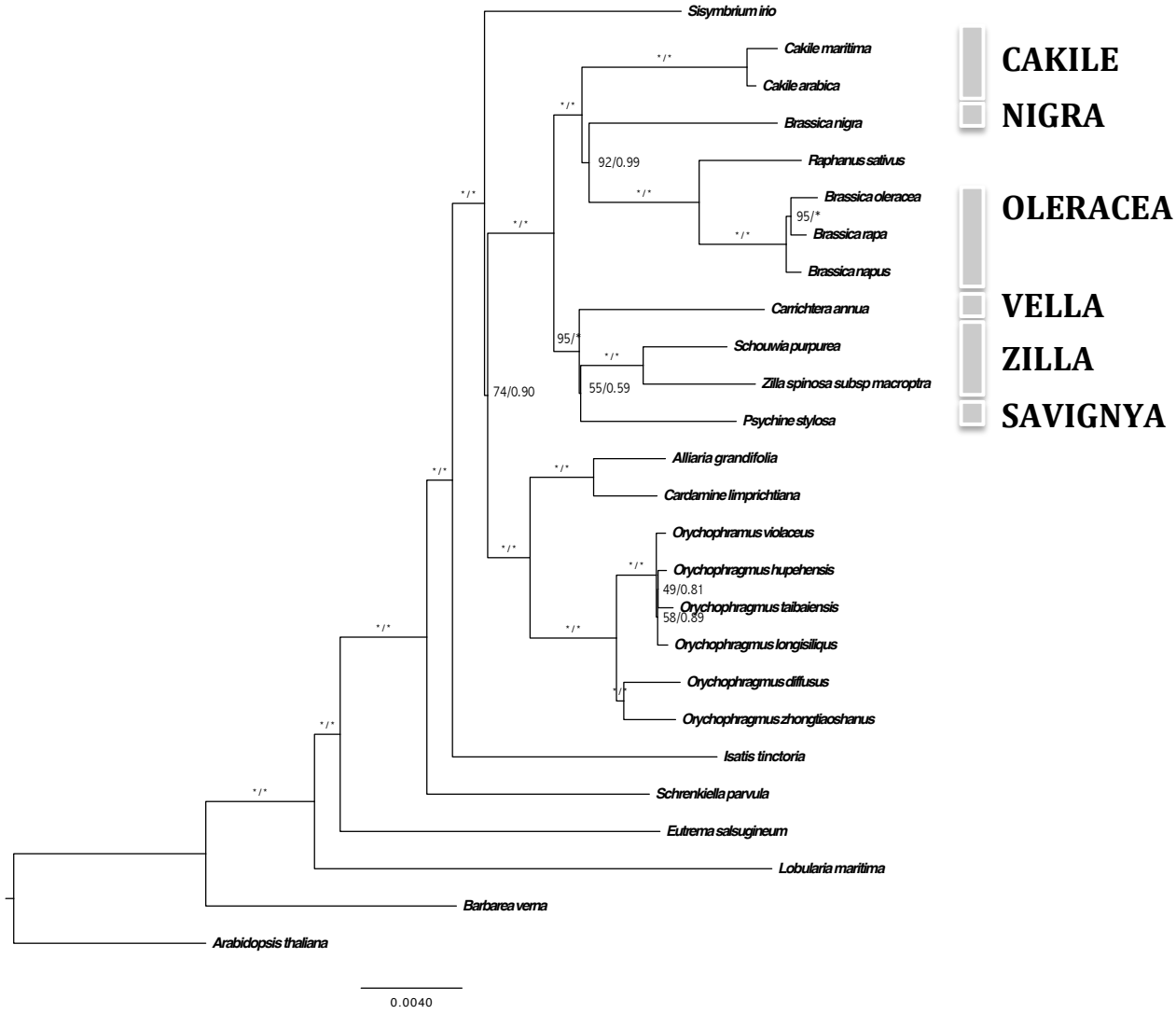
Phylogenetic reconstruction of the Brassiceae using chloroplast coding-sequences

Using a concatenated alignment of 74 chloroplast coding sequences (66,906bp) constructed by combining published sequences as well as original data from our 22 transcriptome assemblies, we obtain a well-resolved phylogeny of the Brassiceae's main lineages (Fig. 2). ML and BI analyses yielded the same tree topology. Both bootstrap values (BP) and posterior probabilities (PP) strongly supported the Brassiceae monophyly (BP = 100, PP = 1) as well as the phylogenetic relationships among the six investigated Brassiceae clades (Fig. 2). The most basal splitting event within the tribe separated the clades "Vella+Zilla+Savignya" (BP=95, PP=1.0) and "Cakile + Nigra + Oleracea" (BP=100, PP=1.0). Vella, Zilla and Savignya are three closely related subtribes and the choloroplast tree cannot allow us to infer their divergence order (clade "Savignya+Zilla": BP=55, PP=0.59). Cakile+Nigra+Oleracea form a strongly supported monophyletic group (BP=100, PP=1.0), defined here as the core Brassiceae. *Raphanus sativus*, whose subtribe assignation is ambiguous in the literature (Yang et al. 2002, Seol et al. 2015), appears clearly to be more related to the Oleracea lineage than to the Nigra lineage based on chloroplast sequences (BP=100, PP=1).

Sisymbrium irio, *Orychophragmus sp.*, the two Sinalliaria species (*A. grandifolia* and *C. limprichtiana*) and the Brassiceae tribe form a strongly supported monophyletic group (BP=100, PP=1). All species of the genus Orychophragmus together with the two Sinalliaria species also form a robust monophyletic group and seem to be more related to the Brassiceae tribe than *S. irio*, however with low BP values (BP=74, PP=1). Our data confirm therefore the position of the genera Sisymbrium and Orychophragmus as sister clades of the Brassiceae tribe.

From the concatenated alignment filtered on alignment gaps (37,431 bp), ML and BI analyses yielded the same tree topology than with the full matrix (Supplementary Fig. S1). The only effect of missing sites removal is a slight loss of resolution for the less supported nodes.

Fig. 2. Maximum likelihood phylogeny of the tribe Brassiceae and its relatives based on the analysis of 74 concatenated chloroplast genes. The six investigated clades (subtribes) are indicated on the right side of the figure. Numbers on branches correspond to bootstrap support values (BP) and Bayesian posterior probabilities (PP). (*) indicates nodes with maximal support (BP = 100 / PP = 1.0).



Phylogenetic reconstruction of the Brassiceae using nuclear coding sequences

The species trees obtained for the ALL-matrix (concatenation of the LF, MF1 and MF2 matrices) and for the individual LF, MF1 and MF2 matrices are presented in Fig. 3a and Fig. 3b, respectively. With both the ML and BI analyses, the intra-Brassiceae topology was congruent among the LF, MF1, MF2 and ALL species trees (Fig. 3a and Supplementary Fig. S2). Concerning the intra-Brassiceae phylogenetic relationships, each topology was very robust with maximal support values at each node (Fig. 3a, Supplementary Fig. S2). Vella is the first diverging clade followed by the Zilla clade and finally the Savignya clade sister group of the core brassiceae (Oleracea + Nigra + Cakile). These three clades, Vella, Zilla and

Savignya, formed a monophyletic group in the chloroplatic phylogeny. Cakile, Nigra and Oleracea clades shared a common ancestor and the two latter are most closely related two each other, in agreement with our chloroplast phylogeny (Figure 3a, Supplementary Fig. S2).

The phylogenetic position of *O. violaceus* and *S. irio* was variable according to the dataset used. Indeed, either *O. violaceus* or *S. irio* were the sister species of the Brassiceae tribe in the MF1 and LF trees, respectively (Fig. 3b), whereas in the MF2 tree, *S. irio* and *O. violaceus* were forming a monophyletic group (Fig. 3b). The position of these two species in the ALL tree was similar to that of the MF1 tree (Fig. 3a). We produced as well four other data matrices by filtering on the proportion of alignment gaps (LF-, MF1-, MF2- and ALL-filtered-matrix), because four species exhibited a high proportion of missing data (Table 3). In order to assess the potential effect of missing data on the phylogenetic inferences, we performed the phylogenetic analyses on these filtered matrices. For each matrix, trees obtained from ML and BI analyses yielded the same topology (Supplementary Fig. S3). The only notable difference between the individual LF, MF1, MF2 filtered and non-filtered analyses, beside the fact that support values dropped when datasets got smaller, concerns the MF1 tree in which *S. irio* and *O. violaceus* appear as a monophyletic group in the filtered analyses (Fig. 3b, Supplementary Fig. S3). Concerning the ALL-filtered species tree, the outgroups *S. irio* and *O. violaceus* have their phylogenetic position exchanged compared to the ALL species tree (Fig. 3a, Supplementary Fig. S3).

The species tree obtained from the HOMOLOGOUS-matrix (836,958 bp) displaying all three lineages LF, MF1 and MF2 is reported in Figure 4. The fully resolved topology of the Brassiceae species in each of the three sub-trees was the same and congruent with all other inferred topologies (ALL, LF, MF1 and MF2). Interestingly, *O. violaceus* and *S. irio* appear to be most closely related to the LF parental subgenome than to the other two subgenomes (Fig. 4), and their relative position is congruent with the LF tree (Fig. 3b, Fig. 4, Supplementary Fig. 2). In the same way, the MF2 subgenome equally diverged from *S. irio* and *O. violaceus*, which is congruent with the topology of the MF2 tree (Fig. 3b, Fig. 4, Supplementary Fig. 2). The MF1 subgenome, sister group to the MF2, equally diverged as well from *S. irio* and *O. violaceus*; this result is only congruent with the MF1-filtered tree (Fig. 3b, Fig. 4, Supplementary Fig. 2).

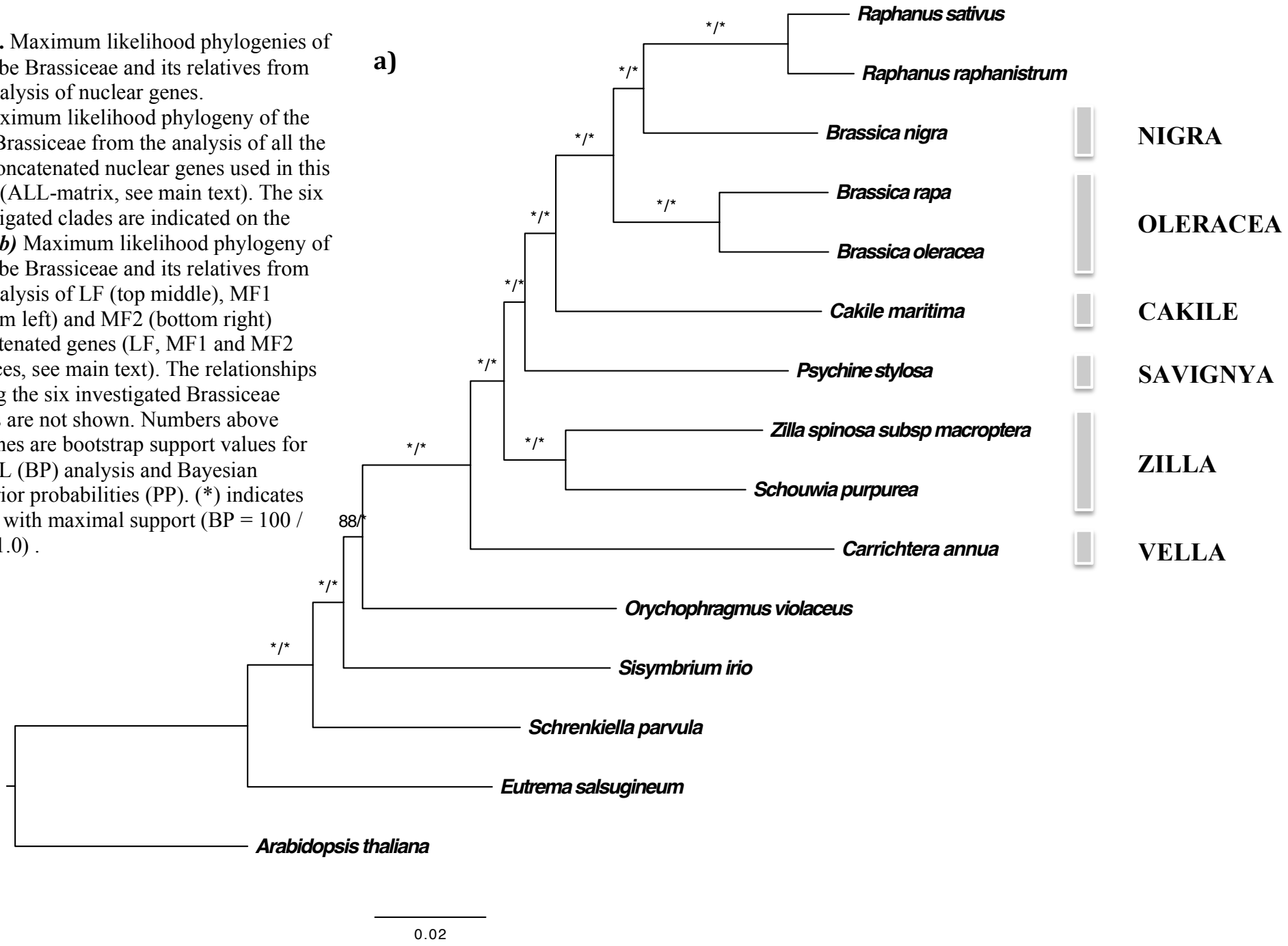


Fig. 3. Maximum likelihood phylogenies of the tribe Brassiceae and its relatives from the analysis of nuclear genes.
a) Maximum likelihood phylogeny of the tribe Brassiceae from the analysis of all the 863 concatenated nuclear genes used in this study (ALL-matrix, see main text). The six investigated clades are indicated on the right. ***b)*** Maximum likelihood phylogeny of the tribe Brassiceae and its relatives from the analysis of LF (top middle), MF1 (bottom left) and MF2 (bottom right) concatenated genes (LF, MF1 and MF2 matrices, see main text). The relationships among the six investigated Brassiceae clades are not shown. Numbers above branches are bootstrap support values for the ML (BP) analysis and Bayesian posterior probabilities (PP). (*) indicates nodes with maximal support (BP = 100 / PP = 1.0) .

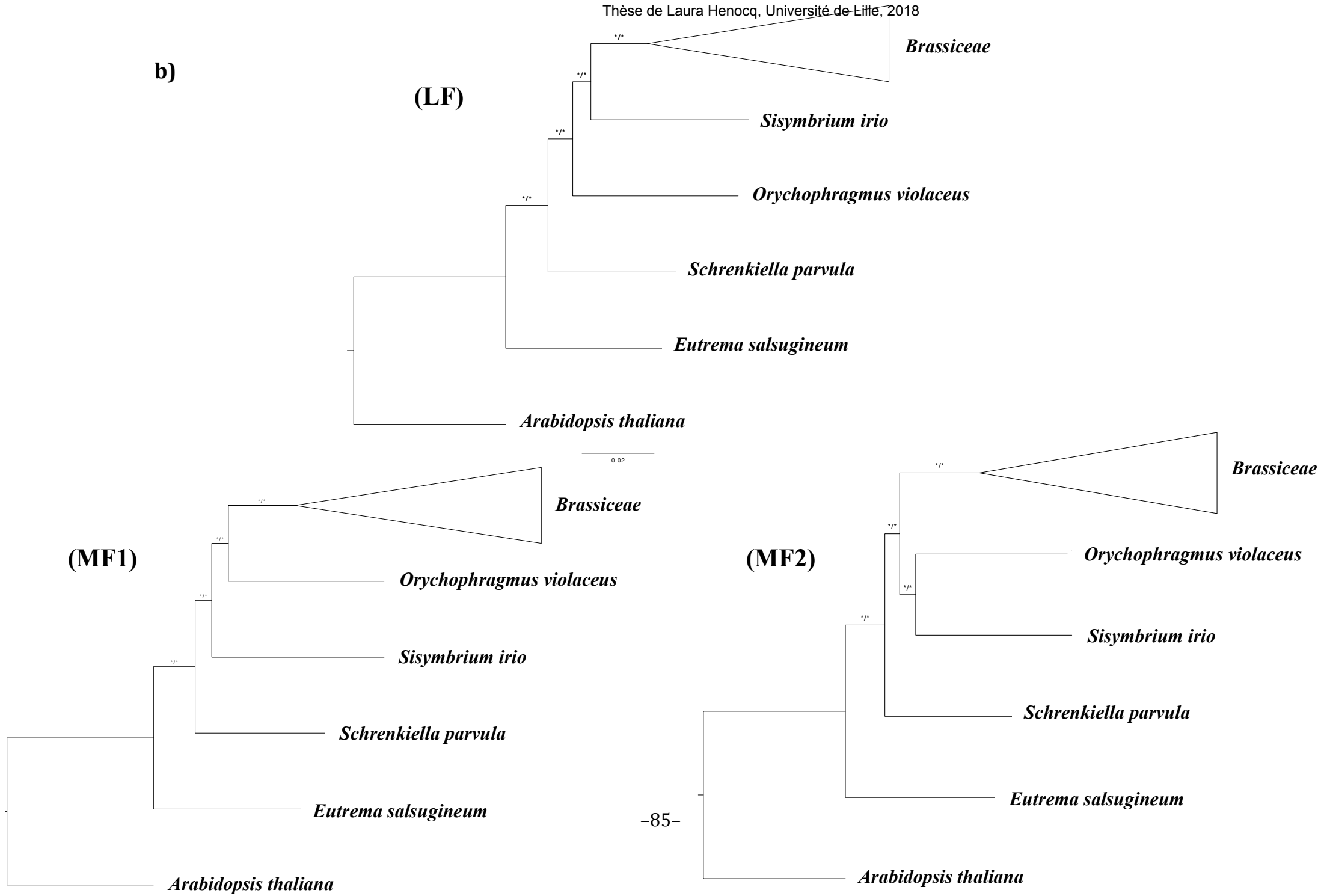
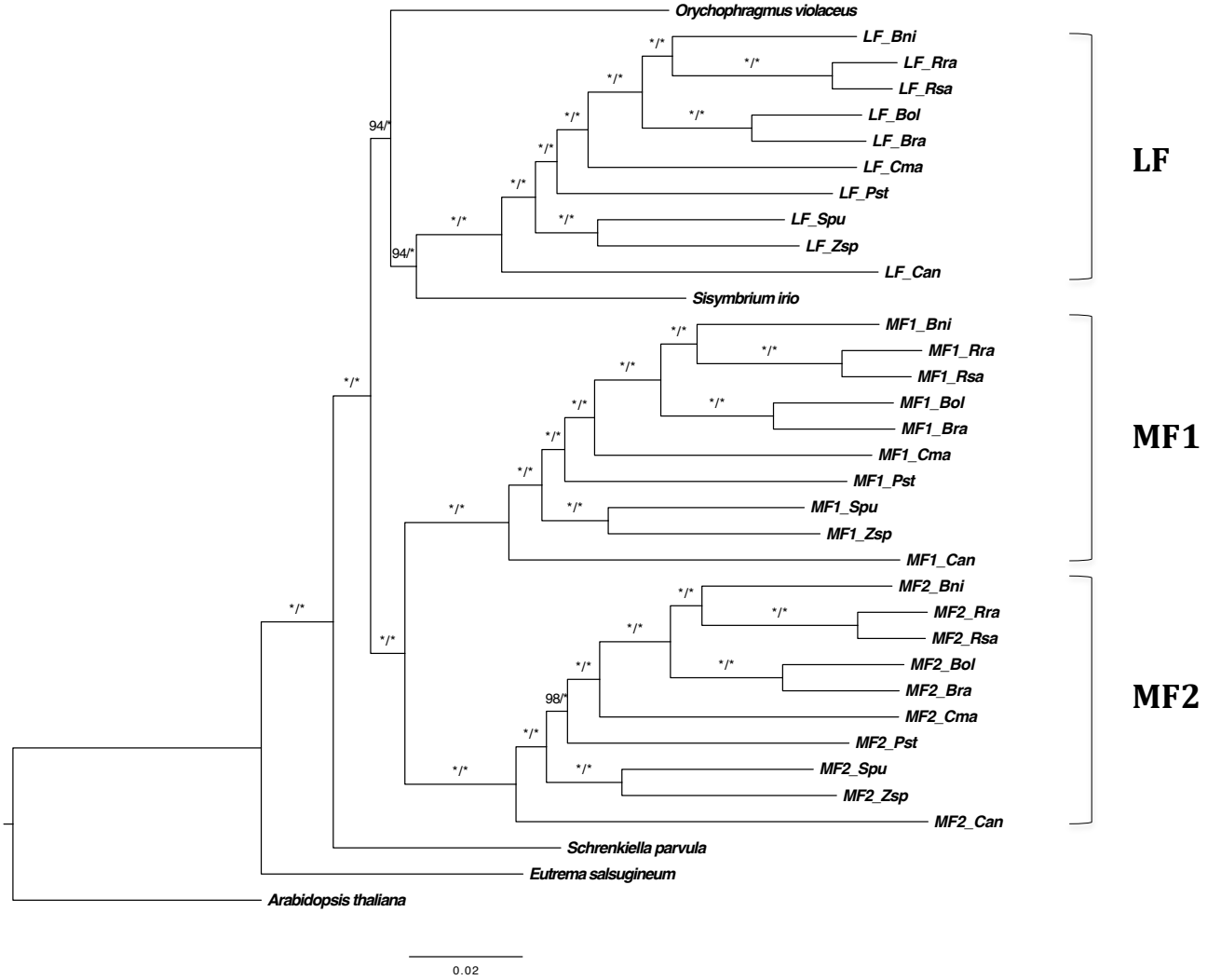

b)
(LF)
/
Brassiceae
Sisymbrium irio
Orychophragmus violaceus
Schrenkiella parvula
Eutrema salsugineum
Arabidopsis thaliana
0.02
(MF1)
Brassiceae
Orychophragmus violaceus
Sisymbrium irio
Schrenkiella parvula
Eutrema salsugineum
Arabidopsis thaliana
(MF2)
Brassiceae
Orychophragmus violaceus
Sisymbrium irio
Schrenkiella parvula
Eutrema salsugineum
Arabidopsis thaliana

Fig. 4. Maximum likelihood phylogeny of the tribe Brassiceae from the analysis of the HOMOLOGOUS-matrix (see main text), showing three sub-trees corresponding to the three subgenomes present in the Brassiceae species (LF, MF1, MF2). Numbers above branches are bootstrap support values for the ML analysis (BP) and Bayesian posterior probabilities (PP). (*) indicates nodes with maximal support (BP = 100 / PP = 1.0). Bni: *Brassica nigra*, Rsa: *Raphanus sativus*, Rra: *R. raphanistrum*, Bol: *B. oleracea*, Bra: *B. rapa*, Cma: *Cakile maritima,* Pst: *Psychine stylosa,* Spu: *Schouwia purpurea*, Zsp: *Zilla spinosa subsp macroptera*.



Discussion

Drawing the evolutionary history of a mesopolyploid group using fully retained genes

After the hexaploidization event experienced by the Brassiceae tribe, some genes have been preferentially retained in triplets in *B. rapa* and *B. oleracea,* with one copy belonging to each of the three annotated subgenomes (Wang et al. 2011a; Cheng et al. 2012; Liu et al. 2014; Murat et al. 2015). Indeed, several studies have shown that some genes for which their product is highly connected within networks appear to be resistant to the gene fractionation process following polyploidization (Liu et al. 2014; Murat et al. 2015). Accordingly, Moghe et *al.* (2014) identified many genes that have been retained in triplicates in *Brassica rapa* and another Brassiceae species, *Raphanus raphanistrum*. Moreover, Haudry et *al.* (2013) reported that genes kept in three copies in *B. rapa* were five times more likely to be also kept in three copies in *Leavenworthia alabamica*, a species belonging to a lineage that experienced an independent whole genome triplication at a similar temporal period than Brassica, and, similarly, Mandáková et al. (2017) reported convergent patterns of gene loss and retention among 13 independent mesopolyploid species distributed within the Brassicaceae family. Therefore, as we hypothesized that the Brassica WGT event was shared by all Brassiceae species (according to Lysak et al. 2005, and Lysak et al. 2007), we expected a high rate of recovery of triplets in the other Brassiceae species studied here. However, in our dataset, many copies of the Brassica homoeologous triplets were missing in each investigated Brassiceae species, regardless of the type of data (genomic or transcriptomic, Table 4). The absence of one or more copies in a given species can be explained for instance by a lineage-specific loss of one or more copies during the gene fractionation process, by the absence or low level of expression of one or more copies in the sampled plant material (for transcriptomic data) due to differential patterns of gene expression among homoeologous copies (Cheng et al. 2012) or the physiological timing, and/or by the insufficient recovery of raw sequencing reads or their misleading assembly for one or more copies. Lineage-specific gene conversion events between homoeologs could also explain that we failed to recover some of the Brassica homoeologous copies used as references (Wang et al. 2009; Wang et al. 2011b; Wang and Paterson 2011; Scienski et al. 2015). However, our analyses showed that the topology estimates that we obtained do not seem to be impacted by the amount of missing data. Hence, it appears that our approach, at least when applied to the mesohexaploid

Brassiceae clade, is not strongly impacted by the incompleteness of transcriptomic or genomic datasets, irrespective of the possible causes of such incompleteness.

By recovering the Brassica's gene copies orthologs in several species belonging to the major Brassiceae clades, we attempted to avoid the orthology and paralogy conflation issue by using a tree-based orthology inference approach, with the overall objective to determine the phylogenetic relationships among the main Brassiceae clades and provide the first phylogeny of Brassiceae based on a large number of nuclear genes. Our original methodological approach allowed us to infer the evolutionary history of the mesopolyploid Brassiceae tribe. The strength of the method comes from the fact that each sub-genome is used independently to assess the phylogeny of the studied taxonomic group. Comparison between sub-genome trees can be made and congruence assessed. In the present case, the three sub-genome trees were strictly identical with strong branch support, which increases our confidence in the result. The phylogenetic relationships among the presumably extinct parental lineages (represented by the 3 sub-genomes LF, MF1 and MF2) could also been satisfactorily resolved. From our results, it appears that the MF1 and MF2 parental lineages were more closely related to each other than to the LF lineage (Fig. 4). More surprisingly, the results suggested that the three parental lineages were not monophyletic and cannot all be considered as extinct lineages. Indeed, the LF parental lineage and two extant outgroup species (*Sisymbrium irio* and *Orychophragmus violaceus*) seem to share a common ancestor, whereas the two other parental lineages (MF1 and MF2) belong to another monophyletic group, sister group to the former (Fig. 4). These results show that the phylogenetic relationships between the Brassiceae tribe and the other Brassicaceae species depend on the parental sub-genome under consideration. This will always be the case when the parental lineages of a mesopolyploid clade are not monophyletic. These results help understanding why in many previous studies the relative phylogenetic positions of the Brassiceae tribe and the *Sisymbrium* and *Orychophragmus* genera were difficult to assess and mostly incongruent (Warwick et al. 2002, Warwick and Sauder 2005).

Our method follows the same overall approach as that of Yang *et al.* (2014): we used similarity scores to infer putative homologs and we performed alignments before reconstructing phylogenetic trees and identifying the ortholog groups (ortholog sub-trees). The main difference is that we used only the annotated triplets from our reference species and not their entire coding genome. This allows reconstructing the phylogenetic relationships among each parental species and between the mesopolyploid clade and the extant outgroup

species. By definition, our method depends entirely on the quality of the sub-genomes construction for each of the mesopolyploid reference species. Most of the published studies used a synteny approach for assigning chromosomal fragments to their parental genome. For a given mesopolyploid genome, all chromosomal fragments identified as being orthologous to a chromosomal fragment of a diploid reference genome are partitioned in two (for mesotetraploid), three (for mesohexaploid) or more subgenomes (Schnable et al. 2011; Wang et al. 2011a; Tang et al. 2012; Liu et al. 2014; Murat et al. 2015; Renny-Byfield et al. 2015). However, there are some substantial differences among studies in the way homoeologs are assigned to a given sub-genome. This is illustrated by the various numbers of homoeolog triplets reported in *B. rapa*: 1,578 in Wang *et al.* (2011a), 1,675 in Cheng *et al.* (2012) and 506 in Murat *et al.* (2015). To date, no evaluation has been done on the performance of the various methodologies. The database chosen in our study comes from the *B. oleracea* genome analysis of Liu et al. (2014) that was based on the same sub-genome assignment method as those used by Wang et al. (2011a) and had reported 1,344 triplets that are present in both species *B. rapa* and *B. oleracea*. Cheng *et al.* (2012) found 1,675 fully retained homoelogs that is slightly higher than the result of Wang *et al.* (2011a). In both studies, the genome of *A. thaliana* was chosen as a representative of the ancestral karyotype of Brassiceae (PCK – proto-Calepinae karyotype), but the identification of syntenic genes follows two markedly different approaches, which probably explains the slight difference in the resulting number of triplets. In Murat *et al.* (2015), the PCK ancestral genome was constructed with the genomes of five sequenced extant Brassicaceae species (*A. thaliana*, *A. lyrata*, *C. rubella*, *B. rapa* and *T. parvula*), containing 21,035 genes (against the 23,716 analysed genes of *A. thaliana* in Cheng *et al.* (2012)). The lower number of triplets found in Murat *et al.* (2015) compared to the studies of Wang *et al.* (2011a) and Cheng *et al.* (2012) can be partially explained by the use of a different genome, with lower gene content, as a representative of the ancestral karyotype of Brassiceae. We investigated to which extent the database of Liu *et al.* (2014) used in our study was concordant with the database of Murat *et al.* (2016) and found that only 38% of the triplets reported in Murat *et al.* (2016) were present in the database of Liu *et al.* (2014) with the same annotation for each of the three subgenomes. For a large part, the discordance found was due either to inversions in the subgenome annotation between MF1 and MF2 (25% of genes) or to genes found in triplicate in one study but only in duplicate or singleton in the other (37% of genes). All these observations called for some caution in the interpretation of our results. We therefore conducted again all phylogenetic analyses on a restricted dataset containing only the concordant triplets between the two studies (181 triplets,

154,545 bp, 607,778 bp, 210,426 bp, 204,578 bp and 192,774 bp for the HOMOLOGOUS, ALL, LF, MF1 and MF2 matrices, respectively), and we found similar results concerning the topology of the Brassiceae phylogeny, however with lower branch support due to the lower matrix size (Fig S4). Yet the phylogenetic relationships between the Brassiceae and their close relatives were not recovered as expected probably due to a lack of phylogenetic information (Fig S4). We strongly recommend, while conducting studies on mesopolyploid species, to evaluate the congruence between reference studies used for sub-genomes reconstructions.

A well-resolved phylogeny of the main Brassiceae clades

Our results represent the first phylogenetic study of the Brassiceae tribe that uses large numbers of nuclear gene sequences and representatives from the major Brassiceae subclades. Up to now, and due to its mesopolyploid origin, the Brassiceae were studied using one or few chloroplast (mainly matK) and/or nuclear (mainly ITS) genes. Therefore, phylogenetic results were quite variable according to the studies (Warwick and Sauder 2005, Arias and Pires 2012, Willis et al. 2014). However at least two lineages, the "Nigra" and the "Rapa/Oleracea" lineages, were reported in several past molecular systematics studies based on variability of cpDNA and mDNA restriction profiles (Pradhan et al. 1992) and RFLP data of cpDNA (Warwick and Black 1991). Based on ITS restriction sites polymorphism and the chloroplast region *trnL*, Warwick & Black (1991, 1993, 1994, 1997) and Warwick & Sauder (2005) identified seven major groups that were moderately supported: the "Oleracea", "Nigra", "Cakile", "Crambe", "Savignya", "Zilla", and "Vella" lineages. The "Zilla", "Savignya" and "Vella" clades seemed to be basal in the phylogenies but with no strong support. Later, by using four rapidly evolving non-coding chloroplast regions, Arias & Pires (2012) provided a fully resolved phylogeny of the tribe and suggested, in addition to the seven aforementioned clades, a new sub-tribal classification through the identification of a new African clade "Henophyton" that had not been previously sampled. In that study, the Savignya lineage appears curiously as a sister group to the Oleracea clade, but this was not confirmed by more recent studies using a large number of species but few genes (Hall et al. 2011; Willis et al. 2014).

In the present study, the Brassiceae representatives appear as a monophyletic group (BP=100, PP=1) in both nuclear and chloroplast gene phylogenies, which confirms previous results (Warwick and Black 1997; Hall et al. 2011; Arias and Pires 2012; Willis et al. 2014).

Bootstrap values and posterior probabilities both strongly support the relationships among the major clades of the Brassiceae tribe (Fig. 2, Fig. 3, Fig. S2). The first diverging lineages of the tribe are the clades “Vella”, “Zilla” and “Savignya”. In our chloroplast phylogeny, they form a monophyletic group, which is congruent with a previous work based on morphological data (Warwick and Black 1994), with “Savignya” and “Zilla” being monophyletic and sister group to “Vella”. On the nuclear topology, for the first time the “Vella” clade (*Carrichtera annua*) appears as the first diverging clade followed by the “Zilla” then the “Savignya” clades (Fig. 3a). The basal position of the three clades is congruent with some previous works based on chloroplast markers and ITS (Warwick and Black 1994; Crespo et al. 2000; Warwick and Sauder 2005; Hall et al. 2011; Willis et al. 2014). The conflicting nlrDNA and cpDNA results concerning these three clades might be explained for instance by an introgression of the “Vella” chloroplast genome in the “Zilla” and “Savignya” clades shortly after their divergence. However, a strong scenario about these events cannot be proposed as the phylogenetic relationships among the “Vella”, “Zilla” and “Savignya” clades are not recovered with certainty with our available chloroplast data. The basal phylogenetic position of *Carrichtera annua* confirms all previous analyses based on morphological and molecular data suggesting that this species is closely related to *Vella* L. and is a member of the Vella clade (Crespo et al. 2000; Warwick and Sauder 2005). Our results support that Savignya belongs to the earliest-divergent lineages rather than to the core Brassiceae, in contradiction with the results of Arias & Pires (2012) suggesting that *C. annua* belongs to the Oleracea clade.

Whereas Arias and Pires (2012) had defined the core Brassiceae as a group containing the two sub-clades Nigra+Cakile+Crambe and Savignya+Oleracea, our phylogenetic reconstructions suggest that the core Brassiceae represents a strongly supported monophyletic group (BP=100, PP=1.0 both for nuclear and chloroplast phylogenies) including the “Nigra”, “Cakile” and “Oleracea” clades (Fig. 2), with “Oleracea” and “Nigra” being sister group (BP=100, PP=1.0 and BP=92, PP=1.0 for the nuclear and chloroplast phylogenies respectively) (Fig. 2, Fig. 3a) in agreement with the results of Hall et *al.* (2011) and Willis et *al.* (2014). The phylogenetic relations among these three clades are generally poorly supported (Warwick and Black 1997; Warwick and Sauder 2005) and various results can be found in the literature (Hall et al. 2011; Arias and Pires 2012; Willis et al. 2014). In our species chloroplast phylogeny, *Raphanus sativus* appears to be more closely related to the Oleracea lineage than to the Nigra lineage, as previously reported (Palmer and Herbon 1988;

Warwick and Black 1991; Pradhan et al. 1992; Warwick and Black 1997; Yang et al. 2002), whereas it is more closely related to the Nigra lineage than to the Oleracea lineage in our phylogeny based on nuclear genes, as reported by Yang et al. (2002) based on nuclear gene data. The conflicting results between nuclear and chloroplast DNA sequences support the previously formulated hypothesis that Raphanus was derived from homoploid hybridization between the Oleracea and the Nigra clades (Yang et al. 2002). This hypothesis is corroborated by our findings that Nigra and Oleracea appear to be two closely related lineages. We suggest here the inclusion of *Raphanus sativus* in the Nigra lineage rather than in the Oleracea lineage. As it was previously shown (Hall et al. 2011; Willis et al. 2014), the phylogeny of the tribe Brassiceae based on nuclear genes is markedly different from the phylogeny based on chloroplast genes and both are necessary to understand the evolution of the group.

Two possible sister groups to the Brassiceae tribe have been proposed in the literature. The first consists in the genus Orychophragmus which comprises at least six species endemic to China and Korea (Al-Shehbaz and Guang 2000; Hu et al. 2015; Hu et al. 2016). Chromosome painting approaches have shown that this genus has experienced an historical WGD event (Lysak et al. 2005; Lysak et al. 2007). It is generally considered as a closely related genus to the tribe Brassiceae (Al-shehbaz 1985; Arias and Pires 2012) and even sometimes included within the Brassiceae (Warwick and Sauder 2005). The second candidate as a sister group to Brassiceae is the genus Sisymbrium, which did not experience a WGD event (Lysak et al. 2005), and has been reported as more closely related to the Brassiceae than to the Calepina/Coringia lineage (Arias et al. 2014). In our results, the relationships between *O. violaceus*, *S. irio* and the Brassiceae tribe depend on the WGT parental lineage considered. The species tree inferred from the concatenation of all genes coming from the three subgenomes (ALL-matrix) shows a good signal for a grouping of the Brassiceae tribe with *O. violaceus* (BP=88, PP=1, Fig. 3a), but this may be an artefact. Indeed, the ALL-matrix dataset is a concatenation of sequences from the three different genomes, which may have followed three distinct evolutionary histories. This is highlighted in the ALL-filtered species tree, where *S. irio* is more closely related to the Brassiceae tribe than *O. violaceus*, because all removed missing sites belonged essentially to MF1 and MF2 genes rather than to the LF genome (see Table 4, 40 to 42% of recovered genes within *C. annua*, *P. stylosa*, *S. purpurea* and *Z. spinosa subsp macroptera* belonged to the LF subgenome). The evolutionary history of the LF lineage is thus over-represented compared to those of the two other subgenomes. Consequently, it appears that concatenating genes from the three subgenomes should not

impact the phylogenetic relationships among the Brassiceae members, but may mislead the phylogenetic relationships between the tribe and outgroup species (Brassicaceae). This demonstrates that the phylogenetic inference from each subgenome separately (here, LF, MF1 and MF2 matrices) appears to be a necessary condition to successfully assess the evolutionary history of a mesopolyploid group. In any case, it seems that the phylogenetic relationship among Orychophragmus, Sisymbrium and the parental lineages of the Brassiceae tribe are quite difficult to assess. Given the branch lengths in the nuclear tree (Fig.4), speciation events between these five lineages happened rather rapidly and incomplete lineage sorting events may render difficult to recover the true evolutionary history of these clades. Nevertheless, our results suggest that one of the parental lineage, LF, may have been more closely related to the Sisymbrium and Orychophragmus clades, than to the other two parental lineages (MF1 and MF2), suggesting that the clade of the parental lineage LF did not go extinct as previously assumed.

With the chloroplast genomic data, we aimed at providing the first plastome phylogeny for the Brassiceae tribe, assessing potential hybridization/introgression events (see above) but as well determining which of the maternal parental genome got transmitted during the WGT events. However, due to a lack of phylogenetic resolution between the Brassiceae and their sister groups (BP = 74, PP = 0.9) in the chloroplast tree, no firm conclusion can be drawn. Moreover the obtained topology does not concord with any of the LF, MF1 nor MF2 nuclear topologies (Fig. 2, Fig. 3b, Fig. 4).

For the chloroplast gene analysis, the topology obtained from the “filtered” matrix is identical to the “non-filtered” topologies with some differences concerning branch supports. (Supplementary Fig. S1). For nuclear markers, only the LF topologies were identical regardless the matrix used “filtered” or “non-filtered”, whereas for the ALL, MF1 and MF2 topologies slight differences could be observed (Supplementary Fig. S3). They mainly concern the phylogenetic position of the sister genera Sisymbrium and Orychophragmus. For these present analyses missing data had a moderate impact on the inference of the topologies. Given the high number of missing data in most of the investigated Brassiceae species, we could not exclude some lineage-specific gene losses following the WGT event in the Brassiceae tribe. This observation confirms the need to assigning homoeologs before inferring the phylogenetic relationships of the tribe using nuclear genes, because differential gene losses might give rise to orthology and paralogy conflation.

In our study, developing a new methodological framework dedicated to the determination of mesopolyploid clades phylogeny, we provide the first fully resolved nuclear phylogeny of the main clades of the Brassiceae tribe. Our methodology allows assigning the orthologous genes to their original genomic origins and appears to be a well suitable tool for investigating the evolutionary history of mesopolyploid groups, even with incomplete data like transcriptome assemblies. However, this method requires a genomic reference with well assigned parental subgenomes, which is getting more and more accessible, with the rapid increase of the number of genome sequenced, and with the availability of powerful tools to reconstruct genome rearrangements following ancient allopolyploid events.

Supplementary Figures.

Fig. S1. Maximum likelihood chloroplast phylogeny of the tribe Brassiceae and its relatives with alignment gaps removed (see main text). The six investigated clades are indicated on the right as well as the genus Orychophragmus. Numbers at branches represent bootstrap support values and Bayesian posterior probabilities (BP/PP). (*) indicates nodes with maximal support (BP = 100 / PP = 1.0).

Fig. S2. Maximum likelihood nuclear phylogeny of the tribe Brassiceae and its relatives from the analysis of LF (left), MF1 (middle) and MF2 (right) sub-genomes concatenated genes (see main text). Numbers at branches represent bootstrap support values and Bayesian posterior probabilities (BP/PP). (*) indicates nodes with maximal support (BP = 100 / PP = 1.0).

Fig. S3. Maximum likelihood or Bayesian nuclear phylogenies of the tribe Brassiceae and its relatives from the analysis of ALL, LF, MF1 and MF2 sub-genomes concatenated genes with alignment gaps removed ("filtered" matrices, see main text). ML phylogeny is represented for the ALL, LF and MF2 matrices, whereas Bayesian phylogeny is represented for the MF1 matrix. Numbers at branches represent bootstrap support values and Bayesian posterior probabilities (BP/PP). (*) indicates nodes with maximal support (BP = 100 / PP = 1.0).

Fig. S4. Maximum likelihood nuclear phylogeny of the tribe Brassiceae from the analysis of the HOMOLOGOUS-matrix constructed from a subset of homoeologous gene triplets presented with same annotation in Liu et al. (2014) and in Murat et al. (2016) (see main text). The tree topology displays three sub-trees corresponding to the three sub-genomes present in the Brassiceae species (LF, MF1, MF2). Numbers at branches represent bootstrap support values for the ML analysis (BP). (*) indicates nodes with maximal support (BP = 100). Bni: *Brassica nigra*, Rsa: *Raphanus sativus*, Rra: *R. raphanistrum*, Bol: *B. oleracea*, Bra: *B. rapa*, Cma: *Cakile maritima,* Pst: *Psychine stylosa,* Spu: *Schouwia purpurea*, Zsp: *Zilla spinosa subsp macroptera*.

Supplementary Tables.

Table S1. List of species used for RNA sequencing using Illumina Hiseq technology and information on sequencing. The number of paired-end reads was assessed using FastQC. The range for each species (min - max) is reported. *Crambe maritima* was not used in this study but will be used in future analyses.

Table S2. Summary statistics for transcriptome assemblies (after CAP3). All statistics were assessed using QUAST and are based on TRINITY contigs of size ≥ 500bp, unless otherwise noted. The range for each species (min - max) is reported. *Crambe maritima* was not used in this study but will be used in future analyses.

Table S3. Mapping statistics of raw reads of *S. irio* mapped onto each coding DNA sequence of the chloroplast genome of *B. nigra* (Seol et al. 2015) using SAMtools (Li and Durbin 2009) and YASS results for each contig of the Spades assemblies. Dash indicates that no sequence has been constructed from mapped reads.

Supplementary Figures.

Fig. S1.

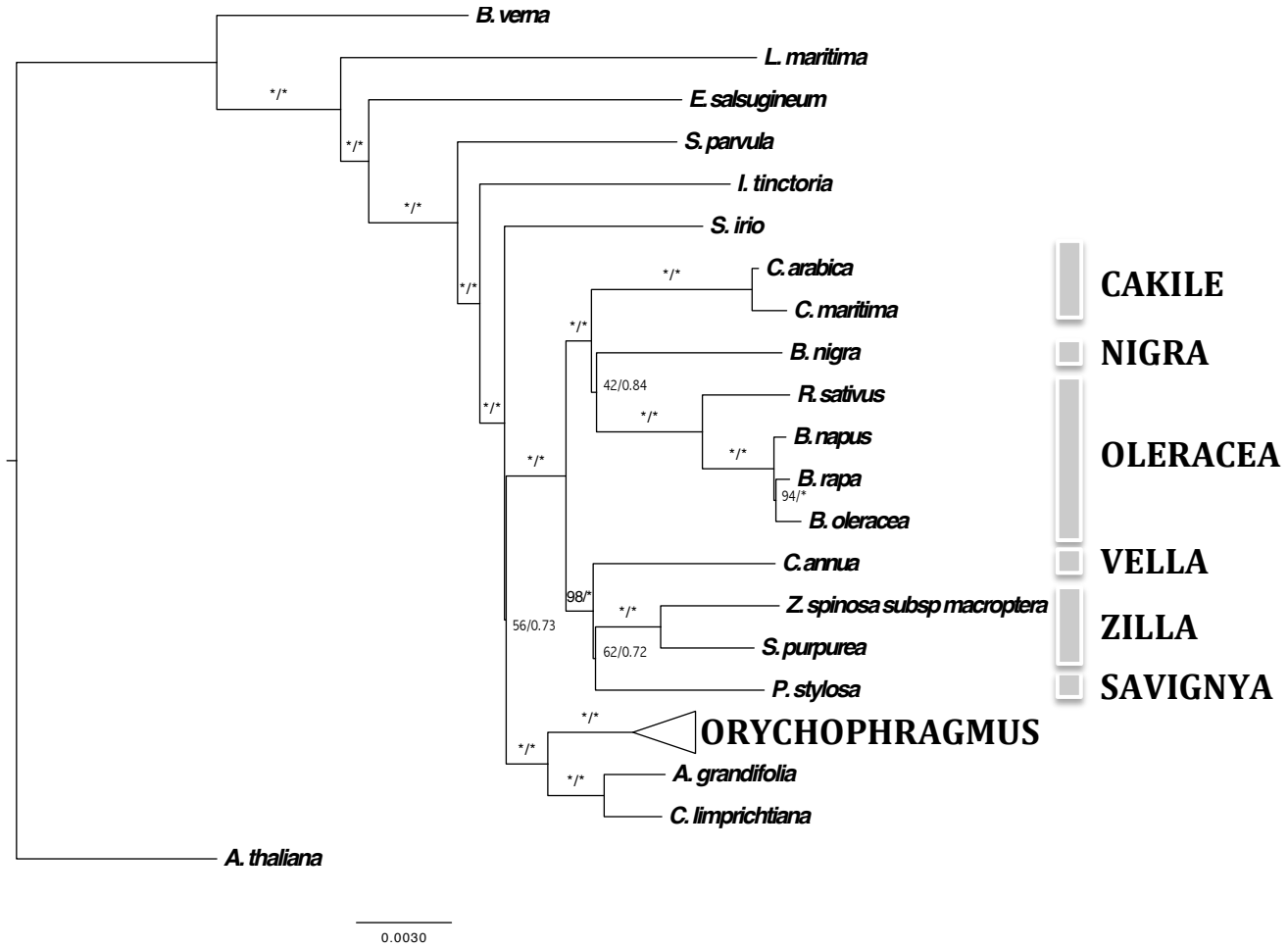
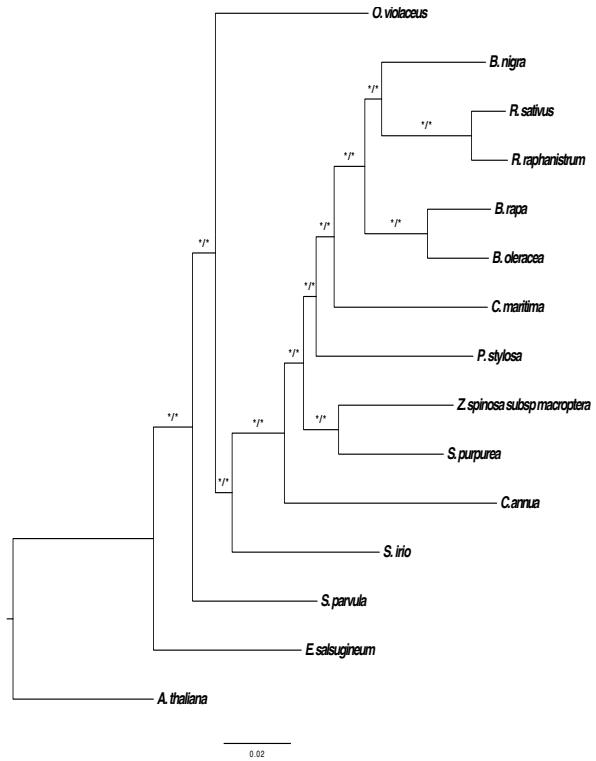
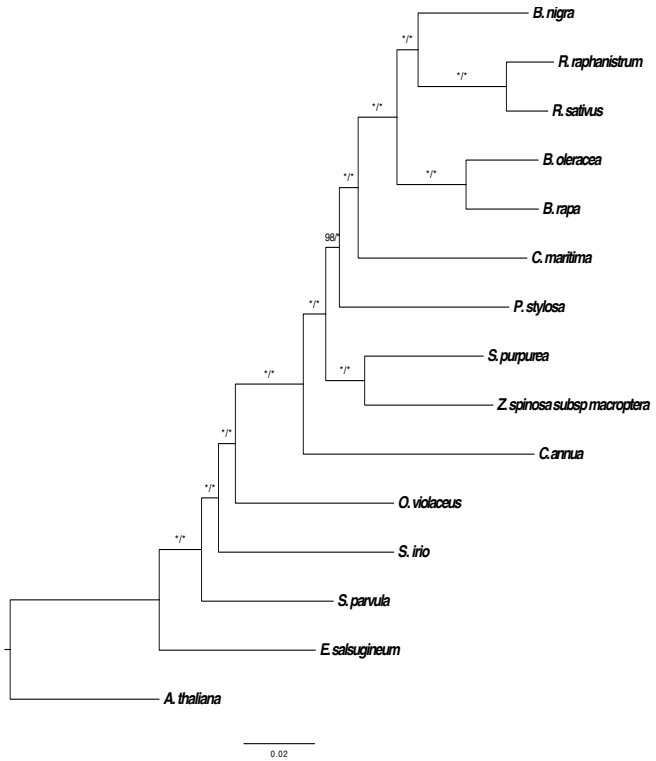


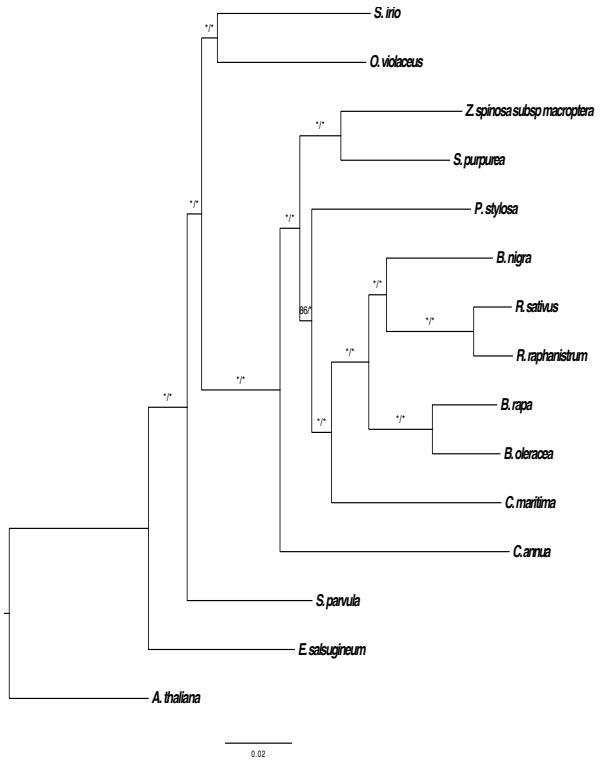
Fig. S2.



(LF)

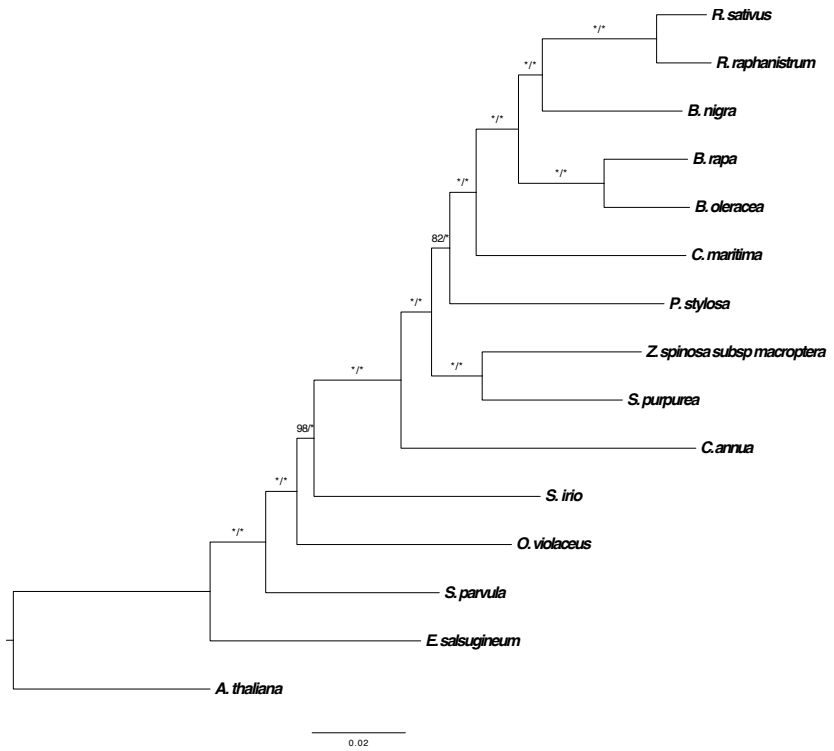


(MF1)

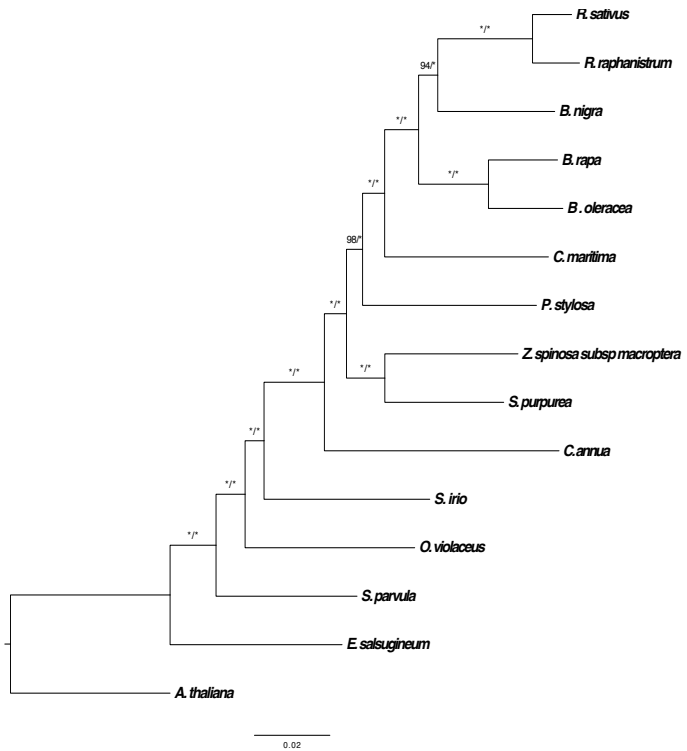


(MF2)

Fig. S3.

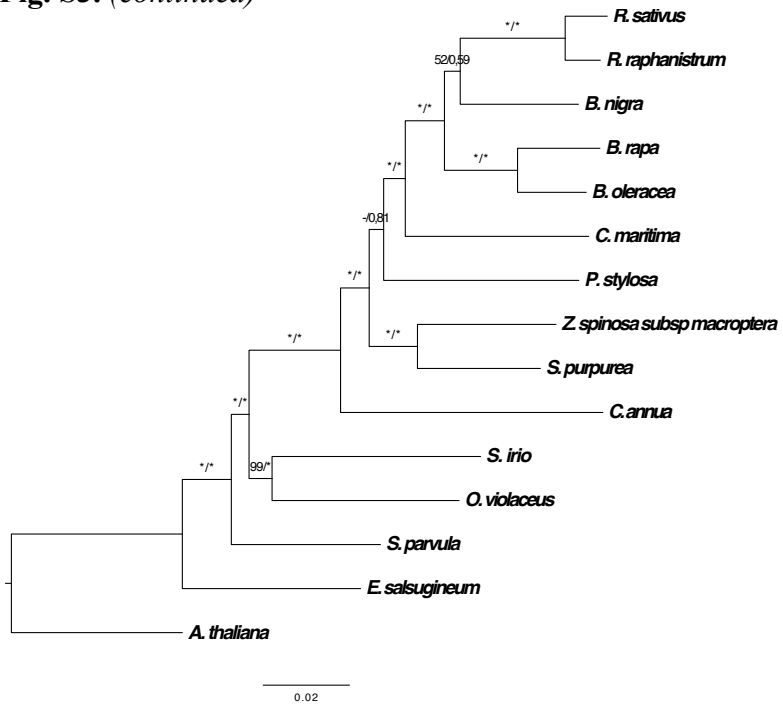


(ALL-filtered)

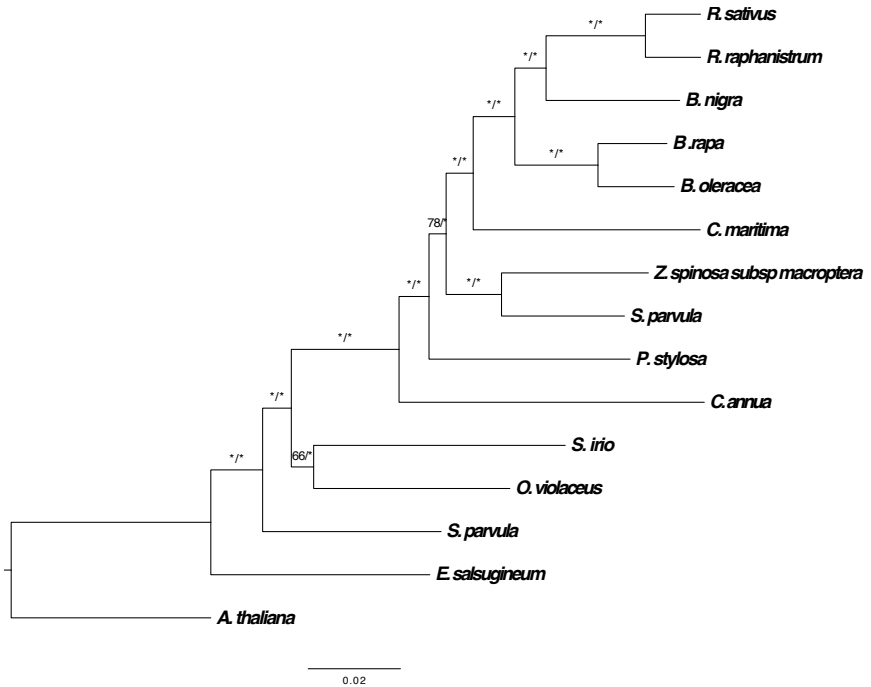


(LF-filtered)

Fig. S3. *(continued)*

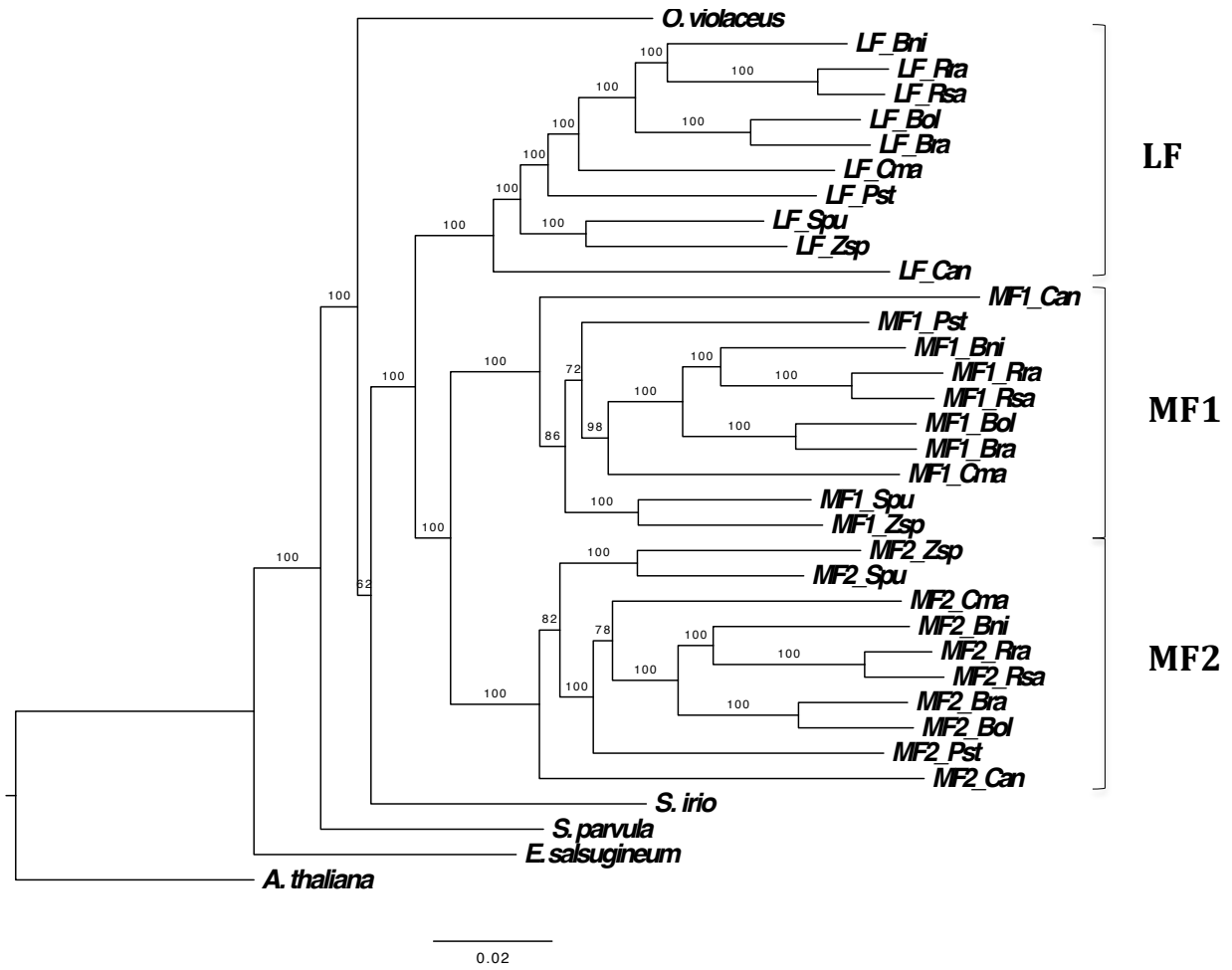


(MF1-filtered)



(MF2-filtered)

Fig. S4.


O. violaceus
LF_Bni
LF_Rra
LF_Rsa
LF_Bol
LF_Bra
LF_Cma
LF_Pst
LF_Spu
LF_Zsp
LF_Can
MF1_Can
MF1_Pst
MF1_Bni
MF1_Rra
MF1_Rsa
MF1_Bol
MF1_Bra
MF1_Cma
MF1_Spu
MF1_Zsp
MF2_Zsp
MF2_Spu
MF2_Cma
MF2_Bni
MF2_Rra
MF2_Rsa
MF2_Bra
MF2_Bol
MF2_Pst
MF2_Can
S. irio
S. parvula
E. salsugineum
A. thaliana
LF
MF1
MF2
0.02

Supplementary Tables.

Table S1.

Source of seeds[a]	Species	Organs (Number of individuals)	RNAseq technology	Read size	Number of raw reads	Number of clean reads
BGBe	*Orychophragmus violaceus*	Flower buds (2)	Illumina HiSeq 2000, paired-end	100bp	14,151,140 - 14,976,841	12,920,785 - 13,756,251
LBG BCN 3548	*Carrichtera annua*	Flower buds (3)	Illumina HiSeq 3000, paired-end	150bp	19,186,051 - 34,288,654	17,927,018 - 32,416,939
GCC BCN 8087	*Schouwia purpurea*	Flower buds (3)	Illumina HiSeq 3000, paired-end	150bp	36,054,912 - 39,167,535	34,141,935 - 36,712,467
GCC BCN 8055	*Zilla spinosa subsp macroptera*	Leaves (2)	Illumina HiSeq 3000, paired-end	150bp	7,070,061 - 8,074,587	6,648,568 - 7,598,227
BGBe BCN 3515	*Psychine stylosa*	Flower buds (2)	Illumina HiSeq 3000, paired-end	150bp	28,874,630 - 39,162,643	27,076,377 - 36,886,118
Embouchure de la Slack, Ambleteuse (Pas-de-Calais)	*Crambe maritima*	Flower buds (3)	Illumina HiSeq 3000, paired-end	150bp	22,250,949 - 30,817,974	20,746,799 - 28,941,942
Digue du Braek, Grande-Synthe (Nord)	*Cakile maritima*	Flower buds (12)	Illumina HiSeq 2500, paired-end	100bp	10,829,811 - 21,637,768	10,207,438 - 20,097,777

[a]BGBe, National Botanic Garden of Belgium, Meise, Belgium ; GCC, Gómez-Campo Collection, Instituto Nacional de Investigaciones Agrarias, Madrid, Spain ; LBG, Leipzig Botanical Garden, Leipzig, Germany. BCN indicates the collection number on herbarium specimens deposited at Herbarium, Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada (DAO).

Table S2.

Species	Total number of TRINITY contigs (≥0bp)	Total number of TRINITY contigs	Total length (bp)	Largest contig (bp)	N50 (bp)	GC (%)
Orychophragmus violaceus	65,359 - 78,128	39,529 - 48,912	61,313,006 – 75,284,383	11,290 - 14,979	1,821 - 1,857	42.56 - 42.65
Carrichtera annua	70,141 - 77,478	40,135 - 43,471	57,324,639 - 63,184,578	14,393 - 15,537	1,699 - 1,757	42.11 - 42.27
Schouwia purpurea	89,921 - 115,887	48,474 - 57,725	70,679,064 - 79,652,025	15,483 - 15,546	1,658 - 1,754	42.57 - 42.85
Zilla spinosa subsp macroptera	52,564 - 56,589	28,710 - 30,767	38,803,080 - 41,875,300	13,332 - 14,382	1,596 - 1,606	43.29 - 43.47
Psychine stylosa	62,623 - 67,639	34,306 - 37,900	49,067,347 - 55,056,638	15,596 - 15,778	1,697 - 1,723	43.14 - 43.40
Crambe maritima	85,592 - 114,672	46,789 - 59,193	64,728,896 - 78,745,514	14,376 - 15,567	1,576 - 1,646	42.88 - 43.51
Cakile maritima	55,157 - 68,295	32,800 - 39,718	46,528,647 - 59,464,324	11,901 - 16,144	1,665 - 1,820	43.33 - 43.84

Table S3. Mapping statistics of raw reads of *S. irio* mapped onto each coding DNA sequence of the chloroplast genome of *B. nigra* (Ref.) using SAMtools (Li and Durbin 2009) and YASS results for each contig of the Spades assemblies. Dash indicates that no sequence has been constructed from mapped reads.

Ref.	No. of raw mapped reads	No. of mapped bases (A)	No. of mismatches (B)	Error rate (B/A)	Length of the contig (bp)	Alignment length (bp)	Identity with Ref. (%)
accD	39552	4271616	115705	2.708694e-02	1483	1470	97.69
atpA	46336	5004288	129859	2.594955e-02	694	694	98.85
atpB	40895	4416660	102041	2.310366e-02	1520	1497	98.53
atpE	9119	984852	20952	2.127426e-02	421	399	98.75
atpF	11724	1266192	50191	3.963933e-02	451	441	98.54
atpH	4826	521208	22584	4.333011e-02	278	246	99.19
atpI	19490	2104920	49290	2.341657e-02	776	750	98.40
ccsA	25801	2786508	82966	2.977418e-02	1023	987	97.06
cemA	18137	1958796	36905	1.884066e-02	719	690	98.99
clpP	11028	1191024	38108	3.199600e-02	337	301	98.01
matK	41062	4434696	147215	3.319619e-02	1606	1575	97.33
ndhA	35263	3808404	98015	2.573650e-02	590	552	98.55
ndhB	111607	12053556	146373	1.214355e-02	631	621	97.10
ndhC	8164	881712	26225	2.974327e-02	389	363	98.62
ndhD	48740	5263920	144438	2.743925e-02	1523	1503	97.87
ndhE	8304	896832	23701	2.642747e-02	337	306	99.35
ndhF	78451	8472708	218559	2.579565e-02	2265	2241	97.81
ndhG	16139	1743012	52522	3.013290e-02	546	531	97.36
ndhH	40486	4372488	106702	2.440304e-02	1213	1182	98.56
ndhI	16890	1824120	26221	1.437460e-02	533	504	99.40
ndhJ	11333	1223964	17335	1.416300e-02	473	466	99.79
ndhK	17673	1908684	46611	2.442049e-02	705	678	98.53
petA	23446	2532168	38558	1.522727e-02	992	963	99.69
petB	16816	1816128	32692	1.800093e-02	669	639	98.90
petD	10415	1124820	34974	3.109298e-02	511	478	98.33
petG	1084	117072	10426	8.905631e-02	-	-	-
petL	218	23544	3419	1.452175e-01	-	-	-
petN	2	216	38	1.759259e-01	-	-	-
psaA	70840	7650720	136823	1.788368e-02	2282	2253	99.07
psaB	71301	7700508	128332	1.666539e-02	2235	1527	98.89
psaC	44029	4755132	103719	2.181201e-02	271	246	98.78
psaI	549	59292	4779	8.060110e-02	142	114	99.12
psaJ	1058	114264	7441	6.512112e-02	-	-	-
psbA	32621	3523068	78245	2.220934e-02	1087	1062	98.78
psbB	41321	4462668	109616	2.456288e-02	1558	1527	98.89
psbC	44029	4755132	103719	2.181201e-02	1453	1422	98.87
psbD	34771	3755268	72248	1.923911e-02	1097	1062	99.15
psbE	5649	610092	18475	3.028232e-02	282	252	99.60
psbF	1338	144504	9959	6.891850e-02	-	-	-
psbH	4342	468936	15261	3.254389e-02	243	222	98.20
psbI	599	64692	5452	8.427627e-02	-	-	-

psbJ	1075	116100	7618	6.561585e-02	-	-	-
psbK	2914	314712	12239	3.888952e-02	208	186	97.85
psbL	1258	135864	12329	9.074516e-02	-	-	-
psbM	980	105840	11408	1.077853e-01	-	-	-
psbN	1620	174960	10169	5.812186e-02	-	-	-
psbT	245	26460	3231	1.221088e-01	-	-	-
psbZ	3728	402624	11288	2.803608e-02	221	189	100
rbcL	44012	4753296	111441	2.344500e-02	1462	1429	98.67
rpl2	41981	4533948	85968	1.896096e-02	468	435	99.77
rpl14	8049	869292	28059	3.227799e-02	392	369	97.83
rpl16	9995	1079460	31304	2.899968e-02	299	278	98.92
rpl20	8355	902340	20411	2.262008e-02	386	354	99.15
rpl22	10918	1179144	39580	3.356672e-02	501	483	97.52
rpl23	13820	1492560	28728	1.924747e-02	311	282	99.55
rpl32	2055	221940	10271	4.627827e-02	183	159	98.11
rpl33	2796	301968	12767	4.227931e-02	229	201	98.51
rpl36	1238	133704	12300	9.199426e-02	-	-	-
rpoA	27306	2949048	73579	2.495009e-02	1016	983	98.47
rpoB	115114	12432312	239980	1.930293e-02	1037	1026	99.12
rpoC1	62358	6734664	138426	2.055426e-02	1649	1615	98.76
rpoC2	132420	14301360	335913	2.348819e-02	4139	4122	98.30
rps2	19884	2148552	39307	1.829465e-02	743	711	99.84
rps3	17234	1861272	40943	2.199732e-02	687	657	99.09
rps4	14982	1618056	35030	2.164944e-02	632	606	98.68
rps7	32491	3509028	46514	1.325552e-02	544	468	100
rps8	8027	866916	26013	3.000637e-02	427	405	98.27
rps11	9693	1057644	32900	3.110688e-02	437	417	98.08
rps12	15427	1666116	41018	2.461893e-02	272	231	100
rps14	7302	788616	20920	2.652749e-02	333	303	99.34
rps15	6151	664308	17581	2.646513e-02	293	267	98.50
rps16	3982	430056	12781	2.971939e-02	253	231	96.97
rps18	5746	620568	13045	2.102106e-02	335	306	99.35
rps19	6266	676728	36281	5.361238e-02	303	279	97.13
ycf1	194002	20952216	794982	3.794262e-02	2467	2478	93.95
ycf2	479813	51819804	667560	1.288233e-02	1261	1261	99.44
ycf3	8288	895104	31369	3.504509e-02	272	230	99.13
ycf4	13228	1428624	34934	2.445290e-02	587	555	99.10
ycf15	12621	1363068	36813	2.700746e-02	254	234	98.29

References

Albrecht B, Scornavacca C, Cenci A, Huson DH. 2012. Fast computation of minimum hybridization networks. Bioinformatics 28:191–197.

Al-Shehbaz IA, Guang Y. 2000. A revision of the chinese endemic Orychophragmus (Brassicaceae). Novon 10:349–353.

Al-Shehbaz IA. 1985. The genera of Brassiceae (Cruciferae; Brassicaceae) in the southeastern united states. J. Arnold Arbor. 66:279–301.

Al-Shehbaz IA. 2012. A generic and tribal synopsis of the Brassicaceae (Cruciferae). Taxon 61:931–954.

Altschul SF, Gish W, Miller W, Myers EW, Lipmanl DJ. 1990. Basic Local Alignment Search Tool. J. Mol. Evol. 215:403–410.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arias T, Beilstein M a., Tang M, McKain MR, Pires JC. 2014. Diversification times among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. Am. J. Bot. 101:86–91.

Arias T, Pires JC. 2012. A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae : Brassiceae): Novel clades and potential taxonomic implications. Taxon 61:980–988.

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. Curr. Opin. Plant Biol. 15:140–146.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham SON, Prjibelski AD, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455–477.

Barker MS, Vogel H, Schranz ME. 2009. Paleopolyploidy in the Brassicales : analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. Genome Biol. Evol. 1:391–399.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. 13:137–144.

Bortiri E, Coleman-derr D, Lazo GR, Anderson OD, Gu YQ. 2008. The complete chloroplast genome sequence of Brachypodium distachyon : sequence comparison and phylogenetic analysis of eight grass plastomes. BMC Res. Notes 1.

Capella-Gutiérrez S, Silla-martínez JM, Gabaldón T. 2009. trimAl : a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One 7:1–9.

Cheng F, Wu J, Wang X. 2014. Genome triplication drove the diversification of Brassica plants. Hortic. Res. 24:1–8.

Cheng F, Wu J, Xie Q, Lysak MA, Wang X. 2013. Deciphering the diploid ancestral genome of the mesohexaploid Brassica rapa. Plant Cell 25:1541–1554.

Couvreur TLP, Franzke A, Al-shehbaz IA, Bakker FT, Koch A, Mummenhoff K. 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol. Biol. Evol. 27:55–71.

Crespo M, Lledó DM, Fay MF, Chase MW. 2000. Subtribe Vellinae (Brassiceae, Brassicaceae): a combined analysis of ITS nrDNA sequences and morphological data. Ann. Bot. 86:53–62.

Dassanayake M, Oh D, Haas JS, Hernandez A, Hong H, Ali S, Yun D, Bressan RA, Zhu J, Bohnert HJ, et al. 2011. The genome of the extremophile crucifer Thellungiella parvula. Nat. Genet. 43:913–918.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary genetics of genome merger and doubling in plants. Annu. Rev. Genet. 42:443–461.

Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-mack J, dePamphilis CW. 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10:1–18.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proc. Natl. Acad. Sci. U. S. A. 112:8362–8366.

Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K. 2011. Cabbage family affairs: the evolutionary history of Brassicaceae. Trends Plant Sci. 16:108–116.

Gabaldón T, Koonin E V. 2013. Functional and evolutionary implications of gene orthology. Nat. Rev. Genet. 14:360–366.

Geiser C, Mandáková T, Arrigo N, Lysak MA, Parisod C. 2016. Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in buckler mustard. Plant Cell 28:17–27.

Glémin S, Scornavacca C, Dainat J, Burgarella C, Viader V, Ardisson M, Sarah G, Santoni S, David J, Ranwez V. 2018. Pervasive hybridizations in the history of wheat relatives (submitted).

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Chen Z, et al. 2013. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 29:644–652.

Guo X, Liu J, Hao G, Zhang L, Mao K, Wang X, Zhang D, Ma T, Hu Q, Al-shehbaz IA, et al. 2017. Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics 18:1–9.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST : quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075.

Hall JC, Tisdale TE, Donohue K, Wheeler A, Al-yahya MA, Kramer EM. 2011. Convergent evolution of a complex fruit structure in the tribe Brassiceae (Brassicaceae). Am. J. Bot. 98:1989–2003.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90, 000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. 45:891–898.

Hohmann N, Wolf EM, Lysak MA, Koch MA. 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. Plant Cell 27:2770–2784.

Hu H, Al-shehbaz IA, Sun Y, Hao G, Wang Q, Liu J. 2015. Species delimitation in Orychophragmus (Brassicaceae) based on chloroplast and nuclear DNA barcodes. Taxon 64:714–726.

Hu H, Hu Q, Al-shehbaz IA, Luo X, Zeng T, Guo X. 2016. Species delimitation and interspecific relationships of the genus Orychophragmus (Brassicaceae) inferred from whole chloroplast genomes. Front. Plant Sci. 7:1–10.

Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-shehbaz I, Edger PP, et al. 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol. Biol. Evol. 33:394–412.

Huang X, Madan A. 1999. CAP3 : a DNA sequence assembly program. Genome Res. 9:868–877.

Huber KT, Moulton V. 2006. Phylogenetic networks from multi-labelled trees. Math. Biol. 52:613–632.

Huber KT, Oxelman B, Lott M, Moulton V. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. Mol. Biol. Evol. 23:1784–1791.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3 : Reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33:1635–1638.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali S, Landherr L, Ralph PE, Jiao Y, Wickett NJ, Ayyampalayam S. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–102.

Jones G, Sagitov S, Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol. 62:467–478.

Kagale S, Robinson SJ, Nixon J, Xiao R, Huebert T, Condie J, Kessler D, Clarke WE, Edger PP, Links MG, et al. 2014. Polyploid Evolution of the Brassicaceae during the Cenozoic Era. Plant Cell 26:2777–2791.

Kitashiba H, Li F, Hirakawa HI, Kawanabe T, Zou Z, Hasegawa YO, Tonosaki KA, Shirasawa SA, Fukushima AKI, Yokoi SH, et al. 2014. Draft sequences of the radish (Raphanus sativus L.) genome. DNA Res. 21:481–490.

Lagercrantz U. 1998. Comparative mapping between Arabidopsis thaliana and Brassica nigra indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics 150:1217–1228.

Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, Soltis PS. 2018. Impact of whole- genome duplication events on diversification rates in angiosperms. Am. J. Bot. 105:1–16.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2 : New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772–773.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–360.

Lemmon AR, Brown JM, Stanger-Hall K, Moriarty Lemmon E. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Syst. Biol. 58:130–145.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows – Wheeler transform. Bioinformatics 25:1754–1760.

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5:1–11.

Lott M, Spillner A, Huber KT, Moulton V. 2009. PADRE : a package for analyzing and displaying reticulate evolution. Bioinformatics 25:1199–1200.

Lou P, Wu J, Cheng F, Cressman LG, Wang X, Mcclung CR. 2012. Preferential retention of circadian clock genes during diploidization following whole genome triplication in Brassica rapa. Plant Cell 24:2415–2426.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. Genome Res. 15:516–525.

Lysak MA, Cheung, K, Kitschke M, Bures P. 2007. Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. Plant Physiol. 145:402–410.

Mandáková T, Joly S, Krzywinski M, Mummenhoff K, Lysak M a. 2010. Fast diploidization in close mesopolyploid relatives of Arabidopsis. Plant Cell 22:2277–2290.

Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. 91:3–21.

Marhold K, Lihová J. 2006. Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Syst. Evol. 259:143–174.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S. 2014. Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish Raphanus raphanistrum and three other Brassicaceae species. Plant Cell 26:1925–1937.

Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, Crollius HR, Salse J. 2015. Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol. 16:1–17.

Van Der Niet T, Linder HP. 2008. Dealing with incongruence in the quest for the species tree : A case study from the orchid genus Satyrium. Mol. Phylogenet. Evol. 47:154–174.

Noe L, Kucherov G. 2005. YASS : enhancing the sensitivity of DNA similarity search. Nucleic Acids Res. 33:540–543.

Oberprieler C, Wagner F, Tomasello S, Konowalik K. 2017. A permutation approach for inferring species networks from gene trees in polyploid complexes by minimising deep coalescences. Methods Ecol. Evol. 8:835–849.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annu. Rev. Genet. 34:401–437.

Oxelman B, Brysting AK, Jones GR, Marcussen T, Oberprieler C, Pfeil BE. 2017. Phylogenetics of allopolyploids. Annu. Rev. Ecol. Evol. Syst. 48:543–557.

Palmer JD, Herbon LA. 1988. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. 28:87–97.

Palmer JD. 1985. Comparative organization of chloroplast genomes. Annu. Rev. Genet. 19:325–354.

Panchy N, Lehti-Shiu M, Shiu S. 2016. Evolution of Gene Duplication in Plants. Plant Physiol. 171:2294–2316.

Parkin I a P, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ. 2005. Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. Genetics 171:765–781.

Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nat. Rev. Genet. 18:411–424.

Popp M, Erixon P, Eggens F, Oxelman B. 2005. Origin and evolution of a circumpolar polyploid species complex in Silene (Caryophyllaceae) inferred from low copy nuclear RNA polymerase introns, rDNA, and chloroplast DNA. Syst. Biol. 30:302–313.

Pradhan AK, Prakash S, Mukhopadhyay A, Pental D. 1992. Phylogeny of Brassica and allied genera based on variation in chloroplast and mitochondrial DNA patterns : molecular and taxonomic classifications are incongruous. Theor. Appl. Genet. 85:331–340.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE : Multiple Alignment of Coding Sequences accounting for frameshifts and stop codons. PLoS One 6:e22594.

Renny-Byfield S, Gong L, Gallagher JP, Wendel JF. 2015. Persistence of subgenomes in paleopolyploid cotton after 60 My of evolution. Mol. Biol. Evol. 32:1063–1071.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3 : Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Ronquist FR, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2 : Efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Roure B, Baurain D, Philippe H. 2012. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

Sang T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 37:121–147.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. PNAS 108:4069–4074.

Schranz ME, Mitchell-olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell 18:1152–1165.

Scienski K, Fay JC, Conant GC. 2015. Patterns of gene conversion in duplicated yeast histones. Genome Biol. Evol. 7:3249–3258.

Smedmark JEE, Eriksson T, Evans RC, Campbell C. 2003. Ancient allopolyploid speciation in Geinae (Rosaceae): evidence from nuclear granule-bound starch synthase (GBSSI) gene sequences. Syst. Biol. 52:374–385.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of Arabidopsis thaliana development. Nat. Genet. 37:501–506.

Seol Y, Kim K, Kang S, Perumal S, Kim C. 2015. The complete chloroplast genome of two Brassica species, Brassica nigra and B. Oleracea. Mitochondrial DNA 28:167–168.

Soltis DE, Albert VA, Leebens-mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Claude W, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. Am. J. Bot. 96:336–348.

Soltis DE, Doyle JJ, Soltis PS. 1998. Molecular Systematics of Plants II. Boston: Kluwer Academic Publishers

Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, Mavrodiev EV, Mei W, Cortez MB, Soltis PS, Gitzendanner MA. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). New Phytol. 202:1105–1117.

Soltis DE, Visger CJ, Soltis PS. 2014. The polyploidy revolution then...and now: Stebbins revisited. Am. J. Bot. 101:1057–1078.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57:758–771.

Stamatakis A. 2006. RAxML-VI-HPC : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stamatakis A. 2014. RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC. 2012. Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics 190:1563–1574.

Vanderpoorten A, Shaw AJ, Cox CJ. 2004. Evolution of multiple paralogous adenosine kinase genes in the moss genus Hygroamblystegium : phylogenetic implications. Mol. Phylogenet. Evol. 31:505–516.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in Arabidopsis. Science 290:2114–2117.

Wang X, Paterson AH. 2011. Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization. Genes. 2:1–20.

Wang X, Tang H, Paterson AH. 2011b. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major poaceae lineages. Plant Cell 23:27–37.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011a. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43:1035-1040

Warwick SI, Al-Shehbaz IA, Price RA, Sauder C. 2002. Phylogeny of Sisymbrium (Brassicaceae) based on ITS sequences of nuclear ribosomal DNA. Can. J. Bot. 80:1002–1017.

Warwick SI, Black LD. 1991. Molecular systematics of Brassica and allied genera (Subtribe Brassicinae , Brassiceae) - chloroplast genome and cytodeme congruence. Theor. Appl. Genet. 82:81–92.

Warwick SI, Black LD. 1993. Molecular relationships in subtribe Brassicinae (Cruciferae, tribe Brassiceae). Can. J. Bot. 71:906–918.

Warwick SI, Black LD. 1994. Evaluation of the subtribes Moricandiinae, Savignyinae, Vellinae, and Zillinae (Brassicaceae, tribe Brassiceae) using chloroplast DNA restriction site variation. Can. J. Bot. 72:1692–1701.

Warwick SI, Black LD. 1997. Phylogenetic implications of chloroplast DNA restriction site variation in subtribes Raphaninae and Cakilinae (Brassicaceae, tribe Brassiceae). Can. J. Bot. 75:960–973.

Warwick SI, Mummenhoff K, Sauder C a., Koch M a., Al-Shehbaz I a. 2010. Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. Plant Syst. Evol. 285:209–232.

Warwick SI, Sauder CA. 2005. Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trn L intron sequences. Can. J. Bot. 83:467–483.

Willis CG, Hall JC, Casas RR De, Wang TY, Donohue K. 2014. Diversification and the evolution of dispersal ability in the tribe Brassiceae (Brassicaceae). Ann. Bot. 114:1675–1686.

Wolfe KH, Sharp PM, Li W. 1989. Rates of Synonymous Substitution in Plant Nuclear Genes. J. Mol. Evol. 29:208–211.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. PNAS 106:13875–13879.

Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, Hu Z, Chen S, Pental D, Ju Y, et al. 2016. The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. Nat. Genet. 48:1225–1234.

Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, Shu S, Prochnik S, Xin M, Ma C, et al. 2013. The reference genome of the halophytic plant Eutrema salsugineum. Front. Plant Sci. 4:1–14.

Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK, Carpenter EJ, Zhang Y, Chen L, Yan Z, Xie Y, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Mol. Biol. Evol. 32:2001–2014.

Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes : improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31:3081–3092.

Yang Y, Tai P, Chen Y, Li W. 2002. A study of the phylogeny of Brassica rapa, B. nigra, Raphanus sativus, and their related genera using noncoding regions of chloroplast DNA. Mol. Phylogenet. Evol. 23:268–275.

Zhang N, Zeng L, Shan H, Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195:923–937.

Ziolkowski P a, Kaczmarek M, Babula D, Sadowski J. 2006. Genome evolution in Arabidopsis/Brassica: conservation and divergence of ancient rearranged segments and their breakpoints. Plant J. 47:63–74.

CHAPITRE II

Triplication de génome entier (WGT) chez les Brassiceae

Bien que les études basées sur l'approche de « Comparison Chromosome Painting » (CCP) semblent indiquer clairement que l'ensemble des espèces de la tribu des Brassiceae a subi un événement de triplication de génome (Lysak et al. 2005; Lysak et al. 2007). Les approches de CCP ont bien révélé la présence de trois copies de chaque bloc génomique ancestral étudié chez les espèces de Brassiceae investiguées (24 au total), contre une seule copie chez l'espèce de référence *Arabidopsis thaliana* – ce qui indique clairement que les espèces de Brassiceae étudiées ont subi un événement de triplication de génome. Cependant, ces seules approches ne permettent pas totalement d'exclure la possibilité que plusieurs évènements indépendants de WGT aient eu lieu au cours de l'histoire évolutive des Brassiceae même si un unique événement de WGT ayant eu lieu chez l'ancêtre commun des Brassiceae reste l'hypothèse la plus parcimonieuse.

A partir d'assemblages de transcriptomes, des méthodes classiques basées sur les distributions des valeurs de taux de substitution synonyme (Ks) ont permis de mettre en évidence l'événement de WGT chez l'ensemble des espèces de Brassiceae investiguées. Ces approches nous ont permis de confirmer la présence d'un événement de WGD (ou WGT, en effet si ces évènements sont rapprochés dans le temps il n'est pas possible de les distinguer par cette approche) dans le passé des espèces investiguées, mais pas de conclure sur le partage entre espèces des évènements de WGD identifiés. C'est pourquoi nous avons complété nos recherches par l'utilisation de l'approche phylogénétique développée dans le chapitre précédent.

A ce jour, les études chez *Brassica rapa* révèlent que l'événement de triplication de génome se compose de deux évènements successifs de duplication de génome (Wang et al. 2011; Tang et al. 2012). Le premier évènement d'allopolyploïdie aurait formé un hybride tétraploïde (4n) issu d'un croisement entre deux individus appartenant à des espèces diploïdes différentes. Le deuxième évènement d'allopolyploïdie aurait eu lieu entre cet hybride allotétraploïde et une troisième espèce diploïde, ce qui aurait créé un nouvel hybride allohexaploïde (l'ancêtre commun des Brassiceae). Chez *B. rapa*, les trois sous-génomes associés aux espèces parentales sont identifiables en utilisant des approches de synténie (conservation de l'ordre des gènes entre deux espèces apparentées). En utilisant les gènes conservés sur chacun des trois sous-génomes de *B. rapa* et de *B. oleracea* (Liu et al. 2014), nous avons cherché à déterminer si l'ensemble des clades de Brassiceae partage le même événement de WGT, par l'utilisation de la méthode exposée au chapitre précédent.

Par l'utilisation de cette approche méthodologique, nous avons également pu explorer une hypothèse proposée par Lysack et al. (2005) concernant l'événement de mésopolyploïdie

de la lignée Orychophragmus. En effet, les auteurs de cette étude ont proposé que le genre *Orychophragmus,* proche phylogénétiquement de la tribu des Brassiceae, et parfois même à l'intérieur de la tribu des Brassiceae selon certaines études phylogénétiques, pourrait être un descendant direct de l'intermédiaire tétraploïde issu du premier événement de spéciation allopolyploïde menant à l'ancêtre hexaploïde des Brassiceae. Les données phylogénétiques produites dans ce second chapitre nous ont permis d'explorer cette hypothèse et de la réfuter.

Enfin, nous avons cherché à savoir si le phénomène de dominance entre les trois sous-génomes parentaux mis en évidence chez Brassica pouvait se généraliser à l'ensemble des Brassiceae et s'il était associé à un relâchement de la pression de sélection sur les gènes appartenant aux sous-génomes dominés ce à quoi l'on s'attend. L'ensemble de ces analyses, leurs résultats ainsi que l'interprétation de ces résultats font l'objet du second chapitre, présenté dans l'article suivant (en préparation pour publication).

Références

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5:1–11.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. Genome Res. 15:516–525.

Lysak MA, Cheung K, Kitschke M, Bures P. 2007. Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. Plant Physiol. 145:402–410.

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC. 2012. Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics 190:1563–1574.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43:1035–1040.

New insights on the mesohexaploid Brassiceae tribe (Brassicaceae): evidence for subgenome dominance.

Authors

Hénocq L.[1], Mazoyer C.[1], Gallina S.[1], Arrigo, N.[2,3], Parisod C.[4], Castric V.[1],Vekemans X.[1], Poux C.[1]

Affiliations

[1]Univ. Lille, CNRS, UMR 8198 – Evo-Eco-Paleo, Lille, France;
[2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA;
[3]Department of Ecology and Evolution, University of Lausanne, CH-1015, Lausanne, Switzerland;
[4]Laboratory of Evolutionary Botany, University of Neuchâtel, Neuchâtel, Switzerland

Abstract

Whole genome duplication events (WGD) are common in flowering plants and presumably played a major role in their diversification. Less studied are the whole genome triplication events (WGT) originating from two successive events of allopolyploidy. Some examples can be found in the Brassicaceae family: in Brassica, an approximately 15 million years old WGT was followed by genomic rearrangements and gene copy losses leading to genome diploidization, as it is widely observed after WGD. It has been observed as well that one of the three parental genomes shows higher levels of gene expression and has retained more genes than the other(s), a phenomenon described as "subgenome dominance". Asymmetry among subgenomes in gene expression and biased gene fractionation are linked because an overall lower level of gene expression within the non-dominant subgenome relaxes the strength of purifying selection and the sensitivity to dosage-constraints. In the present study, we focused on the Brassiceae tribe, to which belongs the genus Brassica, to test whether the patterns of asymmetrical molecular evolution among the subgenomes found in Brassica could be generalized to the Brassiceae tribe. We used RNAseq data from representatives of the main Brassiceae's lineages and used a phylogenomic framework to reconstruct the phylogeny of the tribe - in order to test whether all species share the same WGT and to investigate the phylogenetic relationships among the parental lineages. Based on the obtained phylogeny, we tested the incidence of subgenome dominance within Brassiceae, by estimating and comparing the Kn/Ks values corresponding to the three subgenomes. Altogether, our results confirm that all extant lineages of the Brassiceae tribe have experienced a single WGT and that the moderately fractionated subgenomes show patterns of molecular evolution indicating more relaxed purifying selection than the least fractionated one, demonstrating asymmetric evolution among subgenomes in relation to gene expression levels.

Key words: WGT, allopolyploidy, mesohexaploidy, subgenome dominance, biased genomic expression, biased fractionation

Introduction

Polyploidy is a widespread phenomenon shaping genome evolution in plants. Indeed, there was at least one whole genome duplication (WGD) event in the ancestry of all extant angiosperms and most plant clades are ancient polyploids (Jiao et al. 2011; Arrigo and Barker 2012; Panchy et al. 2016; Van de Peer et al. 2017). In flowering plants, nearly 35% of species are neopolyploids (*i.e.*, recently formed polyploids) and at least 15% of speciation events are accompanied by ploidy increase (Wood et al. 2009). Polyploids arise when rare mitotic or meiotic errors cause the formation of gametes with more than one set of chromosomes: there are called unreduced gametes (Comai 2005). Both autopolyploids and allopolyploids have multiple sets of chromosomes, but the former result from intraspecific genome doubling (*i.e.*, the different sets of chromosomes have the same intraspecific origin), whereas the latter result from interspecific hybridization with genome doubling (*i.e.*, the different sets of chromosomes have different origins). Among neopolyploid vascular plant species, it has recently been estimated that autopolyploids and allopolyploids have similar frequencies of occurrence (Barker et al. 2016), although allopolyploids have long been thought to be predominant. Speciation among diploids and subsequent hybridization can produce allotetraploids via the formation of unreduced gametes (Ramsey and Schemske 1998; Levin 2013; Douglas et al. 2015). However polyploidization events may follow each other increasing the ploidy level as exemplified by the allohexaploid bread wheat *Triticum aestivum* (Marcussen et al. 2014), the recently formed allohexaploid *Mimulus peregrinus* (Vallejo-Marín et al. 2015) and the well-documented brassica species (Wang et al. 2011a). There are two common ways to form allohexaploids: either through a cross between a diploid and a tetraploid individuals forming a triploid bridge (Ramsey and Schemske 1998) undergoing subsequently a genome doubling, or in a two-step process with first the production of an allo or autotetraploid species hybridizing, in a second step, with a distinct diploid species by the mean of unreduced gametes. After going through an allopolyploidization event, a genome thus contains homologous genes from different evolutionary origins, called homoeologs, whereas homologous genes that originated from a whole-genome duplication event in autopolyploids are called ohnologs (Glover et al. 2016).

After a polyploidy event, one generally observes a genome downsizing and an extensive diploidization involving genetic, epigenetic and structural changes leading eventually to a diploid genome (Mandáková et al. 2010). This mainly occurs via neo or sub-functionalization processes, pseudogenization processes (« gene silencing »), genomic

rearrangements (inversions, deletions and translocation) and gene losses (Lynch and Conery 2000; Wendel 2000; Marhold and Lihová 2006; Otto 2007; Tayalé and Parisod 2013). Gene loss and retention following an allopolyploidy event are shaped by the encoding function of genes (dosage-sensitive hypothesis, Birchler and Veitia 2007) but also by the relative "dominance" interactions between the recently merged parental genomes. Indeed, in Brassica, cotton, bread wheat and maize, it has been observed that one of the parental genomes involved in the mesopolyploidy events expresses and retains more genes than the other(s), which is due to dominant gene expression and biased-gene fractionation (*i.e.* bias in gene loss between subgenomes), respectively (Chang et al. 2010; Woodhouse et al. 2010; Schnable et al. 2011; Wang et al. 2011a; Cheng et al. 2012; Yoo et al. 2013; Leach et al. 2014; Renny-Byfield et al. 2015). Subgenome dominance in allopolyploids is thought to be associated with the methylation patterns of the parental subgenomes, particularly the methylation of adjacent transposable elements (Lee and Chen 2001; Chen 2007; Hollister and Gaut 2009; Woodhouse et al. 2014; Cheng et al. 2016; Edger et al. 2017). Asymmetry among subgenomes in gene expression and biased gene fractionation are linked because it is expected that the overall lower level of expression of genes within the non-dominant subgenome causes them to evolve under relaxed purifying selection and to become less sensitive to dosage-constraints (Schnable et al. 2011), which is associated with a higher accumulation of non-synonymous substitutions (Zhang et al. 2015).

In the Brassicaceae family, 37% of species are neopolyploid (Warwick and Al-Shehbaz 2006), which is close to the average figure for angiosperms (Wood et al. 2009). One ancient event of polyploidy has been well characterized in this plant family and is known to have occurred at the onset of the family diversification (Vision et al. 2000; Blanc et al. 2003; Schranz et al. 2006; Barker et al. 2009; Kagale et al. 2014; Edger et al. 2015). This polyploidy event postdates the divergence from the sister Cleomaceae family (Schranz and Mitchell-Olds 2006; Barker et al. 2009; Kagale et al. 2014). Genomes from Brassicaceae species also show evidence for two more ancient events (paleopolyploidy events): gamma and beta events, which probably took place during the flowering plant diversification and within the Brassicales order after the divergence from Caricaceae (*Carica papaya*), respectively (Bowers et al. 2003; Barker et al. 2009; Haudry et al. 2013; Kagale et al. 2014; Panchy et al. 2016). In addition to the three aforementioned paleopolyploidy events, many crucifer genera have experienced additional and more recent polyploidy events (Kagale et al. 2014; Mandáková et al. 2017). For example, all species of the tribe Brassiceae with diploid-like genomes analyzed to date contain three copies of genomic regions orthologous to single copy regions in the *A.*

thaliana genome. This result suggests that the common ancestor of the Brassiceae tribe has experienced two additional, more recent WGD events, generating a whole genome triplication (WGT) (Lysak et al. 2005; Parkin et al. 2005; Ziolkowski et al. 2006; Lysak et al. 2007; Wang et al. 2011a; Cheng et al. 2013; Liu et al. 2014; Moghe et al. 2014). These latter events are younger than the α, β and γ paleopolyploidy events described above, but older than neopolyploid speciation events. They are therefore referred to as mesopolyploidy events and can be detected by comparative genetic and cytogenetic methods (Lysak et al. 2005; Lysak et al. 2007; Mandáková et al. 2010), whereas neopolyploids contain clearly distinguishable parental subgenomes and higher chromosome numbers.

Following the well described mesohexaploidy event experienced by *Brassica rapa*, the observed subsequent patterns of fractionation among its three subgenomes have lead to suggest a two-step fractionation model, where the two subgenomes with heavy gene losses (i.e., the medium fractionated subgenome, MF1, and the most fractionated subgenome, MF2) were in the same nucleus for a longer period of time than the third subgenome (i.e., the least fractionated subgenome, LF) with the fewest gene losses (Wang et al. 2011a; Cheng et al. 2012; Tang et al. 2012; Cheng et al. 2013). Indeed, the LF subgenome has retained approximately twice as many genes as the other two subgenomes (Wang et al. 2011a; Tang et al. 2012; Cheng et al. 2014). More gene deletions were observed in the MF2 subgenome than in LF (10,423 vs. 4853, respectively), more gene singletons in the LF subgenome (5211) than in MF1 (2449), or MF2 (1592), and less nonsynonymous or frameshift mutations in genes located in the LF subgenome than those located in the MF subgenomes (Cheng et al. 2012; Tang et al. 2012). This differential gene losses between the three subgenomes demonstrated in *B. rapa* and its relative species *B. oleracea* may be explained by dominant gene expression between the three subgenomes as the expression levels of genes in the LF subgenome are significantly higher than corresponding syntenic genes in the other two subgenomes (Cheng et al. 2012; Liu et al. 2014). Woodhouse et al. (2014) suggested that the downregulation of gene expression in the MF1 and MF2 subgenomes were caused by the process of silencing of local transposable elements (TEs), as they observed that Brassica homoeologous regions experiencing greater gene loss were enriched in TEs, and surrounding genes had lower expression level compared with their homoeologous counterparts.

Given all the above-mentioned findings in *B. rapa* and *B. oleracea*, a scenario involving a two-step origin of Brassica species has been proposed (Cheng et al. 2012; Tang et al. 2012; Cheng et al. 2013; Cheng et al. 2014; Murat et al. 2015): (1) an 'allo'tetraploidization event involving hybridization between two diploid ancestral lineages

(precursors of the MF1 and MF2 subgenomes, tPCK-like genomes [the translocation Proto-Calepineae Karyotype], Cheng et al. 2013, Murat et al. 2015) that contained one copy each of the Ancestral Karyotype (AK) blocks; followed by (2), a subsequent hybridization of the neotetraploid lineage with a third parental diploid species (precursor of the LF subgenome, tPCK-like genome) giving birth to an allohexaploid hybrid. The first step was assumed to be followed by genomic reshuffling and substantial but slightly biased gene fractionation (subgenome MF2 experiencing higher fractionation than MF1), causing eventually a strong dominance of the third subgenome (LF) after step two, and subsequent strongly biased gene-fractionation leaving the LF subgenome much less fragmented than MF1 and MF2 (Cheng et al. 2012; Tang et al. 2012; Cheng et al. 2014). To date, many investigations about this mesopolyploidy event have been done in *Brassica rapa and Brassica oleracea* using genomic data (Wang et al. 2011a; Liu et al. 2014; Murat et al. 2015) and it has been demonstrated that the investigated Brassica species have experienced the same WGT event. Comparative chromosome painting (CCP) studies have been performed in some other genera of the Brassiceae tribe (Lysak et al. 2005; Lysak et al. 2007) and they showed that all investigated genomic blocks of all studied Brassiceae species were present in triplicates, but the evidence of a single shared WGT event among all members of the Brassiceae tribe is still lacking. Even if the most parsimonious hypothesis is that the WGT occurred in the ancestor of all Brassiceae species, we could not exclude the presence of two (or more) independent WGT events within the tribe.

In this study, devoted to the analysis of transcriptome data from a sample of species covering the Brassiceae tribe phylogenetic diversity, we first tested whether the same complex whole genome triplication (WGT) event was shared among all Brassiceae species. Two distinct approaches were used. The first ones are the largely used intra-genomic Ks distributions representing the divergence among duplicated genes (paralogous and homoeologous genes) (Barker et al. 2009; Tang et al. 2012; Kagale et al. 2014; Geiser et al. 2016; Mandáková et al. 2017). In the intra-genomic Ks distribution, individual peaks represent groups of gene pairs with similar synonymous distances and therefore correspond to large-scale duplication events like genome duplication (Blanc and Wolfe 2004). The second approach used was phylogenomic tree reconstruction (Huerta-Cepas and Gabaldón 2011; Jiao et al. 2011; Li et al. 2015; Marcet-Houben and Gabaldón 2015). Our tree-based method (see Chapter 1) allowed us reconstructing a species phylogeny with both the three parental lineages of the mesopolyploid Brassiceae tribe and extant closely related species. One of the closely related species to the tribe Brassiceae is the mesotetraploid *Orychophragmus*

violaceus (Lysak et al. 2005; Lysak et al. 2007). The second goal of our study was then to determine whether this lineage could represent an extant lineage deriving from the intermediate MF1-MF2 tetraploid ancestor, as suggested by Lysak et al. (2007). We tested as well the incidence of subgenome dominance, within the mesohexaploid lineage, on patterns of molecular evolution genomewide, by analysing K_n/K_s estimates along branches of the reconstructed phylogeny.

Altogether, our results confirm that all extant lineages of the Brassiceae tribe have experienced a single WGT event independently of *Orychophragmus violaceus*. In addition, comparing the transcriptome data of the considered Brassiceae species we brought out convergent retention patterns for some genes but differential gene losses for others. Furthermore, our results show that the moderately fractionated subgenomes (MF1 and MF2) have evolved under a more relaxed selection than the least fractionated subgenome (LF), demonstrating asymmetric evolution among subgenomes in relation to levels of gene expression.

Material & Methods

Transcriptome assemblies and genomic datasets used for the present study

We obtained RNAseq data and constructed 27 individual trancriptome assemblies from five species of the Brassiceae tribe belonging to five well-recognized Brassiceae clades (*Carrichtera annua* from clade Vella; *Schouwia purpurea* and *Zilla spinosa* subsp. *macroptera* from clade Zilla; *Psychine stylosa* from clade Savignya, *Cakile maritima* from clade Cakile, and *Crambe maritima* from clade Crambe; Arias *et al*. 2014) and one outgroup species, *Orychophragmus violaceus* (Table 1). The origin of sampled individuals, the type of collected tissus and the method for obtaining RNAseq data and constructing individual transcriptome assemblies are given in Chapter 1. We also used data from published genomes of five species belonging to two other Brassiceae clades *(Brassica nigra*, *Raphanus raphanistrum* and *R. sativus* from clade Nigra; *B. oleracea* and *B. rapa* from clade Oleracea; Table 1), as well as four additional outgroup species (*Arabidopsis thaliana*, *Eutrema salsugineum*, *Schrenkiella parvula* and *Sisymbrium irio*).

Table 1. Transcriptome assemblies obtained in this study and reference of published genomic dataset analysed. Outgroup species (outside of the Brassiceae tribe) are highlighted in grey. Numbers in brakets display the numbers of individuals sequenced for each species.

Data	Species	Brassicaceae clade[a]/lineage[b]	Brassiceae clade	BioProjects /source
Genomic data (coding sequences)	*Arabidopsis thaliana*	A/I	-	TAIR
	Eutrema salsugineum	B/II	-	PRJNA73205
	Schrenkiella parvula	B/II	-	PRJNA63667
	Sisymbrium irio	B/II	-	PRJNA202979
	Brassica nigra	B/II	Nigra	PRJNA285130
	Brassica oleracea	B/II	Oleracea	Brassicadb.org[c]
	Brassica rapa	B/II	Oleracea	PRJNA59981
	Raphanus raphanistrum	B/II	Nigra	PRJNA209513
	Raphanus sativus	B/II	Nigra	PRJDB1517
Transcriptome assemblies	*Cakile maritima* (12)	B/II	Cakile	-
	Carrichtera annua (3)	B/II	Vella	-
	Crambe maritima (3)	B/II	Crambe	-
	Psychine stylosa (2)	B/II	Savignya	-
	Schouwia purpurea (3)	B/II	Zilla	-
	Zilla spinosa subsp. macroptera (2)	B/II	Zilla	-
	Orychophragmus violaceus (2)	B/II	-	-

[a] according to Huang et al. (2016).
[b] according to Franzke et al. (2011).
[c] genome from Liu et al. (2014).

Inferring ancient polyploidy events from patterns of substitutions among duplicated genes (paralogs and homoeolgs)

Following Blanc and Wolfe (2004), we inferred ancient WGD events from the distributions of synonymous substitutions (*Ks*) among duplicated gene copies (i.e. paralogous and homeologous copies) within each transcriptome assembly (one individual per species, those with the highest base pairs [bp] number [Table S2 in Chapter 1]) and from the cds of *Brassica rapa*, *B. oleracea*, *B. nigra*, *Raphanus raphanistrum* and *R. sativus* (see Table 1 for references). To obtain the *Ks* distribution among duplicated gene copies and detect potential WGD events we used the WGDetect pipeline (developed by Nills Arrigo et al.; available at https://github.com/arrigon/WGDetect) and additional custom scripts as described in Geiser et al. (2016). Only one isoform per transcript, the longest, was selected in each transcriptome assembly before starting the analysis to avoid genes to be represented several times in the data because of alternative splicing. The pipeline performs the analysis in the following three steps.

Firstly, annotation of each transcriptome was performed using BLASTx against the TAIR10 (The Arabidopsis Information Ressource) peptide database (e-value = 1^{e-20}) after filtering out transposable elements and organelle genes. All annotated contigs meeting the e-value criterion were used in subsequent analyses. Then, reading frames were identified and extracted for each genic contig through comparison with the best-match protein from the TAIR10 database using Exonerate (Slater and Birney, 2005).

Secondly, framed contigs showing more than 40% sequence similarity over at least 300bp were grouped into gene family clusters. Transcripts from clusters with at least two members were aligned using MACSE (Ranwez et al. 2011). Corresponding phylogenetic trees were inferred with FastTree (Price et al. 2009) and *Ks* values for each duplicate pair were calculated using the maximum likelihood method implemented in codeml of the PAML package under the F3X4 model (Yang 2007). We discarded redundant *Ks* values in gene families as described in Blanc and Wolfe (2004). Furthermore, *Ks* values > 2 representing old duplication events were removed because duplicated genes losses and substitution saturation can alter the genetic signal. This may leads to artificial peaks in the *Ks* distribution beyond a defined threshold of 2 (Vanneste et al. 2013).

Thirdly, to account for multiple duplication events within each lineage, Gaussian mixture models (GMM) were used to estimate the parameters of the *Ks* distributions by fitting multiple normal distributions to the data (Barker et al. 2009) using the package mixtools in R

cran (Benaglia et al. 2009). The most likely number of normal distributions fitting the observed *Ks* distribution was tested by a parametric bootstrap analysis as described in Geiser et al. (2016), with the maximum number of expected peaks set to 4. Then, the position and the standard deviation of each normal distribution were estimated by the iterative "expectation-maximization" algorithm. To check the result constancy the analyses were performed several times for each species.

Mapping ancient WGD events on a species phylogeny using a tree-based method

Is the WGT experienced by Brassica sp. shared by all Brassiceae clades?

In the Chapter 1, we developed a pipeline in order to decifer the phylogeny of mesopopylploïd clades. To this aim, the pipeline produces a set of trees displaying the potential subgenome lineages. In order to illustrate the method we have used the Brassicaceae dataset from the present study (except *Crambe maritima*). We will first briefly summarise the methods. 1,344 genes identified by Liu et al. (2014) as retained in three homeologous copies in *B. rapa* and *B. oleracea* and therefore deriving from the two subsequent allopolyploidy events were selected. Corresponding orthologous sequences in *A. thaliana*, as determined by Liu et al. (2014), were obtained from TAIR10. Orthologous sequences for each outgroup species (Table 1) were also extracted using BLASTN (Altschul et al. 1990) with the *A. thaliana* sequence as reference (Fig. 1, step1 in Chapter 1). For species from the Brassiceae tribe, we identified homoeologous sequences corresponding to the homoeologous gene triplets of *B. rapa* and *B. oleracea* by applying BLASTN analyses on each transcriptome assembly or on published genome data (Fig.1- step1 in Chapter 1). After applying a set of filters, we could reconstruct 863 homolog trees containing the three homoeologous groups (Fig.1- step3 in Chapter 1) corresponding to each parental subgenome. We will now refer to these trees as "homoeolog trees". In each homoeolog tree, nodes annotation was performed using the ETE v3 toolkit (Huerta-Cepas et al. 2010; Huerta-Cepas et al. 2016) in order to localize each of the homoeologous groups. Thanks to this tree collection, for each Brassiceae species investigated, we estimated the proportion of homoeologous groups for which we recovered, based on our node annotations, the full homoeolog triplets, only two homoeologous copies or a single one.

Does Orychophragmus share an ancient WGD event with the Brassiceae lineage?

According to the hypothesis of a two-step origin of the tribe Brassiceae (two subsequent WGD events associated with allopolyploidy), it is possible that some extant lineages share only the first WGD event with members of the Brassiceae tribe (Lysak et al.

2007). To determine whether the mesotetraploid species *O. violaceus* belongs to such a lineage, we focused on the homoeolog trees in which there were two homoeologous copies for at least one of the two *O. violaceus* individuals used in this study. Among these trees, we determined whether the two homoeologous copies of *O. violaceus* belonged to a monophyletic group or not. If not, we examined whether each copy could be considered as a sister group to a different Brassiceae parental subgenome, in agreement with the hypothesis of one shared WGD event with Brassiceae (Fig. S1). Any deviation from this pattern would be considered as a deviation from the above-mentioned hypothesis, and therefore supporting a fully independent WGD event in the history of *O. violaceus*.

Assessing the phylogenetic relationships among the three parental lineages of the Brassiceae tribe and their closely related genera Orychophragmus and Sisymbrium

For each of the 863 homoeologous trees, we estimated the relative occurrence of different patterns of phylogenetic relationships among the three parental lineages (LF, MF1 and MF2) when they were well supported (*i.e.*, bootstrap support ≥ 80). In the first chapter, mainly interested in reconstructing the Brassiceae's phylogeny, 732 genes were concatenated in a supermatrix in which concatenation of the homoeologous copies was done according to their original subgenome. The sequences from Brassiceae species were extracted separately in each of the three ortholog sub-trees of 732 homoeologous trees, labelled accordingly (*e.g.* species_1_LF, species_1_MF2 and species_1_MF2), and concatenated in a single fasta file along with the corresponding sequence of each outgroup species, resulting in the HOMOLOGOUS supermatrix (Fig.1- step5 in Chapter 1). Only the longest sequence of *O. violaceus* in each homoeologous group containing at least one sequence for this species (695 homoeologous groups) was extracted, because homoeologs for this mesotetraploid species cannot be assigned to the Brassiceae parental subgenomes and labelled accordingly. Picking the longest Orychophragmus sequences could maybe bias the outcome of the analysis; therefore a second approach was used here. Thus, we extracted the longest sequence for this species only in the homoeologous trees in which the sequences of this species were monophyletic (161 homoeologous trees). Best-fit partitioning schemes and models were selected using the rcluster search mode implemented in PartitionFinder 2 (Lanfear et al. 2017), under the corrected Akaike Information Criterion (AICc) as suggested by the authors. Then, we built a species tree using RAxML v8.2.10 (Stamatakis 2014) under the GTR + G model, with a partitioned scheme according to the output of PartitionFinder2 (Lanfear et al. 2017). An automatically determined number of bootstrap replicates was set to assess node

supports (Stamatakis 2006; Stamatakis et al. 2008; Stamatakis 2014). A bayesian analysis was conducted with MrBayes (Ronquist et al. 2012) using a specific model for each DNA partition, according to the best-partition scheme selected by PartitionFinder 2. Two 75,000,000 generations runs were completed with four chains each and trees were sampled every 2000 generations. Plots of the likelihood-by-generation were drawn to check chain convergence also indicated by the average standard deviation of split frequencies smaller than 0.01, the Potential Scale Reduction Factor (PSRF) at 1.0 and the effective sample size (ESS) values greater than 100. The first 25% of trees from all runs were discarded as burn-in. A 50% majority-rule consensus of the trees in the posterior distribution from the two runs was obtained with posterior probabilities (PP).

Testing for the effect of subgenome dominance on patterns of molecular evolution in the mesohexaploid genomes of Brassiceae

In order to test for an effect of subgenome dominance processes on patterns of molecular evolution, i.e. with the hypothesis that the dominated subgenomes evolve under relaxed purifying selection as compared to dominant subgenomes, we estimated the ratio of non-synonymous over synonymous substitutions (ω) using the maximum likelihood method implemented in codeml of the PAML package (Yang et al. 2007). We used the F3X4 model of sequence evolution and the "branch model" allowing two or more distinct ω ratios across branches of a phylogenetic tree with constrained topology. We applied this analysis to the brassiceae's subgenomes phylogeny and the associated framed DNA matrix after removing *O. violaceus* and *S. irio* ("all Brassiceae" data, 863 genes, 334,458 codons) which produced a topology with subgenome LF as a sister clade to the pair MF1-MF2 (Fig. 2). *Orychophragmus violaceus* and *S. irio* were removed because of their phylogenetic position close to the LF parental lineage implying to incorporate more parameters in the analyses that were not useful for our question. Four different branch models were tested (Fig. 2): (1) a model with no variation of ω among branches where one ω value was computed for all branches (ω_0); (2) a model where the three Brassiceae subgenomes together were allowed to evolve with a different ω ratio ($\omega_{LF\text{-}MF1\text{-}MF2}$) from that of the outgroup taxa (ω_0); (3) a model where the LF subgenome was allowed to evolve with a different ratio (ω_{LF}) from that of both the two MF subgenomes ($\omega_{MF1\text{-}MF2}$) and the outgroup taxa (ω_0); and (4) a model where the ω values of the three parental subgenomes (ω_{LF}, ω_{MF1} and ω_{MF2}) were allowed to be both different from each other and from the outgroup taxa (Fig. 2). Because we do not know the exact timing of the WGT event, which occurred somewhere after the split of the three parental

lineages and before the first speciation event within the Brassiceae, we performed two separate analyses for the last three models (Fig. 3). The first analysis consisted to exclude the ancestral branch of each subgenome clade (before Brassiceae radiation) in the calculation of lineage-specific ω values. On the contrary, the second analysis consisted to include it in the calculation. Indeed, if the WGT event occurred rapidly after the split of the parental lineages, the effect of subgenome dominance on patterns of molecular evolution could already be observed along those ancestral branches. A likelihood-ratio test was then performed on results from the codeml analysis to test for an overall effect of the WGT event on the strength of purifying selection (comparison between model 1 and model 2) and to determine the best model of ω variation among Brassiceae parental subgenomes (comparisons between model 2 and 3, and between model 3 and 4).

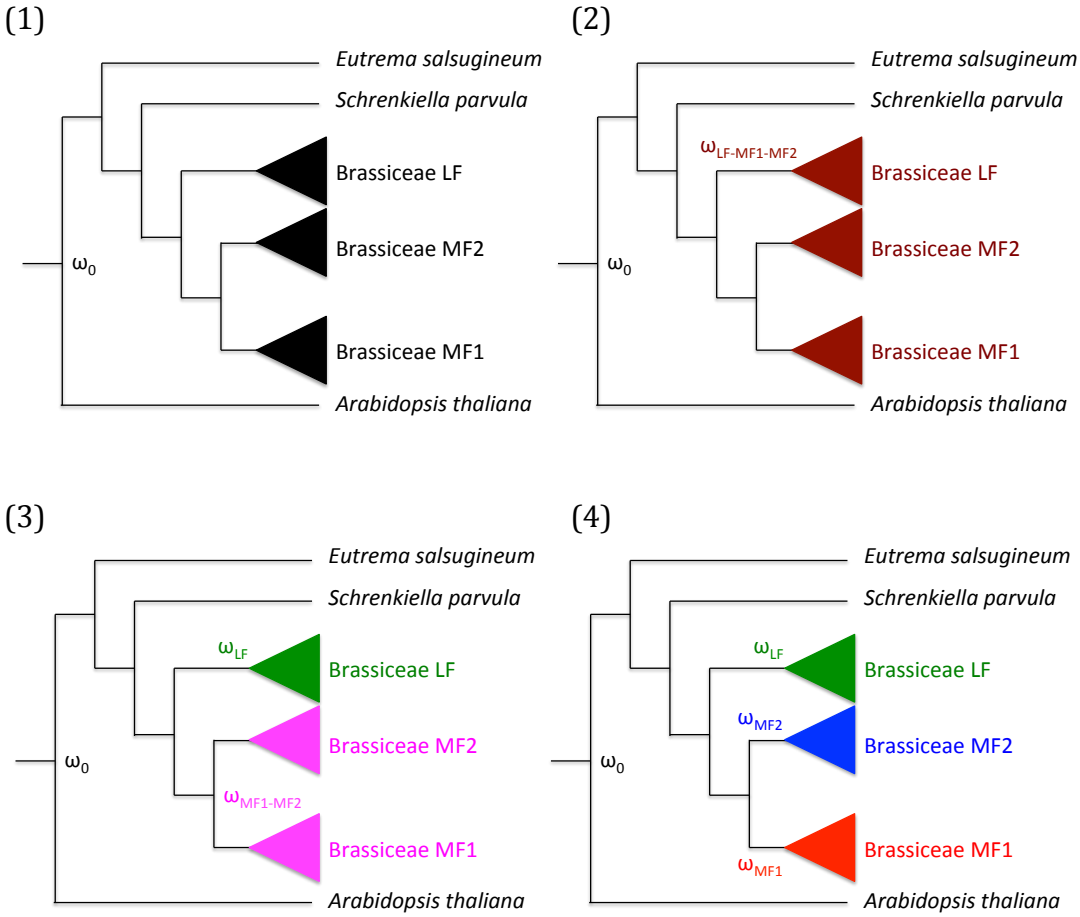


Fig. 2. The four alternative "branch-models" used in the calculation of lineage specific ω ratios for testing the effect of subgenome dominance on patterns of molecular evolution. (1) Model with no variation of ω among branches (a single parameter: ω_0); (2) model with a distinct ω ratio for the Brassiceae species (two parameters: ω_0 and $\omega_{LF\text{-}MF1\text{-}MF2}$); (3) model with a distinct ω ratio for the LF and both MF subgenomes (three parameters: ω_0, ω_{LF} and $\omega_{MF1\text{-}MF2}$); and (4) model with different ω ratios among the different subgenomes (four parameters: ω_0, ω_{LF}, ω_{MF1} and ω_{MF2}).

This overall dataset contains many missing data, especially within the MF1 and MF2 subgenomes, and this could possibly bias the analysis as some missing data may correspond to the most weakly expressed (and thus undetected) transcripts that evolve under more relaxed selection. In order to control for the effect of missing data in the overall dataset, we could remove genes with missing data however the remaining dataset was then too small to be analysed. We therefore performed the analysis on a subset of species for which a whole genome dataset was available, using alternative sources of data to ensure that the triplets were fully represented for each homoeologous group in both the Oleracea and the Nigra clades ("Oleracea & Nigra" dataset, 290 genes, 122,145 codons). For the Oleracea clade, we used sequences from *B. rapa* or *B. oleracea* when data from *B. rapa* was missing. For the Nigra clade, we used sequences from *B. nigra* and *R. sativus or R. raphanistum* when data from one of both was missing.

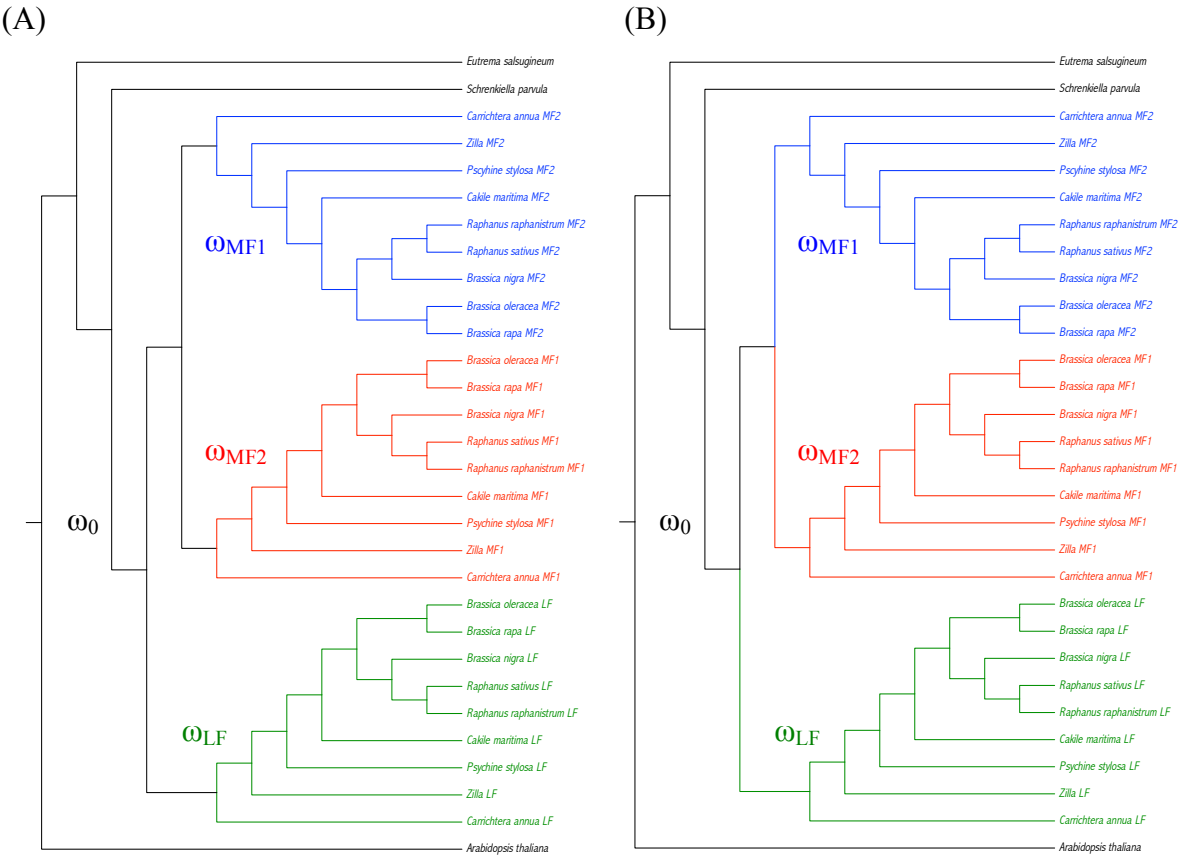


Fig. 3. Two types of analysis were conducted to test whether the ω ratio is different between the three parental sub-genomes (lineages): (A) the "Excluded" analysis and (B) the "Included" analysis, in which the three tested models (see Fig. 2) excludes or includes the branch leading to parental lineage in the calculation of the lineage specific ω ratio, respectively. In green: branches used to compute ω_{LF} value, in red: branches used to compute ω_{MF1} value, in blue: branches used to compute ω_{MF2} value, in black: branches used to compute ω_0 value.

Results

Identification of ancient polyploidy events

The frequency distribution of synonymous substitutions (*Ks*) between paralogous pairs of transcripts was analyzed to detect ancient polyploidy events in each investigated species. Histograms representing the distributions of *Ks* values fitted with Gaussian mixture models for each species are presented in Fig. S2. For each studied species, at least three major peaks were detected, indicating the presence of at least three large-scale duplication events in species of the Brassiceae tribe as well as in the close outgroup *O. violaceus* (Table 1). The mean *Ks* values of the two most ancient events are consistent with estimates for the α and β paleopolyploidy events shared by all Brassicaceae species (0.70 and 1.73, Barker et al. 2009, 0.77 and 2.05, Kagale et al. 2014). The mean *Ks* values of the most recent large-scale duplication event in Brassiceae species (except *Carrichtera annua*) are consistent with the *Ks* estimates obtained for the mesohexaploidy event identified in *Brassica rapa* from genomic and transcriptomic data (0.37-0.38; Tang et al. 2012, Kagale et al. 2014). Furthermore, in the outgroup *O. violaceus*, the mean *Ks* value of the most recent WGD event appears to be slighlty lower than the estimate for the Brassiceae species, which could suggest an independent and more recent WGD event in *O. violaceus*, but the *Ks* value associated with the α event is also smaller in this species compared to those of Brassiceae species, which suggests a different molecular evolutionary rate in *O. violaceus*. In the same way, the mean *Ks* values of the most recent event and the α event in *C. annua* is slightly higher than the estimate of other Brassiceae species, suggesting here again a different molecular evolutionary rate in this species. Our results confirm the presence of at least three large-scale duplication events in each investigated species but, because of molecular evolutionary rate variation among lineages, *Ks* distributions cannot be directly compared.

Table 1. Estimates of the mean (± SD) of individual components of the *Ks* distributions for each investigated species, as obtained by mixture models fitted to normal distributions.

	Mean *Ks* ± SD of mixture model components		
Species	Recent WGD	α WGD	β WGD
Brassica rapa	0.330 ± 0.102	0.760 ± 0.345	1.709 ± 0.182
Brassica oleracea			
Brassica nigra	0.320 ± 0.115	0.814 ± 0.330	1.715 ± 0.174
Raphanus raphanistrum	0.344 ± 0.125	0.850 ± 0.323	1.725 ± 0.168
Raphanus sativus			
Cakile maritima	0.369 ± 0.096	0.827 ± 0.312	1.671 ± 0.201
Crambe maritima	0.356 ± 0.103	0.830 ± 0.298	1.665 ± 0.205
Psychine stylosa	0.377 ± 0.107	0.854 ± 0.310	1.685 ± 0.199
Schouwia purpurea	0.348 ± 0.098	0.805 ± 0.299	1.672 ± 0.202
Zilla spinosa subsp. macroptera	0.360 ± 0.101	0.824 ± 0.303	1.682 ± 0.197
Carrichtera annua	0.420 ± 0.134	0.889 ± 0.317	1.680 ± 0.194
Orychophragmus violaceus	0.285 ± 0.094	0.786 ± 0.298	1.698 ± 0.191

A single shared WGT among all Brassiceae species

Among the 863 homoeologous gene trees obtained with the methodology developed in the first Chapter, we counted the number of homoeologous trees in which at least one homeoelogous copy for a given species was present. For the Brassiceae species with published genome data the number varied from 805 (*R. raphanistrum*) to 863 (*B. rapa*; Table 2). For species with transcriptome data, this number ranged between 638 (*Z. spinosa*) and 799 (*S. purpurea*). These differences reflect the missing data from transcriptome sequencing. Based on these sub-trees annotations we estimated, for each species, the proportion of homoeologous trees for which we recovered either the full triplets (LF, MF1 and MF2 copies), or only two homoeologous copies (LF and MF1, LF and MF2, or MF1 and MF2), or a single copy (LF, MF1 or MF2; Table 2). For example, for *B. rapa* full triplets were present for only 794 out of the 863 homoeologous trees (92%) as a result of the alignment trimming procedure. For all other species with genomic datasets, 23 to 85% of homoeologous trees contained one copy for each of the three subgenomes, whereas 7 to 47% contained one copy for only two of the three subgenomes and 2.6 to 30% contained one copy for only one subgenome, most often the LF subgenome (Table 2). For the species with transcriptomic datasets, 1.6 to 16% of homoeologous trees contained one copy for each of the three subgenomes (triplets), whereas 20 to 44% contained one copy for only two of the three subgenomes and 40 to 78% contained one copy for only one subgenome, most predominantly the LF subgenome (Table 2).

Table 2. For each studied Brassiceae species, the percentage of homoeologous trees in which singletons (LF, MF1 or MF2), duplicates (LF and MF1, LF and MF2 or MF1 and MF2) or triplets (LF, MF1 and MF2) have been found is indicated. Both detailed and summary statistics are given.

	Subgenomes								
	Singletons				Duplicates				Triplets
Species (Number of homoeologous trees*)	In total	LF	MF1	MF2	In total	LF & MF1	LF & MF2	MF1 & MF2	LF, MF1 & MF2
Brassica rapa (863)	0.47	0.00	0.12	0.35	7.53	2.90	3.01	1.62	92.00
Brassica oleracea (859)	0.26	0.81	0.58	1.16	11.88	5.01	3.26	3.61	85.56
Brassica nigra (848)	8.14	3.42	1.77	2.95	33.73	13.21	10.02	10.50	58.14
Raphanus sativus (853)	17.01	6.92	5.28	4.81	46.42	19.34	15.83	11.25	36.58
Raphanus raphanistrum (805)	29.94	11.43	9.94	8.57	47.2	13.91	19.38	13.91	22.86
Cakile maritima (798)	39.97	18.67	11.15	10.15	44.11	15.91	17.92	10.28	15.91
Psychine stylosa (746)	63.94	30.83	17.83	15.28	31.5	10.59	11.93	8.98	4.56
Schouwia purpurea (799)	55.07	27.03	15.27	12.77	38.92	15.89	12.77	10.26	6.01
Zilla spinosa (638)	64.27	28.84	18.97	16.46	33.39	15.05	11.44	6.90	2.35
Carrichtera annua (741)	78.41	34.28	22.00	22.13	19.97	8.91	6.88	4.18	1.62

*Number of homoeologous trees in which there is at least one homoeologous copy for the species under consideration.

Our results show that the recovering of triplets is highly influenced by the type of data used (genome versus transcriptome datasets) and the number of individuals sequenced per species (*e.g.* 12 individuals were sequenced for *Cakile maritima* for which the proportion of triplets is by far the largest, while only two to three individuals were sequenced for the other species). Despite the lower recovery of some homoeologous copies in several species, our results from genome and transcriptome datasets clearly indicate that all investigated species carry homoeologs from each of the three subgenomes and thus are sharing the same WGT event in their history.

Independent ancient WGD event in O. violaceus

Among the 863 homoeologous trees analyzed, 168 trees did not display any sequence of *O. violaceus,* 109 and 305 trees showed a single sequence for one or both individuals, respectively, whereas 281 homoeologous trees (called O-trees, Fig.4) contained two sequences for at least one individual of *O. violaceus*. The homoeologous sequences of *O. violaceus* were monophyletic with a strong support value (Bootstrap percentage [BP]≥80) in

128 of the O-trees and with a weak support (BP<80) in 33, representing overall 57% of the total number of O-trees (Fig. 4). In 120 of the O-trees, the sequences of *O. violaceus* were not monophyletic (43%).

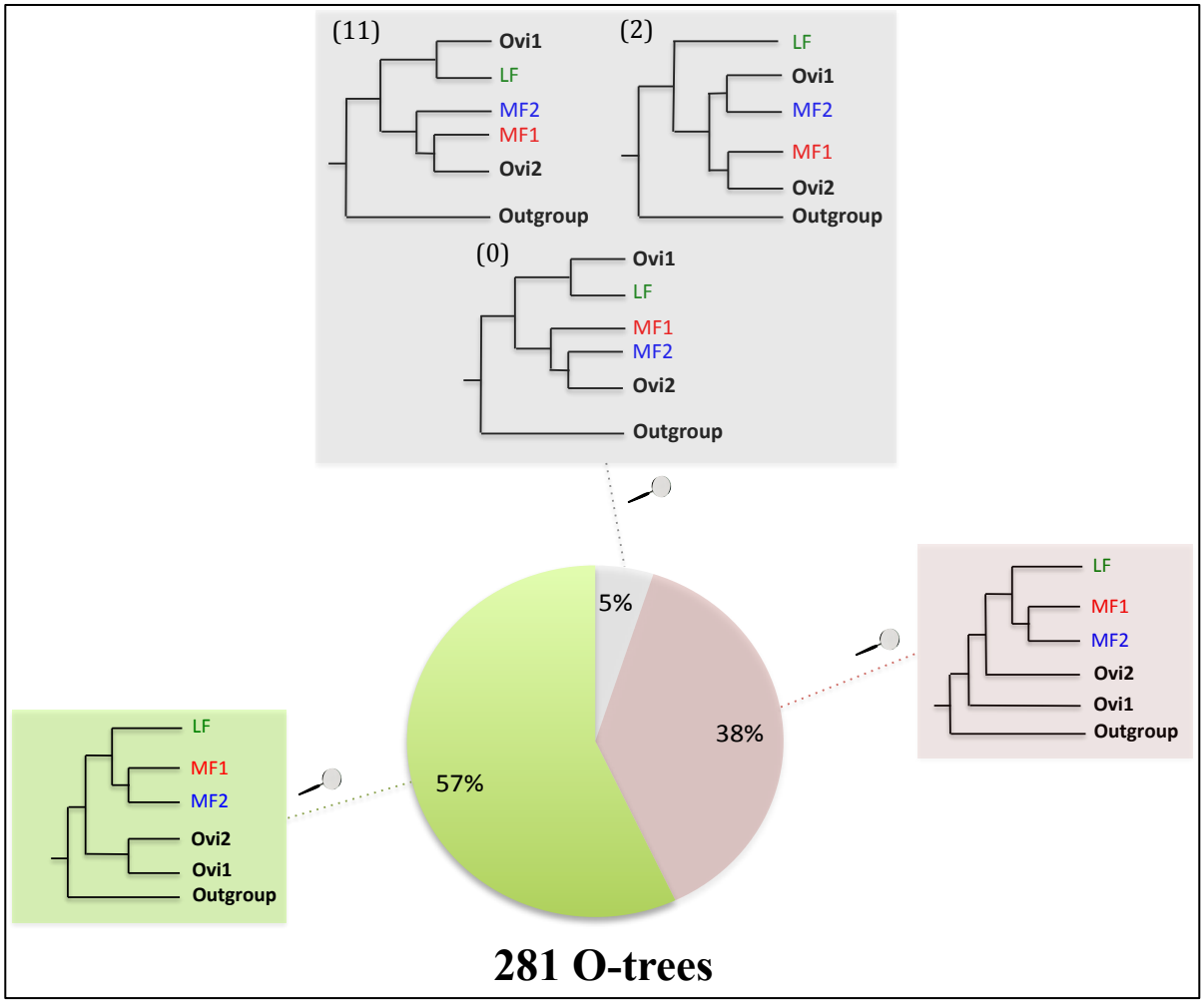


Fig. 4. Pie chart representing the percentage of O-trees (see main text) with a topology supporting (in grey, see Fig. S1) or not (in brown and in green) a shared WGD between *O. violaceus* and the Brassiceae tribe. Expected topologies for a shared (highlighted in grey) or an independent WGD (highlighted in brown and green) are indicated. In 57% of O-trees, sequences of *O. violaceus* are monophyletic as exemplified in the green box, supporting an independent WGD event in this species. In 38% of O-trees, sequences were not monophyletic but the topology did not support a shared WGD event, as exemplified in the brown box. Only 5% of O-trees had a topology consistent with the three expected topologies under a scenario of a shared WGD event between *O. violaceus* and the Brassiceae tribe (see Fig. S1). The number of O-trees supporting each expected topology is indicated in brackets. Ovi1: copy 1 of *O. violaceus*, Ovi2: copy 2 of *O. violaceus*. In the illustration, only one individual of *O. violaceus* is represented for easier comprehension.

Among them, in only 13 (5%) trees, the two sequences were sister groups to two distinct parental Brassiceae subgenomes (LF and MF1 in 11 trees, and MF1 and MF2 in 2 trees), with a high (8) or a weak bootstrap support value (5), whereas the remaining 107 O-

trees did not support any expected topologies (Fig. 4. Fig. S1). Altogether, these results support the hypothesis that the genus Orychophragmus does not share an historical WGD event with the tribe Brassiceae.

Phylogenetic relationships among the three parental lineages of the tribe Brassiceae and their closely related genera

After annotating the three orthologous sub-trees in the 863 homoeologous trees analyzed, we observed that the topology of the three parental lineages was unresolved (BP < 80) for most of them (752/863 gene trees, *i.e.* 87% of trees). For the 111 gene trees where the phylogenetic relationships among parental lineages could be resolved, we observed roughly similar frequencies for the following three scenarios: (1) LF clade first diverging from the ancestor of the MF1 and MF2 clades in 36 gene trees (32%); (2) MF1 clade first diverging from the ancestor of the LF and MF2 clades in 43 gene trees (39%); and (3) MF2 clade first diverging from the ancestor of the MF1 and LF clades in 32 gene trees (29%). Using a supermatrix approach (see Chapter1), 732 out of the 863 genes were concatenated in a supermatrix in which concatenation of the homoeologous copies was done according to their original subgenome. For the Orychophragmus species, two approaches were considered: either genes from 695 trees with at least one *O. violaceus* sequence were used (see Chapter1) or only genes from 161 trees in which multiple *Orychophragmus* sequences were monophyletic were kept for the phylogenetic reconstructions. We obtained two different topologies depending on the treatment of the *O. violaceus* sequences. In both topologies (Fig. 5), the MF1 and MF2 parental lineages were more closely related to each other than to the LF parental lineage (BS=100. PP=1), and *Sisymbrium irio* constituted a sister clade to LF (BS>94. PP=1). The position of *Orychophragmus violaceus* varied between the two topologies, being either a sister clade to the clade {*S. irio*, LF} or being a sister clade to the larger clade {MF1, MF2, LF, *S. irio*} (Fig. 5). These results suggest that *S. irio* and perhaps also *O. violaceus*, can be considered as extant members of a lineage that contained historically the ancestor of the LF subgenome of Brassiceae, and thus cannot be considered formally as Brassiceae outgroups.

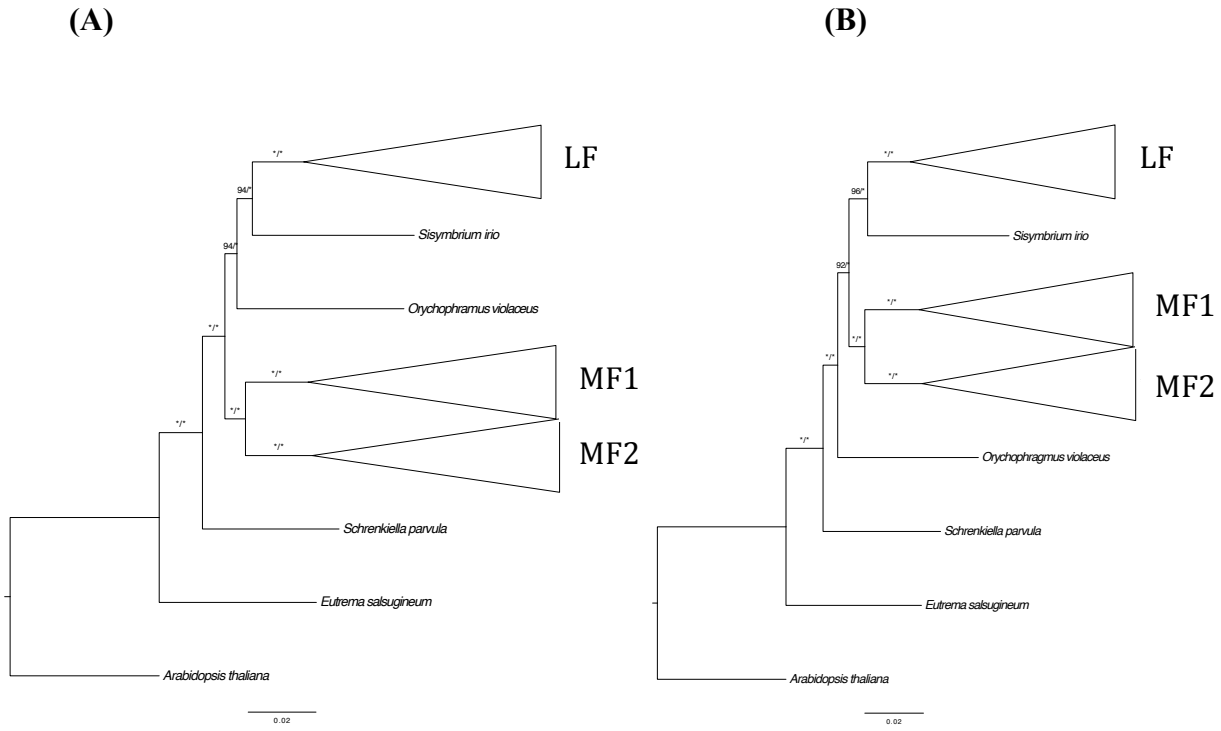


Fig. 5. Maximum likelihood phylogenies of the three parental lineages of the tribe Brassiceae and their closely related species. The three sub-trees (collapsed) correspond to the three subgenomes of the Brassiceae species (LF: top, MF1: middle, MF2: bottom). Numbers above branches represent bootstrap support values for the ML analysis (BP). (*) indicates nodes with maximal support (BP = 100/PP=1). The topology of the Brassiceae tribe in each sub-tree was masked (shown in Chapter 1). (A) Topology obtained using the longest sequence of *O. violaceus* present in homoeologous trees that contained at least one sequence, (B) topology obtained using the longest sequence of *O. violaceus* present in homoeologous trees where sequences of this species were monophyletic.

Testing for the effect of subgenome dominance on patterns of molecular evolution in the mesohexaploid genomes of Brassiceae

For the two alignments and for the analyses including or excluding the ancestral branch leading to each subgenome, the results indicated highly significant variations in the ω ratio among sub-trees (*i.e.* sub-genomes) and the best model (with $P<0.001$) was the model 4, the "LF versus MF1 versus MF2" model (Table 3 and S1). The value of the ω ratio for LF subgenome was consistently lower than the value for both MF1 and MF2 subgenomes. Moreover, the value of the ω ratio for subgenome MF2 was consistently higher than the value for subgenome MF1 (Table 3 and S1). The values of the ω ratio for subgenomes LF, MF1 and MF2 were substantially higher in the analysis with the subgenome ancestral branches included than with those branches excluded, and the reverse was true for the branches connected to the outgroups, suggesting that the period of relaxed purifying selection started somewhat before the diversification of Brassiceae but shortly after the split between the LF,

MF1 and MF2 parental lineages. The effect of including or not the ancestral branches for the LF, MF1 and MF2 subgenomes was much higher for the Oleracea & Nigra than for the whole dataset, most probably because these branches become much longer when only two sister clades within Brassiceae are considered (Fig. 3).

Table 3. Estimated values of the ω ratio for each tested model while including ("Included") or excluding ("Excluded") the ancestral branch of each parental lineage in the calculation of the lineage specific ω (see Fig. 2 & 3). The values of the estimates (ω) for the four models and the *P*-value for the best model (see table S1) are reported. The number of codons corresponding to each DNA alignment is reported under the dataset names. Bold values emphazise the best model estimates with associated statistic probability.

Models	Number of lineages	ω ratio			
		Oleracea & Nigra (122,145 codons)		*all Brassiceae* (334, 458 codons)	
		Sub-genomes ancestral branches		Sub-genomes ancestral branches	
		Excluded	Included	Excluded	Included
(1) One rate	**1** (1: all branches)	$\omega_0 = 0.181$		$\omega_0 = 0.182$	
(2) LF, MF1 and MF2	**2** (1: LF, MF1 and MF2 lineages, 2: all others branches)	$\omega_0 = 0.184$ $\omega_{LF\text{-}MF1\text{-}MF2} = 0.179$ (P<0.01)	$\omega_0 = 0.159$ $\omega_{LF\text{-}MF1\text{-}MF2} = 0.191$ ($P < 0.001$)	$\omega_0 = 0.176$ $\omega_{LF\text{-}MF1\text{-}MF2} = 0.184$ ($P < 0.001$)	$\omega_0 = 0.161$ $\omega_{LF\text{-}MF1\text{-}MF2} = 0.187$ ($P < 0.001$)
(3) LF versus MF1 and MF2	**3** (1: LF lineage, 2: MF1 and MF2 lineages 3: all other branches)	$\omega_0 = 0.184$ $\omega_{LF} = 0.167$ $\omega_{MF1\text{-}MF2} = 0.185$ ($P < 0.001$)	$\omega_0 = 0.159$ $\omega_{LF} = 0.179$ $\omega_{MF1\text{-}MF2} = 0.198$ ($P < 0.001$)	$\omega_0 = 0.176$ $\omega_{LF} = 0.177$ $\omega_{MF1\text{-}MF2} = 0.188$ ($P < 0.001$)	$\omega_0 = 0.162$ $\omega_{LF} = 0.180$ $\omega_{MF1\text{-}MF2} = 0.191$ ($P < 0.001$)
(4) LF versus MF1 versus MF2	**4** (1: LF lineage, 2: MF1 lineage, 3: MF2 lineage, 4: all other branches)	**$\omega_0 = 0.184$** **$\omega_{LF} = 0.167$** **$\omega_{MF1} = 0.173$** **$\omega_{MF2} = 0.197$** **($P < 0.001$)**	**$\omega_0 = 0.159$** **$\omega_{LF} = 0.179$** **$\omega_{MF1} = 0.189$** **$\omega_{MF2} = 0.207$** **($P < 0.001$)**	**$\omega_0 = 0.176$** **$\omega_{LF} = 0.177$** **$\omega_{MF1} = 0.183$** **$\omega_{MF2} = 0.192$** **($P < 0.001$)**	**$\omega_0 = 0.162$** **$\omega_{LF} = 0.180$** **$\omega_{MF1} = 0.187$** **$\omega_{MF2} = 0.195$** **($P < 0.001$)**

Discussion

Shared mesohexaploidy (WGT) among Brassiceae subclades

Previous studies using comparative chromosome painting in a set of species representing all Brassiceae subclades (except the Henopython subclade) have shown that all Brassicaceae ancestral genomic blocks investigated appeared to be triplicated within the Brassiceae tribe (Lysak et al. 2005; Ziolkowski et al. 2006; Lysak et al. 2007), suggesting that all Brassiceae species experienced an historical whole genome triplication event. Although this mesohexaploidy event was detected in the *B. rapa* and *B. oleracea* genomes (Wang et al. 2011; Liu et al. 2014), two species belonging to the same Oleracea subclade, it had to be confirmed with genomewide sequence data for the other major Brassiceae clades. By analyzing distributions of synonymous substitutions among paralogous sequences in original

transcriptome datasets obtained from five additional Brassiceae clades, we confirmed that each of them went through a mesopolyploidy event. It should be noted that the observed peak, representing an allopolyploidy event, indicates the timing of divergence of the parental diploids lineages and not their merger to form the allopolyploid (Kagale et al. 2014). Only one peak is observed in *Brassica rapa* and *B. oleracea*, whereas two allopolyploidy events have occurred, suggesting that the two speciation's events creating the three diploid parental species of *Brassica rapa* and *B. oleracea* were close in time. Only one peak is also observed for all the other Brassiceae species investigated in our study. If we build up on CCP and genomic results and consider that the polyploidy event is a WGT, our results suggest either that they share the same WGT event than *B. rapa* and *B. oleracea* or that they have experienced an independent WGT event but with the same specifity of those of *B. rapa* and *B. oleracea* namely a rapid divergence between the three parental lineages.

However, such approaches are associated with large variance of the estimates of the peaks ages and cannot readily distinguish between WGD and WGT events. Futhermore, our analysis used transcriptome assemblies, which are error-prone, and may slightly overestimate *Ks* values, as it is observed when we compare our *Ks* values obtained from genomic data and from transcriptome data. Higher *Ks* values associated with the most recent large-scale duplication event in Brassiceae species with transcriptome data might also be explained by the fact that most comparisons between duplicated gene pairs represent mostly LF-MF1 and LF-MF2 comparisons, whereas the MF1-MF2 comparisons are probably rare, as MF duplicates were not recovered as often as LF (Table 2). Thus, the inferred peak may rather represent the divergence between the parental LF species and the ancestor of the two parental species MF1 and MF2 than the divergence between these two latter species, which is observed to be more recent (Fig. 5). Hence, even if the most likely scenario is that the ancestor of all extant Brassiceae species has experienced the mesohexaploid event identified in *B. rapa* and *B. oleracea*, several independent but contemporaneous WGT or WGD events in different lineages might not be excluded.

We then used a phylogenomic approach to check for the occurrence of a WGT, and not a WGD, in the history of all Brassiceae subclades, and to determine if they all experienced the exact same mesohexaploid event. However, the use of transcriptome data in complex mesohexaploid lineages raises some difficulties of dataset unbalance, as illustrated by the sharp differences in the rate of recovery of triplets of homoeologous copies according to the type of data used (genomic versus transcriptomic). The lower rate of recovery of

homoeologous copies from transcriptomes is probably due, at least in part, to the weak or null expression of some copies because of the physiological timing and/or the tissue sampled, and as a result of differences in patterns of gene expression among homoeologs, which may also be different depending on the tissue (Cheng et al. 2012; Parkin et al. 2014; Pfeifer et al. 2014). From transcriptomic datasets, we recovered more triplets and more duplicates in *C. maritima* for which 12 individuals have been sequenced. By sequencing several individuals, we probably improved our ability to catch genes whose the expression depends on physiological timing or other processes. Differences in patterns of gene expression among homoeologs would explain our observation of a substantially higher rate of recovery for LF than for MF1 and MF2 homoeologs based on transcriptome datasets (although only homoeologs maintained as triplets in *B. rapa* and *B. oleracea* were analysed), knowing that it has been shown in *B. rapa* and *B. oleracea* that homoeologs from the MF1 and MF2 subgenomes are expressed overall at lower levels than those from the LF subgenome (Cheng et al. 2012, Liu et al. 2014). Additional sources of differences between transcriptomic and genomic datasets may also be associated with less efficient assembly procedures in transcriptomes for homoeologs with low sequence coverage or those with low divergence among homoeologous copies. Beside idiosynchracies associated with the analysis of transcriptome data, we also noted in the analyses based on genomic data that a substantial proportion of homoeologous copies present as triplets in *B. rapa* and *B. oleracea* were present as duplicates, as singletons or were totally absent in the genomes of *B. nigra*, *R. raphanus* and *R. raphanistrum*. Although these differences could be due in part to the lower quality of the genome assemblies of these species from the Nigra subclade, they suggest the occurrence of differences in patterns of gene losses among Brassiceae subclades. This confirm the study of Moghe et al. (2014) who performed a comparative analysis of homoeologs retention in *B. rapa* and *R. raphanistrum* genomes and concluded that most of the losses of homoeologous copies within subgenomes occurred prior to their divergence but several losses postdate the divergence of the two species. Hence, this suggests that the loss of homoeologous copies following a WGD event is a continuous process that proceeds a long time after that event. Another explanation for the apparent heterogeneous gene losses between species could be ectopic gene conversion between homeologs (Wang et al. 2009; Wang et al. 2011; Wang and Paterson 2011; Scienski et al. 2015). Indeed, after gene duplication, gene conversion can affect one of the two paralogous copies. In our homoeologous trees, both copies should then appear as very similar, and fall within the same subgenome sub-tree. According to the procedure we have used, one such copy would be discarded and this would leave a missing

data within the subgenome subjected to gene conversion. For instance, in *B. rapa* and *B. oleracea*, it has been estimated that 8% of duplicates were affected by gene conversion and one-sixth of them have experienced independent conversion events since the split of the two species (Liu et al. 2014).

Although we recovered full triplets in only few homoeologous trees for some species, for the reasons just discussed, our phylogenomic analyses strongly support the scenario of a single shared WGT event among all Brassiceae subclades investigated, as hypothesized by Lysak et al. (2007, 2005). Indeed, based on annotation of subgenome sub-trees in individual homoeolog trees, we found compelling evidence that each species of Brassiceae investigated shares the LF, MF1 and MF2 subgenomes, although with different rates of recovery for the different homoeologs in different subgenomes and different species (Table 2). Moreover, phylogenetic reconstruction based on our supermatrix clearly identified three lineages corresponding to the three parental subgenomes in each species, and identified the genus Sisymbrium as a sister clade to the parental lineage that gave rise to the LF subgenome. Parental species of the Brassiceae tribe are extinct but the extant genus Sysimbrium appears to share a commun ancestor with one of them (the LF diploid parental lineage), which could explain that, in some studies, the Brassiceae tribe and the Sisymbrideae are paraphyletic (Warwick et al. 2002). However, two Sisymbrium species that fell within the tribe Brassiceae have been submitted to nomenclatural changes based on the phylogeny but also on morphological characters (Warwick et al. 2002; Warwick and Al-Shehbaz 2003).

Confirming an independent WGD event in O. violaceus

Surprinsingly, we found various topologies among the O-trees, which could be due to incomplete lineage sorting. This result also suggests that we probably not recovered the two homoeologous copies of *O. violaceus* in all O-trees. Indeed, it is possible that in some O-trees, the two copies found in one or in both *O. violaceus* individuals may be two isoforms (alternative transcripts) or recent duplicated genes (a subcomponent of Trinity does not always correspond to a gene and its splice variants, Grabherr et al. (2011)), thus corresponding to only one parental subgenome of *O. violaceus*. This is supported by our observation that in the 57% of O-trees where the sequences of *O. violaceus* belong to a monophyletic group, two isoforms are often present and the sequences of *O. violaceus* are sister to the clade {MF1, MF2, LF, *S. irio*} (data not shown). This could explain why the position of *Orychophragmus violaceus* varied in our two phylogenetic species trees (Fig. 5), being either a sister clade to the clade {*S. irio*, LF} (first approach) or being a sister clade to

the larger clade {MF1, MF2, LF, *S. irio*} (second approach). This is perhaps because with the first approach the sampled sequences of *O. violaceus* were biased toward a parental lineage that appears to be closed to the clade {*S. irio*, LF} whereas with the second approach, the sampled sequences of *O. violaceus* were biased toward the second parental lineage that appears to be closed to the clade {MF1, MF2, LF, *S. irio*}. However, despite this possible bias, no signal for a shared WGD event between the Orychophragmus lineage and the Brassiceae tribe has been found (Fig. 4).

Contrary to the seducing scenario suggested by Lysak et al. (2005) that the mesotetraploid species *O. violaceus* might be the surviving ancestral genome of the first allopolyploidy event experienced by the ancestor of the Brassiceae tribe, our data suggest that *O. violaceus* has experienced an independent and probably more recent WGD event than the Brassiceae tribe. Furthermore, as also noted by Lysak et al. (2005), the geographical distribution of the species outside the diversity center of the Brassiceae tribe in the southwestern Mediterranean region makes *Orychophramus* a less likely candidate as tetraploid bridge between a diploid lineage and the mesohexaploid Brassiceae (Gómez-Campo 1980; Al-Shehbaz and Guang 2000; Hu et al. 2015).

Two step-model of genome merging

According to the two-step model of allohexaploidy proposed by Tang et al. (2012) for the history of the Brassica lineage, the MF1 and MF2 parental genomes would have merged in a first step, and the resulting tetraploid lineage, after some genomic rearrangements, would have merged in a second step with the LF parental genome. This hypothesis has been suggested to account for the strong biased fractionation and biased patterns of expression of homeologous copies observed in *B. rapa*, where substantially higher expression level and gene retention had been observed in the LF subgenome as compared to MF1 and MF2 (Wang et al. 2011). Hence, it was assumed that subgenome dominance of LF in the mesohexaploid genome of *B. rapa* was due to the fact that the MF1 and MF2 subgenomes had already experienced partial gene losses and rearrangements, at the time they merge with LF. Further support for this hypothesis comes from the observation that regions with particularly low gene densities within either the MF1 or MF2 subgenomes were associated with higher level of gene retention in the other subgenome, suggesting some sort of compensation in the ancestral tetraploid hybrid MF1-MF2 (Wang et al. 2011, Tang et al. 2012). However, interactions between parental genomes in allopolyploids may induce epigenetic changes including DNA methylation (Chen 2007), and such changes have been documented in many plants (Salmon et

al. 2005; Parisod et al. 2009; Song and Chen 2015). It is known that DNA methylation plays a crucial role in the regulation of gene expression and the different methylation patterns of progenitors genomes could probably explain some of the systematic homoeolog expression bias found in several allopolyploids species. Thus, biased gene expression and biased gene retention in allopolyploids is not necessarily associated with the order of genome merging events. Even if our phylogenomic analysis suggests that the MF1 and MF2 parental species were closely related to each other, and that the LF ancestor was a more divergent lineage, that does not exclude a scenario with a first allopolyploid hybridization between the LF and the MF1 (or MF2) species and then, a second allopolyploid hybridization between the newly formed tetraploid hybrid and the MF2 (or MF1) species.

Subgenome dominance in the mesohexaploid tribe Brassiceae

In our study, we tested for differences in selection pressure among subgenome lineages, which is expected in case of subgenome dominance in relation to differential levels of overall gene expression. Indeed, it is expected that the underexpressed, non-dominant subgenome undergoes a relaxation in the strength of purifying selection and thus shows a higher accumulation of non-synonymous substitutions (Zhang et al. 2015) or deletions (Schnable et al. 2011). To date, in the Brassiceae tribe, some investigations were performed by counting the number of non-synonymous SNPs on each of the three subgenomes of *Brassica rapa*, which has revealed a higher number of non-synonymous SNPs on the most fractionated subgenomes (MF1 and MF2) compared to the least fractionated subgenome (LF) (Cheng et al. 2012). Furthermore, Tang et al. (2012) have reported more short exonic deletions in the MF subgenomes than in the LF subgenome. However, their analyses could not estimate quantitatively the differences in selection pressure among subgenomes.

Here, based on a phylogenetic approach composed of three outgroup species and each of the three subgenomes of all studied Brassiceae species using a dataset of 732 homoeologous groups, we estimated separately the patterns of molecular evolution in outgroup lineages and in the lineage of each parental subgenome. In agreement with our hypothesis, the analyses revealed a consistent and significant subgenome bias in the ω ratio, with a lower ω ratio for the LF (0.180) subgenome compared to the MFs subgenomes (0.187 and 0.195). Moreover, the selection pressure relaxation was stronger on the MF2 subgenome than on the MF1 subgenome, in agreement with the higher number of genes from MF1 with dominant expression compared to MF2 (Cheng et al. 2012). Interestingly, our analysis also allowed us to demonstrate an overall increase in the ω ratio following the mesohexaploid

event (overall comparison of the ratio between diploid outgroups and the Brassiceae lineage). The functional redundancy of genes following a whole genome duplication or triplication event is a breeding ground for the relaxation of the purifying selection pressure on some copies (Lynch and Conery 2000; Otto and Whitton 2000; Conant and Wagner 2003; Tang et al. 2012), which may eventually lead to pseudogenization or neo and sub-functionalization. Indeed, Parkin et al. (2014) found that 83% of *B. oleracea* triplicates had different expression patterns depending on the type of tissue, suggesting a functional diversification of the duplicated genes. An analysis in *Raphanus raphanistrum* and *B. rapa* showed that 13.1% and 18.7% of the gene pairs have undergone asymmetric evolution, and that these doublets have a 1.5 × higher ω ratio than duplicate pairs evolving at uniform rates, which suggests that the functional divergence observed in almost one fifth of the doublets could have occurred through an asymmetric divergence of the sequences (Moghe et al. 2014). The rapid accumulation of independent mutations in each genes copy can open the way for asymmetrical evolution of sequences over time.

The results obtained from the second dataset based on genomic data only ("*Oleracea & Nigra*" dataset), are similar to those obtained from our transcriptomic data. In transcriptomic data ("*all Brassiceae*" dataset), genes with low or null level of expression — mostly biased toward the lower expressed MF subgenomes (Cheng et al. 2012, Parkin et al. 2014) — might not be recovered, which could impact the test for dN/dS variation among parental lineages. Our results suggest that this bias seems to be minimal considering the strong signal for ω variation among the three subgenomes that has been detected from the transcriptomic dataset.

As expected, in Brassiceae, the underexpressed, non-dominant subgenomes MF1 and MF2 undergo a relaxed purifying selection pressure and might have evolved faster than the dominant subgenome (LF), which might have caused asymmetric evolution between the subgenomes (Conant & Wagner, 2003), as it has been observed in the allotetraploid cotton, although there is no evidence for genome-wide expression dominance in this species (Zhang et al. 2015). Gene expression levels are one of the best predictors of patterns of molecular evolution in most organisms, where more highly expressed genes evolve slower and lower expressed genes evolve faster due to stronger selective constraints on highly expressed genes (Pál et al. 2001; Yang and Gaut 2011; Zhang and Yang 2015). Our results are consistent with these theoretical and empirical observations.

Conclusion

Based on genes retained in all the merged parental subgenomes of Brassiceae, we have inferred the phylogenetic position of the brassica WGT. Our phylogenetic results are consistent with the previous two-step model of genome merger, but we do not exclude the alternative hypothesis of differential methylation patterns in progenitors. Furthermore, we confirmed that the mesotetraploid *Orychophragmus violaceus* has experienced an independent and probably more recent WGD. Assymetrical evolution between subgenomes has been revealed in all investigated Brassiceae species, where the subgenomes MFs underwent a relaxed purifying selection pressure. The highly expressed genes, from the LF subgenome, kept a strong purifying selection pressure due to strong constrainst in amino-acid substitution, as it was previously reported in brassica. The differential gene losses between Brassiceae species reported here suggest that gene losses following a WGD continue throughout the evolution of the lineage.

Supplementary Figures.

Fig. S1. Expected topologies of homoeologous trees supporting a shared WGD between *O. violaceus* and the Brassiceae tribe. Three topologies could be observed in a case of shared WGD (A, B and C), with each of the two *O. violaceus* copies being sister to a different Brassiceae parental lineage. Any different topologies would not support the hypothesis of a shared WGD and would be considered supporting an independent WGD in *O. violaceus*. Only one individual of *O. violaceus* is represented for easier comprehension.

Fig. S2. Frequency distribution of *Ks* values calculated between paralogous/homoeologous genes for five Brassiceae species (one per clade) and *O. violaceus*. The three coloured lines correspond to individual normal components extracted using mixture model analyses and are used to infer independent WGD (or WGT) events. Red, green and blue lines: more recent, intermediate and more ancient large-scale duplication events, respectively. X-axis: Ks values, Y-axis: number of paralogous gene pairs.

Supplementary Tables.

Table S1. Statistical tests (Likelihood Ratio tests) of ω ratio variation among parental sub-genomes using gene copies triplets for all the studied species belonging to the clades Oleracea and Nigra and for all the studied Brassiceae species. Each of the four tested models includes (analysis "Included") or excludes (analysis "Excluded") the ancestral branch of parental lineage in the calculation of the lineage specific ω ratio (see Fig. 3).

Fig. S1.

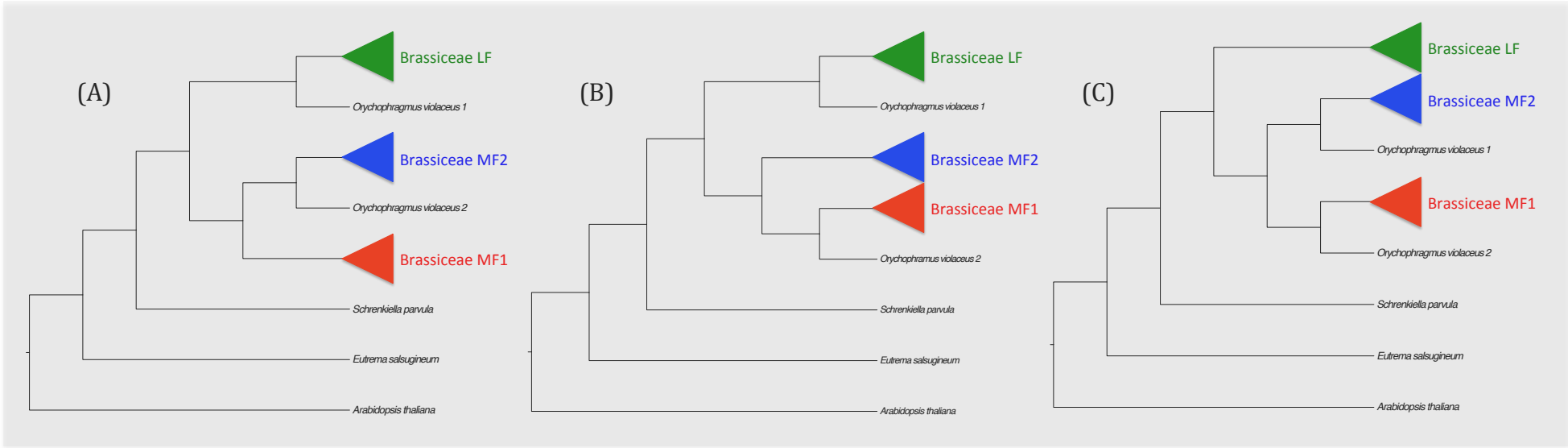
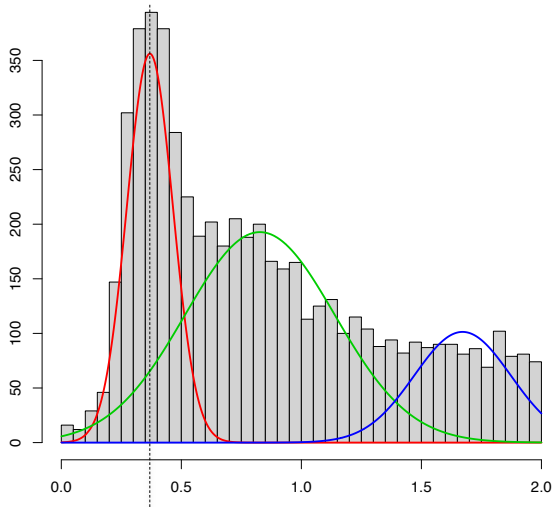
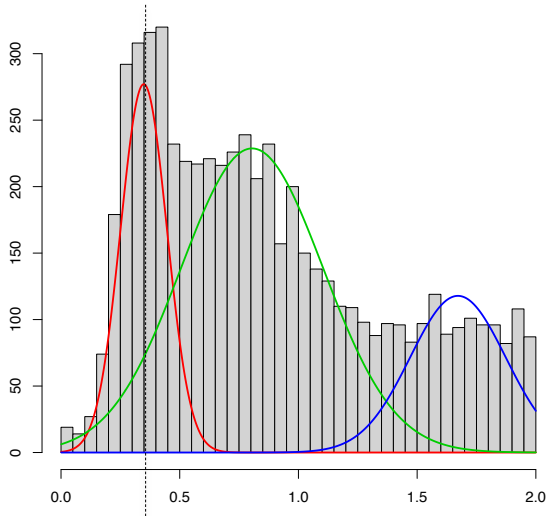

(A)
Brassiceae LF
Orychophragmus violaceus 1
Brassiceae MF2
Orychophragmus violaceus 2
Brassiceae MF1
Schrenkiella parvula
Eutrema salsugineum
Arabidopsis thaliana
(B)
Brassiceae LF
Orychophragmus violaceus 1
Brassiceae MF2
Brassiceae MF1
Orychophragmus violaceus 2
Schrenkiella parvula
Eutrema salsugineum
Arabidopsis thaliana
(C)
Brassiceae LF
Brassiceae MF2
Orychophragmus violaceus 1
Brassiceae MF1
Orychophragmus violaceus 2
Schrenkiella parvula
Eutrema salsugineum
Arabidopsis thaliana

Fig. S2.

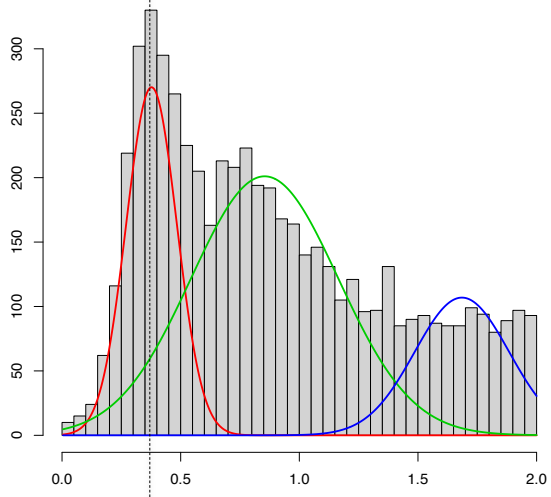
Cakile maritima



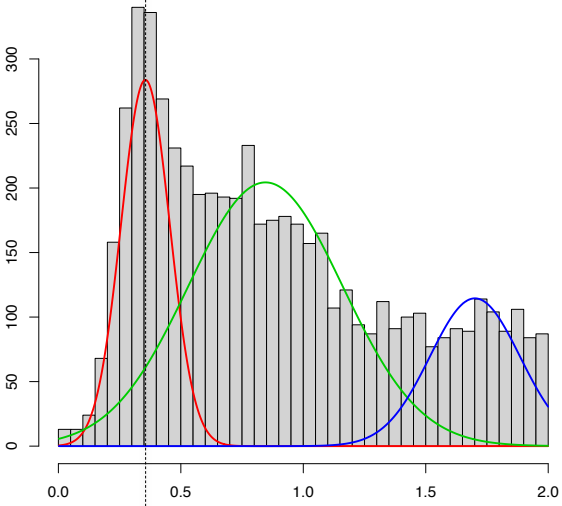
Schouwia purpurea



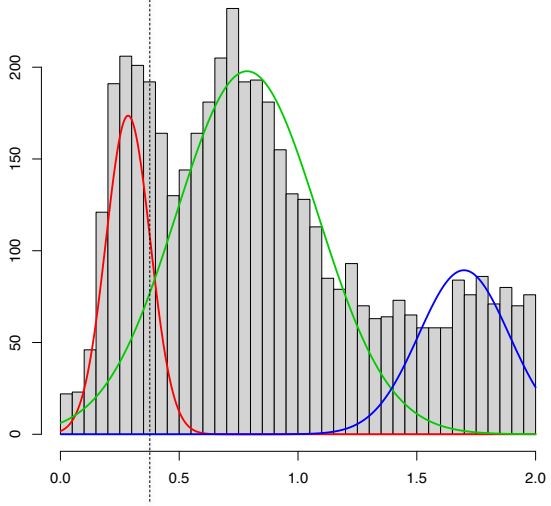
Psychine stylosa



Crambe maritima



Orychophragmus violaceus



Carrichtera annua

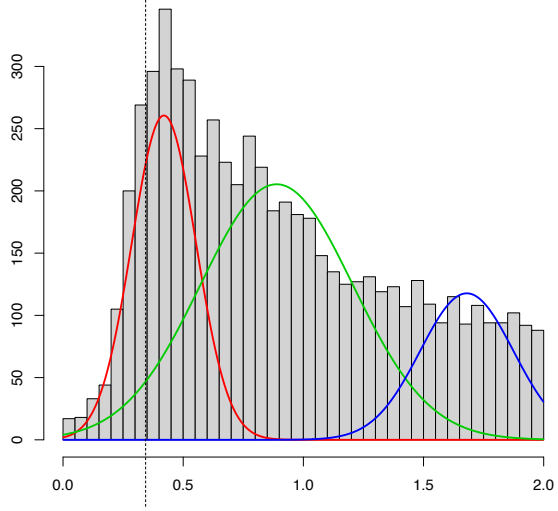


Table S1.

Analysis	Alignment	Models (null/alternative)	χ^2	ddl	P-value*
Excluded	Oleracea & Nigra[a]	One rate / LF, MF1 and MF2	8.919	1	**0.003** **
		LF, MF1 and MF2 / LF versus MF1 and MF2	50.13	1	**0.000** ***
		LF versus MF1 and MF2/ LF versus MF1 versus MF2	60.58	1	**0.000** ***
	all Brassiceae[b]	One rate / LF, MF1 and MF2	36.17	1	**0.000** ***
		LF, MF1 and MF2 / LF versus MF1 and MF2	85.10	1	**0.000** ***
		LF versus MF1 and MF2/ LF versus MF1 versus MF2	38.82	1	**0.000** ***
Included	Oleracea & Nigra[a]	One rate / LF, MF1 and MF2	294.3	1	**0.000** ***
		LF, MF1 and MF2 / LF versus MF1 and MF2	76.43	1	**0.000** ***
		LF versus MF1 and MF2/ LF versus MF1 versus MF2	44.44	1	**0.000** ***
	all Brassiceae[b]	One rate / LF, MF1 and MF2	433.7	1	**0.000** ***
		LF, MF1 and MF2 / LF versus MF1 and MF2	98.15	1	**0.000** ***
		LF versus MF1 and MF2/ LF versus MF1 versus MF2	38.70	1	**0.000** ***

*Probability of Type I error (the rejection of the null hypothesis that is true)

References.

Al-Shehbaz IA, Guang Y. 2000. A revision of the chinese endemic Orychophragmus (Brassicaceae). Novon 10:349–353.

Altschul SF, Gish W, Miller W, Myers EW, Lipmanl DJ. 1990. Basic Local Alignment Search Tool. J. Mol. Evol. 215:403–410.

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. Curr. Opin. Plant Biol. 15:140–146.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2016. On the relative abundance of autopolyploids and allopolyploids. New Phytol. 210:391–398.

Barker MS, Vogel H, Schranz ME. 2009. Paleopolyploidy in the Brassicales : Analyses of the Cleome Transcriptome Elucidate the History of Genome Duplications in Arabidopsis and Other Brassicales. Genome Biol. Evol. 1:391–399.

Benaglia T, Chauveau D, Hunter DR, Young DS, Chauveau D, Young DS. 2009. mixtools : An R package for analyzing finite mixture models. J. Stat. Softw. 32:1–29.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis : from classical genetics to modern genomics. Plant Cell 19:395–402.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. 13:137–144.

Blanc G, Wolfe KH. 2004. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. Plant Cell 16:1667–1678.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438.

Chang PL, Dilkes BP, Mcmahon M, Comai L, Nuzhdin S V. 2010. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. 11:1–17.

Chen ZJ. 2007. Genetic and epigentic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu. Rev. Plant Biol. 58:377–406.

Cheng F, Sun C, Wu J, Schnable J, Woodhouse MR, Liang J, Cai C, Freeling M, Wang X. 2016. Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. New Phytol. 211:288–299.

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One 7:1–9.

Cheng F, Wu J, Wang X. 2014. Genome triplication drove the diversification of Brassica plants. Hortic. Res. 24:1–8.

Cheng F, Wu J, Xie Q, Lysak MA, Wang X. 2013. Deciphering the diploid ancestral genome of the mesohexaploid Brassica rapa. Plant Cell 25:1541–1554.

Comai L. 2005. The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6:836–846.

Conant GC, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. 13:2052–2058.

Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Hazzouri KM, Wang W, Platts AE, et al. 2015. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. PNAS 112:2806–2811.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proc. Natl. Acad. Sci. U. S. A. 112:8362–8366.

Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-W, Bewick AJ, Ji L, Platts AE, Bowman MJ, et al. 2017. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell 29:2150–2167.

Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K. 2011. Cabbage family affairs: The evolutionary history of Brassicaceae. Trends Plant Sci. 16:108–116.

Geiser C, Mandáková T, Arrigo N, Lysak MA, Parisod C. 2016. Repeated Whole-Genome Duplication, Karyotype Reshuffling, and Biased Retention of Stress-Responding Genes in Buckler Mustard. Plant Cell 28:17–27.

Glover NM, Redestig H, Dessimoz C. 2016. Homoeologs : What are they and how do we infer them ? Trends Plant Sci. 21:609–621.

Gómez-Campo C. 1980. Morphology and morpho-taxonomy of the tribe Brassiceae. In: Tsunoda, S., Hinata, K. & Gómez-Campo C, editors. Brassica crops and the wild allies: Biology and breeding. Tokyo: Japan Scientific Societies Press. p. 3–31.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Chen Z, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–654.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90, 000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. 45:891–898.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements : A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19:1419–1428.

Hu H, Al-shehbaz IA, Sun Y, Hao G, Wang Q, Liu J. 2015. Species delimitation in Orychophragmus (Brassicaceae) based on chloroplast and nuclear DNA barcodes. Taxon 64:714–726.

Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-shehbaz I, Edger PP, et al. 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol. Biol. Evol. 33:394–412.

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE : a python environment for tree exploration. BMC Bioinformatics 11:1–7.

Huerta-Cepas J, Gabaldón T. 2011. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics 27:38–45.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3 : Reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33:1635–1638.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali S, Landherr L, Ralph PE, Jiao Y, Wickett NJ, Ayyampalayam S. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–102.

Kagale S, Robinson SJ, Nixon J, Xiao R, Huebert T, Condie J, Kessler D, Clarke WE, Edger PP, Links MG, et al. 2014. Polyploid evolution of the Brassicaceae during the Cenozoic Era. Plant Cell 26:2777–2791.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2 : New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772–773.

Leach LJ, Belfield EJ, Jiang C, Brown C, Mithani A, Harberd NP. 2014. Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. BMC Genomics 15:1–19.

Lee H-S, Chen ZJ. 2001. Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. PNAS 98:6753–6758.

Levin DA. 2013. The timetable for allopolyploidy in flowering plants. Ann. Bot. 112:1201–1208.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. Sci. Adv. 1:e1501084.

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5:1–11.

Lynch M, Conery S. 2000. The Evolutionary Fate and Consequences of Duplicate Genes.

Lysak MA, Cheung K, Kitschke M, Bures P. 2007. Ancestral Chromosomal Blocks Are Triplicated in Brassiceae Species with Varying Chromosome Number and Genome Size. Plant Physiol. 145:402–410.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. Genome Res.15:516–525.

Mandáková T, Joly S, Krzywinski M, Mummenhoff K, Lysak M a. 2010. Fast diploidization in close mesopolyploid relatives of Arabidopsis. Plant Cell 22:2277–2290.

Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. 91:3–21.

Marcet-Houben M, Gabaldón T. 2015. Beyond the whole-genome duplication : phylogenetic evidence for an ancient interspecies hybridization in the Baker's Yeast lineage. PLoS Biol. 7:1–26.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science 345.

Marhold K, Lihová J. 2006. Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Syst. Evol. 259:143–174.

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S. 2014. Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish Raphanus raphanistrum and Three Other Brassicaceae Species. Plant Cell 26:1925–1937.

Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, Crollius HR, Salse J. 2015. Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol. 16:1–17.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annu. Rev. Genet. 34:401–437.

Otto SP. 2007. The Evolutionary Consequences of Polyploidy. Cell 131:452–462.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.

Panchy N, Lehti-Shiu M, Shiu S. 2016. Evolution of Gene Duplication in Plants. Plant Physiol. 171:2294–2316.

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M-A, Ainouche M. 2009. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol. 184:1003–1015.

Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ. 2005. Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. Genetics 171:765–781.

Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, et al. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. Genome 15:R77.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nat. Rev. Genet. 18:411–424.

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Mayer KFX, Olsen O-A. 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. Science (80-.). 345.

Price MN, Dehal PS, Arkin AP. 2009. FastTree : Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Mol. Biol. Evol. 26:1641–1650.

Ramsey J, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst. 29:467–501.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE : Multiple Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. PLoS One 6:e22594.

Renny-Byfield S, Gong L, Gallagher JP, Wendel JF. 2015. Persistence of subgenomes in paleopolyploid cotton after 60 My of evolution. Mol. Biol. Evol. 32:1063–1071.

Ronquist FR, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2 : Efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Salmon A, Ainouche M, Wendel JF. 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). Mol. Ecol. 14:1163–1175.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. PNAS 108:4069–4074.

Schranz ME, Lysak MA, Mitchell-Olds T. 2006. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. 11:535–542.

Schranz ME, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell 18:1152–1165.

Scienski K, Fay JC, Conant GC. 2015. Patterns of gene conversion in duplicated yeast histones. Genome Biol. Evol. 7:3249–3258.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinforma. Bioinforma. 6:31.

Song Q, Chen ZJ. 2015. Epigenetic and developmental regulation in plant polyploids. Curr. Opin. Plant Biol. 24:101–109.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57:758–771.

Stamatakis A. 2006. RAxML-VI-HPC : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stamatakis A. 2014. RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC. 2012. Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics 190:1563–1574.

Tayalé A, Parisod C. 2013. Natural pathways to polyploidy in plants and consequences for genome reorganization. Cytogenet. Genome Res. 140:79–96.

Vallejo-Marín M, Buggs RJA, Cooley AM, Puzey JR. 2015. Speciation by genome duplication : Repeated origins and genomic composition of the recently formed allopolyploid species Mimulus peregrinus. Evolution 69:1487–1500.

Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. Mol. Biol. Evol. 30:177–190.

Vision TJ, Brown DG, Tanksley SD. 2000. The Origins of Genomic Duplications in Arabidopsis. Science 290:2114–2117.

Wang X, Paterson AH. 2011. Gene conversion in angiosperm genomes with an emphasis on genes duplicated by polyploidization. Genes (Basel). 2:1–20.

Wang X, Tang H, Bowers JE, Paterson AH. 2009. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. Genome Res. 19:1026–1032.

Wang X, Tang H, Paterson AH. 2011b. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major poaceae lineages. Plant Cell 23:27–37.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011a. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43:1035–1040.

Warwick SI, Al-Shehbaz IA, Price RA, Sauder C. 2002. Phylogeny of Sisymbrium (Brassicaceae) based on ITS sequences of nuclear ribosomal DNA. Can. J. Bot. 80:1002–1017.

Warwick SI, Al-Shehbaz IA. 2003. Nomenclatural notes on Sisymbrium (Brassicaceae). Novon 13:265–267.

Warwick SI, Al-Shehbaz IA. 2006. Brassicaceae: Chromosome number index and database on CD-Rom. Plant Syst. Evol. 259:237–248.

Wendel JF. 2000. Genome evolution in polyploids. Plant Mol. Biol. 42:225–249.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. PNAS 106:13875–13879.

Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. PNAS 111:5283–5288.

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M. 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biol. 8:1–15.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among Arabidopsis genes. Mol. Biol. Evol. 28:2359–2369.

Yang Z. 2007. PAML 4 : Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591.

Yoo M-J, Szadkowski E, Wendel JF. 2013. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity 110:171–180.

Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. Nat. Rev. Genet. 16:409–420.

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, et al. 2015. Sequencing of allotetraploid cotton (Gossypium hirsutum L . acc . TM-1) provides a resource for fiber improvement. Nat. Biotechnol. 33:531–540.

Ziolkowski PA, Kaczmarek M, Babula D, Sadowski J. 2006. Genome evolution in Arabidopsis/Brassica: conservation and divergence of ancient rearranged segments and their breakpoints. Plant J. 47:63–74.

CHAPITRE III

Evolution du locus d'auto-incompatibilité dans la tribu allo-hexaploïde des Brassiceae (Brassicaceae).

Une perte drastique de diversité phylogénétique au locus S a été observée dans les genres Brassica et Raphanus (Lim et al. 2002; Sato et al. 2002; Edh et al. 2009). La richesse allélique à ce locus reste forte mais tous les allèles se regroupent dans deux clades (les allèles de classe I et les allèles de classe II, voir section 3.3.2). Par opposition, la diversité phylogénétique au locus S chez Arabidopsis est très élevée, ce qui suggère qu'un goulot d'étranglement majeur aurait eu lieu chez l'ancêtre commun des genres Brassica et Raphanus, suivi d'une re-diversification allélique sous l'effet de la sélection fréquence-dépendante négative agissant sur le système SI. De plus, la localisation génomique du locus S chez *Brassica rapa* (Brassiceae, clade Oleracea) est distincte de la position génomique ancestrale observée chez *Arabidopsis sp.* (voir section 3.4.4 et FIGURE 15). Ces observations faites au locus S suggèrent un scénario évolutif original : cette perte de diversité phylogénétique au locus d'auto-incompatibilité et le changement de localisation génomique du locus S pourraient être liés à l'évènement de triplication du génome qui pourrait avoir touché l'ancêtre commun de la tribu des Brassiceae (Vekemans et al. 2014). L'objet de ce troisième chapitre de thèse est ainsi de vérifier si les différents clades de Brassiceae partagent les mêmes signatures de goulot d'étranglement et de changement de localisation génomique du locus-S que celles observées chez *Brassica rapa*. Plus précisément, nous avons cherché à caractériser (1) la localisation génomique du locus-S dans les génomes publiés de *Brassica rapa*, *B. oleracea* et *Raphanus sativus*, en lien avec la localisation génomique des sous-génomes parentaux issus de l'évènement WGT et (2) la diversité allélique du gène *SRK* chez les Brassiceae ainsi que les relations phylogénétiques entre allèles S, de manière à tester l'hypothèse d'un partage du goulot d'étranglement identifié précédemment chez Brassica. Ces analyses se sont portées sur des données à la fois génomiques et transcriptomiques.

Références

Edh K, Widen B, Ceplitis A. 2009. Molecular population genetics of the SRK and SCR Self-Incompatibility genes in the wild plant species Brassica cretica (Brassicaceae). Genetics 181:985–995.

Lim S, Cho H, Lee S, Cho Y, Kim B. 2002. Identification and classification of S haplotypes in Raphanus sativus by PCR-RFLP of the S locus glycoprotein (SLG) gene and the S locus receptor kinase (SRK) gene. Theor 104:1253–1262.

Sato K, Nishio T, Kimura R, Kusaba M, Suzuki T, Hatakeyama K, Ockendon DJ, Satta Y. 2002. Coevolution of the S-Locus genes SRK, SLG and SP11/SCR in Brassica oleracea and B. rapa. Genetics 162:931–940.

Vekemans X, Poux C, Goubet PM, Castric V. 2014. The evolution of selfing from outcrossing ancestors in Brassicaceae: what have we learned from variation at the S-locus? J. Evol. Biol. 27:1372–1385.

Title: Testing for an association of whole genome triplication, loss of diversity and genomic translocation of the self-incompatibility genes in the tribe Brassiceae (Brassicaceae).

Authors

Hénocq L.[1], Genete M.[1], Castric V.[1], Vekemans X.[1], Poux C.[1]

Affiliations

[1]Univ. Lille, CNRS, UMR 8198 – Evo-Eco-Paleo, Lille, France;

Abstract

Whole genome duplication events are common in flowering plants and especially within the Brassicaceae family. For example, the common ancestor of the Brassiceae tribe has experienced two successive events of allopolyploidy, generating a whole genome triplication. Ancient polyploidy events are generally followed by a diploidization process involving genetic and epigenetic changes, as well as structural changes leading to a diploid genome. Furthermore, the dynamic of transposable elements is disturbed, which can lead to an increase in translocation events. In one lineage of Brassiceae, a drastic loss of phylogenetic diversity, with all S-alleles that clustered into only two distinct phylogenetic clades (class I and class II S-alleles), and a genomic translocation have been observed at the self-incompatibility locus (S locus). We suspect that these patterns are associated with the allopolyploidy events. By analysing the S locus diversity and the genomic organization of the S locus among Brassiceae clades, we aim at determining (1) whether the bottleneck observed at the S-locus in Brassica and Raphanus is shared between other members of the Brassiceae tribe; (2) whether this bottleneck is contemporaneous with the whole genome triplication event that occurred in the ancestor of the tribe and (2) whether these events are also associated with the translocation of the S-locus. We also aim at determining in which extent the genomic organization is conserved among S-haplotypes.

Introduction

The Brassicaceae self-incompatibility locus (S locus) is involved in a self-recognition system that prevents self-fertilization (Takayama and Isogai 2005; Nasrallah 2017). This system requires two genes – one for the female (pistil, *SRK*) side and the other for the male (pollen, *SCR*) side – that interact as a key lock system, and allele-specific interaction between male and female molecules prevents self-fertilization to avoid inbreeding depression. In Brassica and Raphanus, there is a paralogous sequence of the S-domain of *SRK*, called *SLG*, that is located in the S-locus but its role is unclear (Suzuki et al. 2000; Takasaki et al. 2000). Because of the frequency-dependant selection that occurred at the S-locus, a large number of S-alleles is found within populations at the both components of this system (Wright 1939). Indeed, SI species of Arabidopsis and Capsella have SRK alleles distributed into many lineages, with highly diverged sequences. Trans-specific and even trans-generic polymorphism between several self-incompatible Brassicaceae species suggests that the diversification at the S-locus has occurred before the diversification of the family.

In self-incompatible species of Brassica and Raphanus that have experienced a whole genome triplication event (Wang et al. 2011; Liu et al. 2014; Moghe et al. 2014), previous phylogenies of the SI-related genes (*SRK*, *SLG*) have indicated that S-alleles cluster into only two clades of sequences (class I and class II S-alleles) and that the diversification of the S-alleles occurred prior to the divergence of the two genera (Lim et al. 2002; Sato et al. 2002). The same S-haplotypes are distributed in different Brassica species (Kusaba et al. 1997; Sato et al. 2002), which are called trans-specific S-haplotypes. The clustering of the S-alleles into two phylogenetic clades in Brassica and Raphanus suggests that a strong genetic bottleneck probably occurred in the ancestor of Brassica and Raphanus (Castric and Vekemans 2007; Edh et al. 2009), but the examination of the S-alleles diversity in other members of the Brassiceae tribe, particularly in other Brassiceae clades, has not yet being performed. A bottleneck in population size associated with the founding of new polyploid lineages (Otto 2007), generating a strong genetic bottleneck at the genome scale, should dramatically reduce the number of S-alleles (Ramsey and Schmeske 1998). Here, we hypothesized that the bottleneck observed on the phylogeny of the S-alleles of Brassica and Raphanus could be associated with the WGT event experienced by the Brassiceae tribe. In this case, we should find only class-I and class-II S-alleles in the entire Brassiceae tribe. Moreover, the S-locus occupies different chromosomal regions in *Brassica rapa* and in Arabidopsis spp. (Fobis-Loisy and Gaude 2004; Chantha et al. 2013; Kim et al. 2016), suggesting that a genomic

translocation occurred in Brassica, but investigations in other Brassiceae clades are still lacking. Once more, the genomic translocation of the S-locus could be associated with the WGT event since translocation events may be frequent in polyploids due to the disturbance of the dynamic of transposable elements in polyploids (Parisod et al. 2010; Vicient and Casacuberta 2017).

In the present study, we thus aim at determining (1) whether the bottleneck observed at the S-locus in Brassica and Raphanus is shared between other members of the Brassiceae tribe; (2) whether this bottleneck is contemporaneous with the whole genome triplication event that occurred in the ancestor of the tribe and (2) whether these events are also associated with the translocation of the S-locus. Because a previous study revealed different genomic organization between S-haplotypes in Brassiceae (Kim et al. 2016), we also aim at determining in which extent the genomic organization of the class-I and class-II S-haplotypes is conserved among several Brassiceae species.

Material & Methods

Structural gene organization in the Arabidopsis and Brassica S-locus regions within Brassica rapa and Brassica oleracea

Comparative genetic mapping of the S-locus in Brassica and *Arabidopsis lyrata* has shown that the S-locus occupies different chromosomal regions in Brassica spp. and in Arabidopsis spp. (Fobis-Loisy and Gaude 2004). In *Arabidopsis thaliana*, the S-locus is located between the *B80*/*U-Box* gene (*At4g21350*) flanking the S-locus on one side, and the *ARK3* gene (*At4g21380*) flanking the S-locus on the other side (Goubet et al. 2012), in the ancestral genomic Block U on chromosome IV (Schranz et al. 2006), which is recognized as the ancestral S-locus location (Chantha et al. 2013; Vekemans et al. 2014). In Brassica, the S-locus is located on the *B. rapa* A07 chromosome (Kim et al. 2016, Fig. 1), in a region that is homologous to a chromosomal region on chromosome I of *A. thaliana* belonging to the ancestral genomic Block E.

We compared the structural gene organization between *A. thaliana*, *B. rapa* and *B. oleracea* genomes in the two genomic regions corresponding to both the "Arabidopsis" S-locus region, and the "Brassica" S-locus. For the “Arabidopsis” S-locus region, we considered the region located between genes *AT4G21110* and *AT4G21440* that are framing the S-locus, on chromosome IV of *A. thaliana* (Table S1). We extracted the three homoeologous regions (within subgenomes LF, MF1 and MF2) that are homologous to the “Arabidopsis” S-locus region in the genome of *B. rapa* using the search tool « syntenic gene » of the BRAD

database (Cheng et al. 2011). The two flanking genes of each of the three homoeologous regions of *B. rapa* were mapped onto the genome of B*rassica oleracea* (To1000 cultivar, Parkin et al. 2014) using the blast search tool of Ensembl (Zerbino et al. 2018) in order to localize and extract the three homoeologous regions in this species (Table S1). For the "Brassica" S-locus in *B. rapa* genome, according to Kim et al. (2016), the sequenced Chiifu cultivar contains the S-60 haplotype which is located on chromosome A07 (Wang et al. 2011). By blast search using the BRAD database (Cheng et al. 2011) and sequences of *B. rapa SRK-60* and *SLG-60* (Genbank AB097116.1) we identified that these corresponded to the annotated genes *Bra004179* (*SRK* exons 1, 2 and 3), *Bra004180* (*SRK* exons 4, 5, 6 and 7) and *Bra004182* (*SLG*) (Table S2). In this genome, the S-locus region is located in the ancestral genomic Block E within subgenome LF (Wang et al. 2011, Cheng et al. 2011) between the two flanking genes *Sll2* (*AT166680*) and *AtPPa* (*AT1G66690*) (Kim et al. 2016). We could identify a homoeologous region containing a few of the S-locus neighbouring genes within the MF1 subgenome located on chromosome A02, but no homoeologous region located on the MF2 subgenome could be detected (Table S1). We selected two genes (*Bra004159*, *Bra004203*) framing the "Brassica" S-locus region at a sufficient distance to correctly characterize the synteny between homoeologous genes in the LF and MF1 subgenomes (Table S1). These two flanking genes were mapped onto the genome of *B. oleracea* (Parkin et al. 2014) in order to extract the S-locus region in this species. As in *B. rapa*, two homoeologous regions, one on *B. oleracea* chromosome C06 (LF subgenome) and the other on chromosome C02 (MF1 subgenome), were identified (Table S1). The *SRK* gene (corresponding to *B. oleracea S13* haplotype) was located on the ancestral block E region of chromosome C06 belonging to the LF subgenome. The homologous genomic sequence of *A. thaliana* has been extracted using the genome browser of the BRAD database (Cheng et al. 2011). For each of the two genomic locations ("Arabidopsis" and "Brassica" S-loci), the full genomic sequences of *B. rapa*, *B. oleracea* and *A. thaliana* were aligned and compared using the global pairwise and multiple sequence alignment procedure (Brudno et al. 2003) implemented in mVISTA (Dubchak 2007).

Genomic location of the S-locus region within the Brassiceae tribe

In order to determine if the genomic translocation of the S-locus region found in *B. rapa* was specific to this species, to the Brassiceae clade Oleracea, or to the entire Brassiceae tribe, we examined published genome sequences of additional Brassiceae species. For other species without sequenced genomes, we designed primers in conserved regions of the *ARK3*

and the *U-Box* genes flanking the Arabidopsis S-locus region to amplify the intergenic region (Table 1) and therefore test for the presence of the S-locus at the "Arabidopsis" position. Using the BRAD database (Cheng et al. 2011), we localized the two genes *ARK3* and *B80/U-Box* in *B. rapa* and extracted the DNA sequence between the end codon of the *B80/U-Box* gene (in reverse position, Fig. S1) and the start codon of the *ARK3* gene (in reverse position) ([A01:6,112,142..6,118,190]). We performed the same procedure for three other published genomes of Brassiceae species available in the BRAD and Ensembl databases: *Brassica oleracea* ([C01: 8,793,569..8,799,620]), *Brassica nigra* ([B2:36,699,723..36,706,368]) and *Raphanus sativus* ([Rsa1.0_00770.1:32,777..39,840]). The four genomic sequences were aligned using MUSCLE (Edgar, R.C., 2004) using the default strategy, and the size of the intergenic region was computed for each species.

As the entire S-locus has been deleted in the Arabidopsis S-locus region in the two Brassica genomes *B. rapa* and *B. oleracea*, *ARK3* and the *U-Box* have become neighbouring genes, with an intergenic region smaller than about 1,25 Kb (see results). If the S-locus region in another Brassiceae species is located at the same genomic position than in Arabidopsis, no amplification should occur between the *ARK3* and the *U-Box* primers, because of the large size of the S-locus. On the contrary, we should expect amplification of the *ARK3* - *U-Box* intergenic region if the species shares the S-locus region translocation with *Brassica*. For this purpose, the *Brassica rapa* sequences of the *B80/U-Box* and *ARK3* genes (Wang et al. 2011, Cheng et al. 2011) were mapped onto each transcriptome assembly described in the chapter 2. The best-hit contig of each individual assembly was then extracted and aligned separately for each of the two genes using MUSCLE (Edgar 2004) with the default strategy (see Table 1 for the list of species used for the design of each primer). Primers were designed by hand in conserved regions of the two genes and checked with the software Oligo Analyzer version 3.1 (PrimerQuest® program, IDT, Coralville, Iowa, USA. Accessed 12 December, 2018. http://www.idtdna.com/SciTools). The *ARK3*-Forward primer was designed at the end of the gene (in forward orientation) whereas the *B80/U-Box*-Reverse primer was designed at the beginning of the gene (in reverse orientation), given the gene organization within the S-locus region in Arabidopsis (see Fig. S1). In *B. rapa,* the region between the two primers has a size of 2,618 base pairs (bp) (517bp in the U-Box gene, 1,245bp in the intergenic region and 856bp in the ARK3 gene). Leaf material of an individual for each of the following species was sampled (see Table S1 in chapter 1 for provenance of seeds): *Cakile maritima* (Brassiceae, clade Cakile), *Schouwia purpurea* (Brassiceae, clade Zilla), *Psychine stylosa* (Brassiceae, clade Savignya), *Carrichtera annua* (Brassiceae, clade Vella) and

Orychophragmus violaceus (Brassicaceae, not belonging to the Brassiceae tribe). Leaf samples were dried and conserved in silica gel until DNA extraction. DNA was extracted from 15-20mg of leaf tissue using the Santoni extraction protocol, following the standard protocol outlined in the manufacturer's handbook. Two DNA samples of *Brassica insularis* (clade Oleracea) were also used for amplification. Intergenic sequences were obtained by polymerase chain reaction (PCR) amplification with primers *ARK3-F* and *U-Box-R* (Table 1). The amplification reactions contained 5µl of 5X Q5 Reaction Buffer, 0.5µl of dNTPs, 0.3µl of each primer, 0.25µl of Q5® Hot Start *Taq* High-Fidelity DNA Polymerase (BioLabs® *inc.*), 14,65 µl of H_2O and 4µl of DNA in a 25µl reaction volume. Amplifications were performed on a MasterCycler EpGradient Eppendorf thermocycler and the cycling scheme included 95° for 3 min, 35 cycles of 95° for 50 sec, 51.6° for 45 sec, 72° for 3 min, and, finally, 72° for 10 min. PCR products were purified using the NucleoSpin® Gel and PCR Clean-up kit (MACHEREY-NAGEL) with multiple elution steps to increase recovery. Samples were sequenced with the BigDye3.1 sequencing kit (Applied Biosystems) and loaded on a 3130 capillary sequencer. Sequences were edited and cleaned using CodonCode Aligner© version 5.1.5 (CodonCode Corporation, Dedham, Massachusetts). The four genomic sequences along with the sequenced PCR fragments were aligned and compared using the global pairwise and multiple sequence alignment procedure (Brudno et al. 2003) implemented in mVISTA (Dubchak 2007) to confirm the homology of the genomic region among species.

Table 1. Primers used for the amplification of the genomic region flanking by the two genes *B80/U-Box* and *ARK3*, and species for which sequences could be obtained from the individual transcriptome assemblies[T] or published genomes[G] in order to obtain consensus sequences and define the primers. Due to the orientation of the two genes and the gene order (see Fig. S1), the primer designed in the *ARK3* gene is forward whereas the primer designed in the *B80/U-Box* gene is reverse.

	ARK3- Forward (Position in the coding sequence of *B. rapa*)	*B80/U-Box*-Reverse (Position in the coding sequence of *B. rapa*)
Primers	5' – GGT ATT GCT AGA GGR CTT CTA TAT CT – 3' (1876-1901)	5'-CYA AAC CCA CTT TGT TAT CRT CTT C-3' (493-517)
Species used for the primer design	*Brassica rapa* [G], *Carrichterra annua* [T], *Schouwia purpurea* [T], *Orychophragmus violaceus* [T], *Arabidopsis thaliana* [G].	*Brassica rapa* [G], *Cakile maritima* [T], *Psychine stylosa* [T], *Carrichterra annua* [T], *Schouwia purpurea* [T], *Zilla spinosa subsp macroptera* [T], *Orychophragmus violaceus* [T], *Sisymbrium irio* [G], *Arabidopsis thaliana* [G].

Phylogenetic relationships among *SRK* sequences

Phylogenetic relationships between Arabidopsis and Brassica SRK sequences

Based on published as well as unpublished (some sequences from *A. halleri*) data, we compiled 152 *SRK* coding sequences from two SI species of the Arabidopsis genus (*Arabidopsis halleri*, *A. lyrata*) and two Brassica species (*Brassica rapa* and *B. oleracea*), as well as 5 *ARK3* (an *SRK* paralog) sequences from SI and SC Brassicaceae species (*A. halleri*, *A. lyrata*, *Eutrema salsugineum*, *Capsella rubella*) and *Cleome hassleriana* (Cleomaceae). The ARK sequences belonging to the same gene family as SRK were used as outgroup. We aligned those sequences with MACSE with default options (Ranwez et al. 2011). The resulting alignment was examined by eye in order to check its quality and was trimmed manually in order to keep only the S-domain of the *SRK* gene. The resulting framed alignment contained 1,518bp. Starting from a partitioning scheme containing 3 partitions (a separate partition for each codon position), the best partitioning schemes and evolutionary models were selected using the rcluster search mode implemented in PartitionFinder 2 (Lanfear et al. 2014; Lanfear et al. 2017), under the corrected Akaike Information Criterion (AICc). The resulting partitioning scheme contained 3 partitions (one for each codon position) with a GTR+Γ+I model and was used in the following analyses.

Phylogenetic reconstructions were performed by maximum likelihood (ML) using RAxML v.8.2.10 (Stamatakis 2006; Stamatakis et al. 2008; Stamatakis 2014). The number of bootstrap replicates to assess node supports was automatically determined (Stamatakis 2014). Bayesian analyses (BI) were performed using MrBayes v.3.2.6 (Ronquist et al. 2012) using a specific model for each DNA partition according to the best partition-scheme selected by PartitionFinder 2. Two independent runs with 10,000,000 generations were completed with one cold and three hot chains, starting from a random tree. Trees and parameters were sampled every 200 generations. Plots of the likelihood-by-generation evolution were drawn to check chains convergence. It was also supported by an average standard deviation of split frequencies between runs smaller than 0.01 and effective sample size (ESS) values reaching above 100. The first 25% of trees from both runs were discarded as burn-in. A majority-rule consensus of the remaining trees was performed to obtain the final tree with posterior probabilities (PP).

Diversity and phylogenetic positions of SRK sequences from Brassiceae species external to the Oleracea clade

In chapter 1, transcriptome sequencing via HiSeq technology was performed on five species of the Brassiceae tribe belonging to five well-recognized Brassiceae clades (*Carrichtera annua* from clade Vella; *Schouwia purpurea* and *Zilla spinosa* subsp. *macroptera* from clade Zilla; *Psychine stylosa* from clade Savignya, *Cakile maritima* from clade Cakile, and *Crambe maritima* from clade Crambe; Arias & Pires, 2012) and one outgroup species, *Orychophragmus violaceus*. RNAseq reads were obtained for 27 individuals, with 2 to 12 individuals per species (Table S1 in chapter 1). The origin of sampled individuals, the type of collected tissues and the method for obtaining RNAseq data and clean reads are given in chapter 1.

An exhaustive database containing all published *SRK* sequences (except those containing only the kinase domain) and several paralog sequences (*SLG*, *ARK3*, *SUI*, *SLR2*) of Brassiceae species (Table S3, including sequences from *B. rapa*, *B. oleracea*, *Raphanus raphanistrum* and *R. sativus*) was built in order to search for homologous *SRK* sequences from our RNAseq data. For this purpose we used an unpublished bioinformatics pipeline under development in the lab (NGSGenotyp developed by Mathieu Genete) that executes the following procedures: (1) mapping cleaned reads from each individual onto each reference sequence using Bowtie 2 version 2.2.6 (Langmead and Salzberg 2012); (2) computing alignment statistics against each reference sequence using SAMtools (Li and Durbin 2009); (3) assembling mapped reads using the *de novo* assembly procedure of the dipSPAdes assembler (v.3.10.1, with the following parameters –k 21, 41, 81 and –careful); (4) aligning the reconstructed contigs together with all reference sequences using the DNA local alignment tool YASS (v1.14) and MUSCLE (Edgar 2004); and (5) reconstructing a phylogenetic tree for each individual with PHYML (Guindon and Gascuel 2003) in order to identify which assembled contigs correspond to putative *SRK* sequences.

In order to compare the diversity and phylogenetic divergence of *SRK* sequences from the major clades of Brassiceae, we compiled the aforementioned published *SRK* and paralogous sequences from *B. rapa*, *B. oleracea*, *R. raphanistrum* and *R. sativus*, together with (i) our selected contigs from transcriptome data of Brassiceae species and *Orychophragmus violaceus*, (ii) four unpublished S-domain *SRK* sequences from *Sinapis arvensis* (“Nigra” clade) obtained previously in our lab (Fléchon et al. 2012), (iii) five *ARK3* sequences from Brassicaceae species, and finally (iv) a set of *A. halleri* and *A. lyrata* representative sequences from each major phylogenetic clade of *SRK* sequences. As this data

set contained sequences from NGS data and also sequences from putatively non-functional S-haplotypes, we performed the nucleotide alignment using MACSE with different stop codon and frameshift costs between the reference alignment and the dataset to be aligned, in order to account for potential sequencing errors in transcriptome contigs (Ranwez et al. 2011). More precisely, *SRK* and *SRK*-like sequences from RNAseq data were aligned using the Arabidopsis and Brassica *SRK* alignment (1,518bp) as backbone. The resulting alignment was examined by eye in order to check its quality and was manually corrected and trimmed in order to keep only the S-domain of the *SRK* or paralogous sequences. The resulting alignment contained 1,584bp and 228 sequences. We used PartitionFinder 2 as explained above to determine the best partitioning scheme and evolutionary models (Lanfear et al. 2014; Lanfear et al. 2017). Then, we performed the same previous phylogenetic analyses using the best partition scheme and models produced by PartitionFinder 2 (3 partitions with a GTR+Γ+I model). All contigs that had an ambiguous phylogenetic position were mapped onto the *B. rapa* genome using the BRAD database (Cheng et al. 2011, Wang et al. 2011) in order to check for potential unsampled paralogs.

Annotation of the S-locus genomic region in different S-haplotypes from Brassica and Raphanus

In analogy with Goubet et al. (2012) for the Arabidopsis S-locus, we define the "Brassica" S-locus as the region strictly embedded between the two flanking (but not functionally involved in self-incompatibility) genes *SLL2* (*Bra004178*) and *AtPPa* (*Bra004183*). In order to compare the organization of the S-locus and its flanking regions among different S-haplotypes, we looked for synteny in the extended S-locus genomic region lying between the start codon of the *SP1* gene (*Bra004174*) flanking the S-locus on one side, and the stop codon of the *SP7* gene (*Bra004186*) flanking the S-locus on the other side (Table S2). We compared previously reported entire sequences of the S-locus from three *B. rapa* S-haplotypes (S-8, S-46 and S-60, Table 2), with two additional sequences that we extracted from the two published *B. oleracea* genomes (Liu et al. 2014; Parkin et al. 2014) and one extracted from the *Raphanus sativus* genome (Kitashiba et al. 2014) according to Kim et al. (2016). For each compiled S-haplotype sequence (Table 2), we performed gene annotation by using BLASTN (Altschul et al. 1990) with the annotated genomic sequences of the *B. rapa* genome used as reference (Wang et al. 2011, Cheng et al. 2011 BRAD, Table S2). In addition, the sequences of the two S-haplotypes-linked small RNA (sRNA), *SMI* (*SP11 METHYLATION INDUCER*), present in both class I and class II S-haplotypes (Tarutani et al.

2010), and *SMI2* (*SP11 METHYLATION INDUCER 2*), present only in class II S-haplotypes (Yasuda et al. 2016), were mapped onto each S-locus region for the annotation. Because of its small size and its high nucleotide diversity, *SCR*/*SP11* might not be detected by using one or few *SCR*/*SP11* sequences. Therefore, a database of known *SCR*/*SP11* proteins from *B. rapa, B. oleracea* and *Raphanus sativus* was aligned on each S-locus region using BLASTN. We used an R custom script (R development Core Team, 2009) from Vekemans et al. (2014) to visually represent these annotations.

Table 2. Sequences of the S-locus region used in this study and their source. In order to obtain the sequences coming from the two genomic assemblies of *Brassica oleracea*, the reference sequence of the two genes flanking the S-locus region in *Brassica rapa*, Bra004174 and Bra004186 (see Table S2), were mapped onto each genomic assembly using tools available in the BRAD database (Cheng et al. 2011) and the Ensembl database (Zerbino et al. 2018). Then, the sequence between the two hits was extracted from the assemblies (see the column "genomic location").

Species	Haplotypes	Allelic class	Data type	No. accession / genomic location or scaffolds	Reference
Brassica rapa	BrS-60	II	genomic	[A07: 20,491,331..20,576,853]	Wang et al. (2011)
	BrS-46	I	cloning	AB257128	Takuno et al. (2007)
	BrS-8	I	cloning	AB257127	Takuno et al. (2007)
Brassica oleracea	unknown	unknown	genomic	Scaffold000336:401,645..423,013 Scaffold000196:1,570..186,424	Liu et al. (2014)
	unknown	unknown	genomic	[C06:32,017,843..32,143,043]	Parkin et al. (2014)
Raphanus sativus	S-19	I	genomic	Rsa1.0_01586.1 Rsa1.0_06919.1 Rsa1.0_07263 Rsa1.0_01510.1	Kitashiba et al. (2014) Kim et al. (2016)

Results

Structural gene organization in the Arabidopsis and Brassica S-locus regions within Brassica rapa and Brassica oleracea

The selected region of the "Arabidopsis" S-locus region and the identified homologous regions in each of the three subgenomes of *B. rapa* and *B. oleracea* were aligned and compared. In both Brassica species, homology was found in the flanking regions of the core "Arabidopsis" S-locus region in each of the three subgenomes, with substantially higher gene conservation in the LF subgenome than in the MF1 and MF2 subgenomes (Fig. 1), indicating that the extracted genomic regions in these two species are homologous to the selected region of the "Arabidopsis" S-locus region. Homology between *A. thaliana* and Brassica sequences was found to a large extent in the exonic regions, but also to some extent in intronic and some intergenic regions (Fig. 1). The MF2 region was inverted in the *B. rapa* and *B. oleracea* genomes (Table S1) and showed the lowest level of conservation compared to the LF and MF1 regions (Fig. 1, Table S1). However, we did not find any homology in the core "Arabidopsis" S-locus itself (between *B80/U-box* and *ARK3*) in either subgenome (Fig. 1), neither for *B. rapa* or *B. oleracea*, demonstrating that the S-locus was absent from the ancestral Arabidopsis position in all three subgenomes. Moreover, the two S-locus flanking genes on each side, *B80/U-box* and *ARK3*, were absent in MF1 and MF2.

In *B. rapa*, the S-locus is a single-copy region located within the LF subgenome on chromosome A07. We compared the structural gene organization in this region with that in *B. oleracea* (on chromosome C06) and *A. thaliana* (on chromosome 1), as well as with the homoeologous region in the MF1 subgenome of both Brassica species (chromosomes A02 and C02, respectively; note that, as stated above, the homoeologous region is missing in the MF2 subgenome). Our results show that the S-locus region is the same in both species *B. rapa* and *B. oleracea*, and is located on the ancestral genomic block E between the *Sll2* and *AtPPa* genes that are orthologous to the Arabidopsis genes *AT1G66680* and *AT1G66690*, respectively (Fig. 2, Table S2).

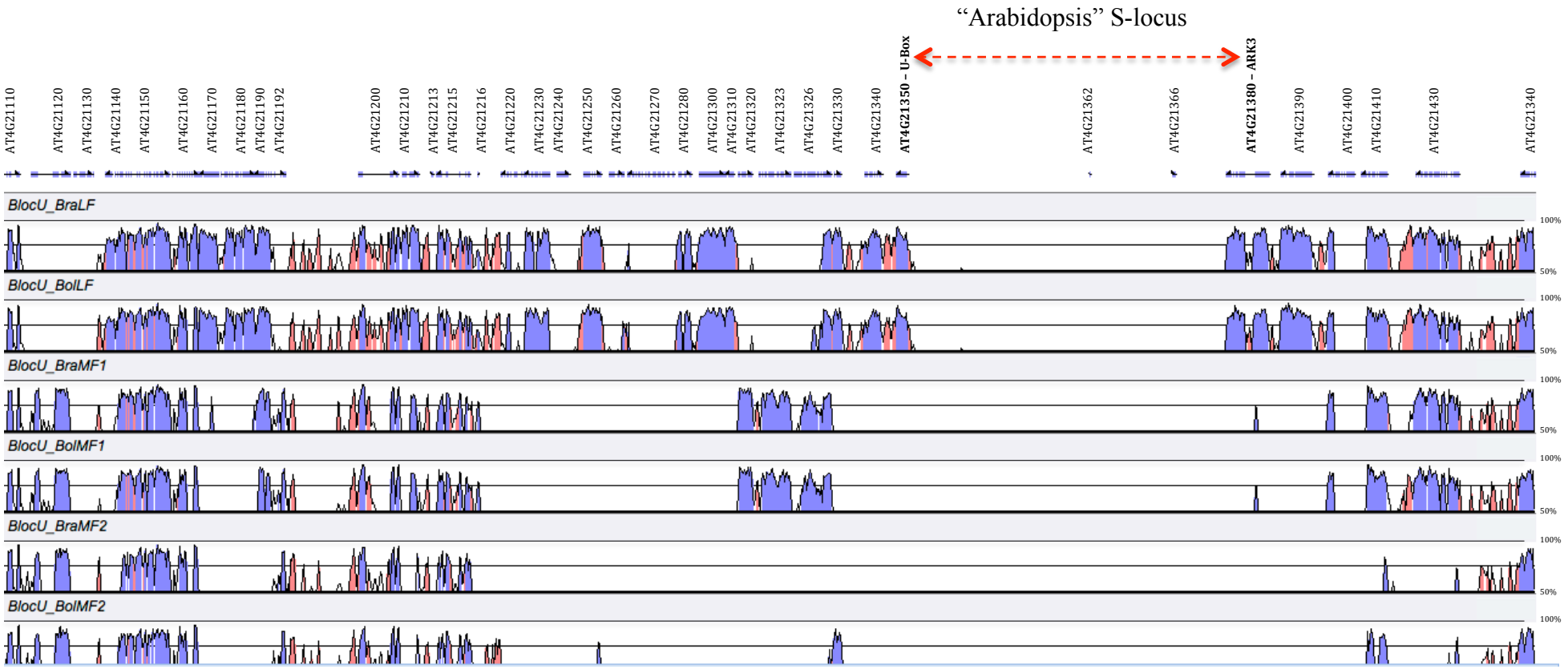


Fig. 1. Characterization of the genomic region homologous to the "Arabidopsis" S-locus region (see main text and Table S1 for details about the region) in each of the three subgenomes of *Brassica rapa* (BlockU_BraLF, BlockU_BraMF1 and BlockU_BraMF2) and *Brassica oleracea* (BlockU_BolLF, BlockU_BolMF1 and BlockU_BolMF2). The figure displays the VISTA alignment showing the level of sequence conservation (homology) in a selected region of the "Arabidopsis" S-locus region and the homologous regions in the *B. rapa* and *B. oleracea* genomes. The *A. thaliana* sequence was used as the reference (top). The core "Arabidopsis" S-locus region flanking by the two genes *U-Box* and *ARK3* (in bold) is indicated with a red arrow. Pink areas correspond to the intergenic and intronic regions whereas purple areas correspond to the exonic regions.

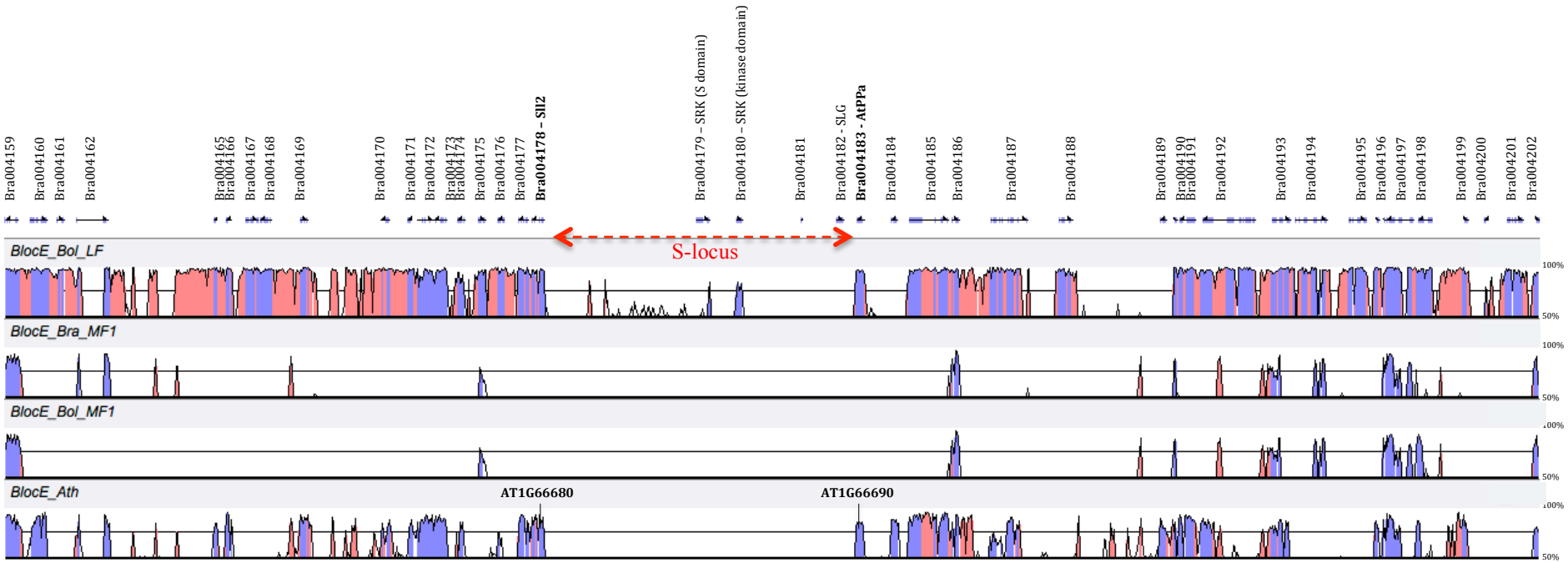


Fig. 2. Characterization of the genomic region homologous to the "Brassica" S-locus region (see main text and Table S1 for details about the region) in the MF1 subgenome of *Brassica rapa* (BlockE_Bra_MF1) and the LF and MF1 subgenomes of *Brassica oleracea* (BlockE_Bol_LF and Block E_Bol_MF1) and in the genome of *Arabidopsis thaliana* (BlockE_Ath). The figure displays the VISTA alignment showing the level of sequence conservation (homology) in a selected region of the "Brassica" S-locus region and the homologous regions in the *B. rapa*, *B. oleracea* and *A. thaliana* genomes. The *B. rapa* sequence from the LF subgenome was used as the reference (top). The core "Brassica" S-locus region flanking by the two genes Sll2 and AtPPa (in bold) is indicated with a red arrow. *A. thaliana* genes that are orthologous to these two genes are indicated. Pink areas correspond to the intergenic and intronic regions whereas purple areas correspond to the exonic regions.

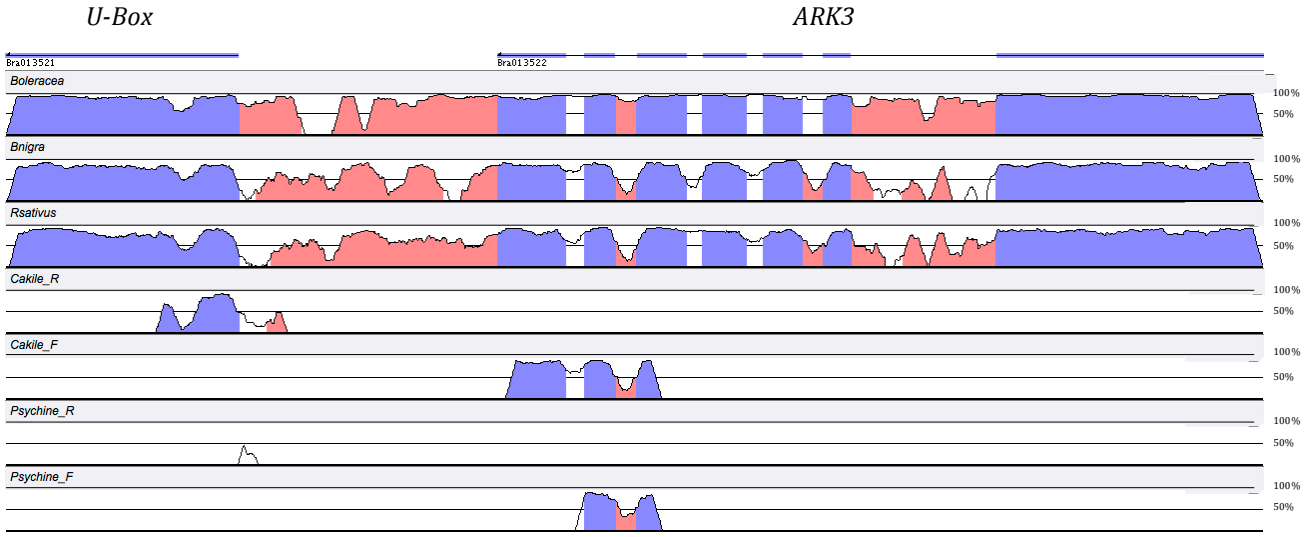
Genomic location of the S-locus region within the Brassiceae tribe

We could successfully amplify the intergenic region between the *B80/U-box* and *ARK3* only for two species, *Cakile maritima* (clade Cakile) and *Psychine stylosa* (clade Savignya), with a fragment size corresponding to those expected (~3kb, Fig. 3), whereas for all other species, PCR amplification failed at any tested temperature and whatever the elongation time. One forward and one reverse DNA fragments were sequenced for *C. maritima* (733 and 630bp, respectively) and *P. stylosa* (421 and 144bp, respectively). From published genomic data, the size of the genomic fragments between the end position of the *B80/U-Box* gene and the start position of the *ARK3* gene in *Brassica rapa*, *B. oleracea*, *B. nigra* and *Raphanus sativus* was estimated as 6,049bp, 6,052bp, 6,046bp and 7,064bp, respectively. The size of the genomic fragment between the start position of the *B80/U-Box* gene and the end position of the *ARK3* gene in these species was estimated as 1,245bp, 1,229bp, 1,382bp and 1,271bp, respectively.

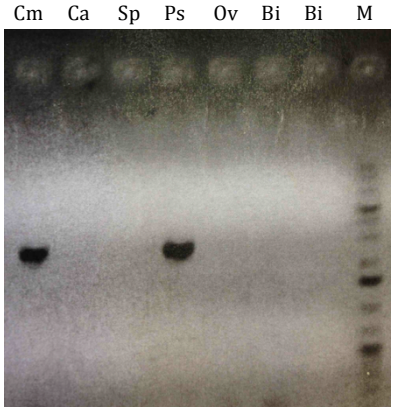
The genomic region that is located between the two genes *B80/U-Box* and *ARK3* displayed a high level of sequence conservation between *B. rapa* and the three species *B. oleracea* (clade Oleracea), *B. nigra* and *R. sativus* (clade Nigra), demonstrating that, as in *B. rapa*, the S-locus is not located at this genomic position in these three Brassiceae species (Fig. 3, A). The short fragments amplified in *C. maritima* and *P. stylosa* (Fig. 3, B) as well as the VISTA alignment confirm that the amplified genomic regions in these two species are homologous to the *B. rapa* sequence, even if only a small proportion of the region was compared (Fig. 3, A). Altogether, these results suggest that the S-locus is not located at the "Arabidopsis" S-locus position in *C. maritima* and probably also in *P. stylosa*.

Fig. 3. Characterization of the Arabidopsis S-locus region in Brassiceae species. A. Analysis of sequence homology in the genomic region flanking by the two genes *B80/U-Box* and *ARK3* in the *Brassica oleracea* (Boleracea), *Brassica nigra* (Bnigra) and *Raphanus sativus* (Rsativus) genomes (genomic data) as well as in the genome of *Cakile maritima* (Cakile_F and Cakile_R) and *Psychine stylosa* (Psychine_F and Psychine_R) (PCR fragments sequences, see main text). The *B. rapa* sequence was used as the reference (top). The *B80/U-Box* (Bra013521) and *ARK3* (Bra013522) genes are indicated with exons (in purple) and introns (in pink) structure. B. Polyacrylamide-gel electrophoresis of the PCR products obtained for *C. maritima* and *P. stylosa* (*Cakile maritima*-Cm, *Carrichtera annua*-Ca, *schouwia purpurea*-Sp, *Psychine stylosa*-Ps, *Orychophragmus violaceus*-Ov, *Brassica insularis*-Bi). DNA sizes are shown on the right (M: 1000bp ladders).

A.



B.



Phylogenetic relationships of SRK sequences from Arabidopsis and Brassiceae species

At first, we compared allelic diversity in Arabidopsis and Brassica species and we investigated their phylogenetic relationships. Comparably high levels of within-species allelic diversity were found in both Arabidopsis and Brassica species, as well as high levels of allele sharing between congeneric species (Fig. 4, Fig S2). However, *SRK* sequences of *A. halleri* and *A. lyrata* were distributed across many ancient lineages, with a combination of highly diverged and highly related sequences, whereas *SRK* sequences of *B. rapa* and *B. oleracea* clustered into only two monophyletic groups, the so-called classes I and II, evolving independently from ancient allelic lineages since they are intermingled within *SRK* Arabidopsis clusters (Schierup et al. 2001; Castric and Vekemans 2007). Indeed, the Brassica class I *SRK* alleles are sister group to *AlSRK10* and *AhSRK35*, as well as to *AlSRK27* and *AhSRK37*, whereas the Brassica class II *SRK* alleles cluster together with a monophyletic group of Arabidopsis *SRK* alleles including *AhSRK41*, *AhSRK22*, *AhSRK34* and *AlSRK09* (Fig. 4, Fig S2). These results are compatible with a scenario involving a very strong bottleneck in the history of the Brassica genus, followed by allelic re-diversification at the S-locus (Edh et al. 2009; Leducq et al. 2014).

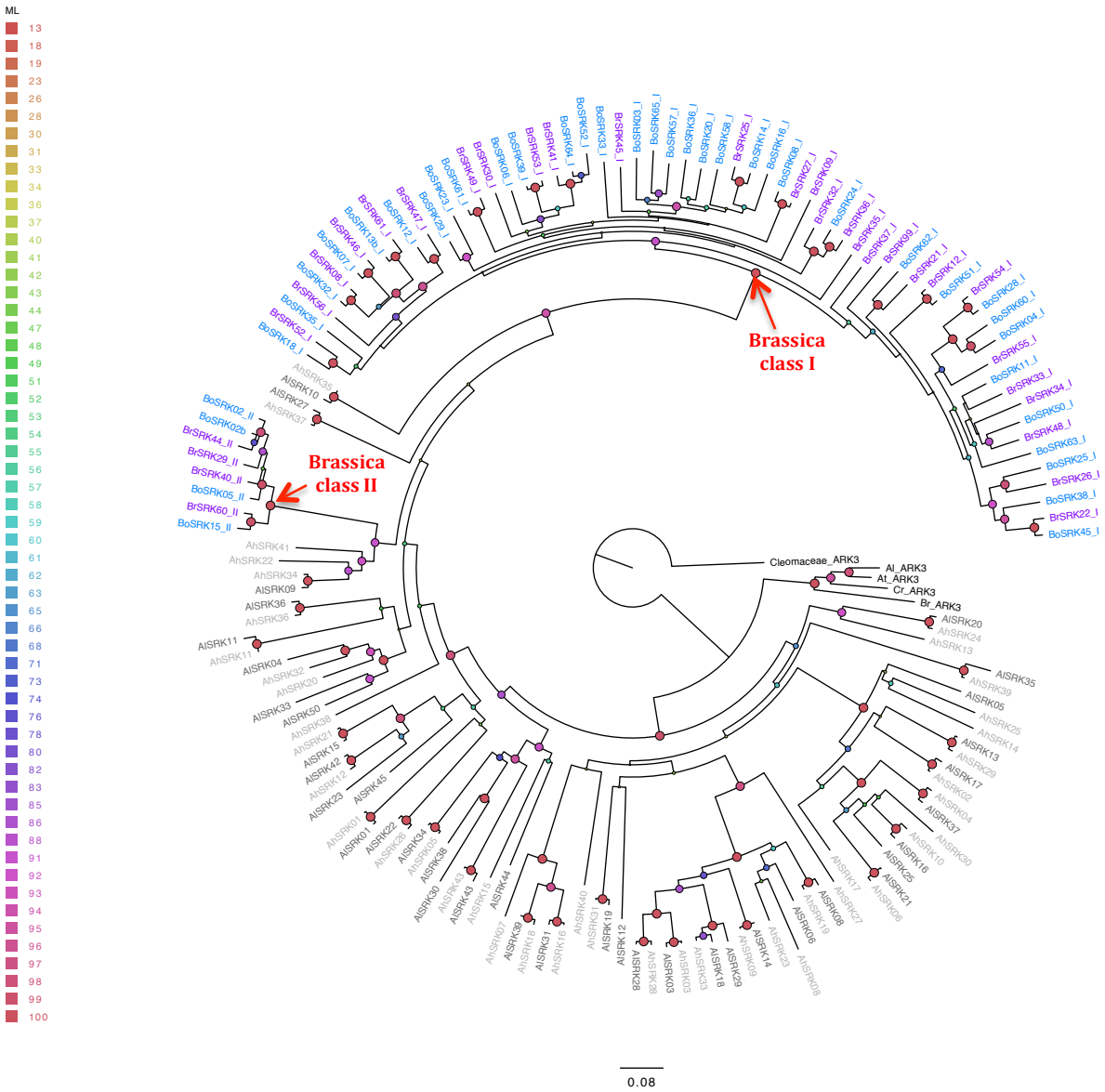


Fig. 4. Phylogenetic relationships inferred from ML analysis between *SRK* sequences of *Brassica rapa* (Br, in purple), *B. oleracea* (Bo, in blue), *Arabidopsis halleri* (Ah, in light grey) and *A. lyrata* (Al, in dark grey) and *ARK3* sequences from multiple Brassicaceae species (*A. thaliana*-At, *A. lyrata*-Al, *B. rapa*-Br, *Capsella rubella*-Cr, in black). The *ARK3* paralog sequence of the species *Cleome hassleriana* (Cleomaceae) was used as outgroup. The two monophyletic groups of Brassica S-alleles sequences are indicated with red arrows. Bootstrap values of each node are indicated by a colour gradient (legend on the right) and the circle size at each node is proportional to the support. Accession numbers of all *SRK* sequences are detailed in Table S3.

In order to test our hypothesis that this genetic bottleneck was associated with the mesohexaploid event in the history of the Brassiceae, we investigated the *SRK* allelic diversity in other members of the Brassiceae tribe. By using all *SRK* and *SLG* sequences of Brassiceae species compiled from published data (Table S3), together with one *SRK* sequence from each major clades of Arabidopsis S-alleles, several *SRK* paralogous sequences present in the Brassica genomes (*ARK3*, *SLR2* and *SUI*) and *SRK* and *SRK*-like sequences reconstructed from our RNAseq data, we constructed a second phylogenetic tree. Both ML and Bayesian phylogenies show that class I *SRK* and *SLG* alleles are shared by all investigated Brassiceae species. Furthermore, *SRK* and *SLG* sequences from all Brassiceae species were intermingled in the phylogenies, suggesting that the allelic diversification of the class I alleles occurred prior to the Brassiceae' diversification (Fig. 5, Fig S3). The putative *SRK* sequences of *Orychophragmus violaceus* cluster with the *SRK* sequence of *Eutrema salsugineum* (Vekemans et al. 2014), suggesting that the genetic bottleneck experienced by Brassica and the other Brassiceae species occurred after the divergence between *O. violaceus* and the Brassiceae tribe. Similarly, the *SRK* sequence of *Sisymbrium irio* falls outside the clas I clade (Fig. 5, Fig S3), implying that the genetic bottleneck occurred after the divergence between this species and the Brassiceae tribe.

Surprisingly, and contrary to the pattern observed in the class I clade, the Brassiceae sequences that clustering into the class II clade were not intermingled and exhibited a more complex pattern with many species-specific allelic clusters and co-clustering of known Brassica *SRK* paralogs (*SUI* and *SUI2* involved in intraspecific unilateral incompatibility in *B. rapa* and located on chromosome A04, and *SLR2 – Bra038635 –* which is located on *B. rapa* chromosome A06) with sequences obtained from other Brassiceae species (Fig. 5, Fig S3). The phylogenetic patterns may suggest that class II alleles diversified only after the divergence of the Brassiceae clades, as for instance all Brassica (clade Oleracea) class II sequences cluster together, the Sinapis and Raphanus (clade Nigra) sequences are basal to the Brassica sequences, and the Cakile (clade Cakile) and Crambe (clade Crambe) sequences are basal to clade Nigra sequences. However, these results should be taken with caution due to the lack of strong phylogenetic support, and to the lack of genetic or physical mapping of the obtained sequences, which does not allow distinguishing true S-locus sequences from paralogs. With these precautions in mind, it is tempting to distinguish two groups of sequences in the *SRK* class II-like clusters: a group of putative *SRK* sequences of class II that would form a putative "core clade" of class II S-alleles (Fig. 5) and another group

corresponding to *SRK* paralogous sequences, that diverged before the class II S-alleles (Fig. 5). Altogether, these results suggest that we probably recovered one or several parologs close to the class II S-alleles in the following Brassiceae species: *Cakile maritima*, *Crambe maritima*, *Psychine stylosa*, *Schouwia purpurea*, *Zilla spinosa* and *Carrichtera annua*. Furthermore, these results also suggest that we probably recovered class II S-alleles only for the following species: *Cakile maritima*, *Crambe maritima*, *Raphanus sativus, Sinapis arvensis, Brassica rapa* and *B. oleracea* that are known to be self-incompatible, or partially self-incompatible (*Crambe maritima*).

Three near identical sequences of *Carrichetera annua* (clade "Vella") and one sequence of *Zilla spinosa* (clade "Zilla") clustered together with the putative *SRK* sequences of *O. violaceus* and the true *SRK* sequence of *E. salsugineum* (Fig. 5, Fig S3), in the vicinity of the Brassica class II cluster. These sequences could belong to a paralog of the *SRK* gene that existed in the ancestor of the Brassiceae and that had been lost in other Brassiceae clades. However, such sequence was not found in *Schouwia purpurea* that belongs to the same clade of *Z. spinosa*.

All sequences obtained from *Carrichtera annua*, both in the class I or class II *SRK* clusters, showed stop codons and/or frameshift mutations (data not shown), suggesting that these *SRK* sequences are probably non-functional pseudogenes in this species, known to be self-compatible (Boaz et al. 1990 and phenotypic results from this thesis in annex). For all other Brassiceae species, even for the known self-compatible species (*Schouwia purpurea*, *Psychine stylosa,* phenotypic results from this thesis), reconstructed sequences from RNAseq data did not show any stop codons and/or frameshift mutations (data not shown). Two sequences from *Cakile maritima* were sister to those of the *SUI* gene of *B. rapa*, which could suggest that this *SRK* paralog described in *B. rapa* is also present in *C. maritima*. However, we have to take some caution due to the lack of strong phylogenetic support.

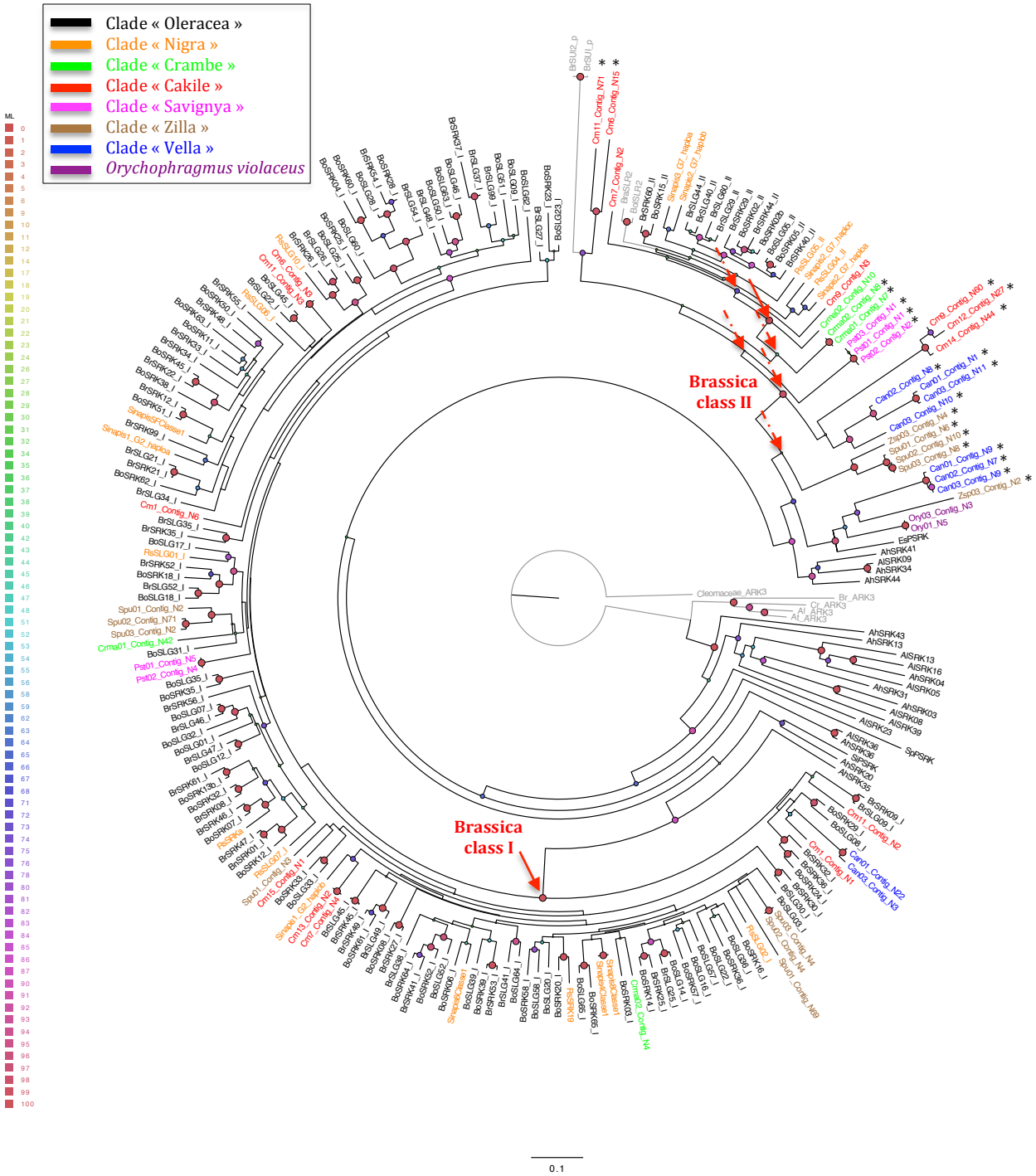


Fig. 5. Phylogenetic relationships inferred from ML analysis between *SRK* and *SLG* sequences. The *SRK* and *SRK*-like sequences were obtained from the following species: *Brassica rapa* (Br), *B. oleracea* (Bo), *Raphanus sativus* (Rs), *Sinapis arvensis* (Sinapis), *Arabidopsis halleri* (Ah), *A. lyrata* (Al) (in black), *Schrenkiella parvula* (Sp), *Sisymbrium irio* (Si), *Eutrema salsugineum* (Es) (in black), *Carrichtera annua* (Can), *Zilla spinosa* (Zsp), *Schouwia purpurea* (Spu), *Psychine stylosa* (Pst), *Crambe maritime* (Crma), *Cakile maritima* (Cma) (see the colour legend) and *Orychophragmus violaceus* (Ory). *ARK3* sequences from multiple Brassicaceae species (*A. thaliana*-At, *A. lyrata*-Al, *B. rapa*-Br, *Capsella rubella*-Cr) are displayed in light grey as well as two Brassica paralogs of *SRK* (*SLR2* and *SUI*). The *ARK3* paralog sequence of the species *Cleome hassleriana* (Cleomaceae) was used as outgroup. The two monophyletic groups of Brassica S-alleles sequences are indicated with

red arrows. The core clade of class II S-alleles (see main text) is indicated with a full line whereas the other putative positions of the clade root are indicated with dotted lines. An asterisk indicates putative *SRK* paralogs (see main text). Bootstrap values of each node are indicated by a colour gradient (legend on the right) and the circle size at each node is proportional to the support. Accession numbers of all *SRK* sequences from other studies are detailed in Table S3.

Structural variation within the S-locus

Our analysis showed that the *SRK* gene of the S-haplotype extracted from the *Brassica oleracea* genome from Parkin et al. (2014) displays a strong homology with the four following class I S-haplotypes: *BoSRK28*, *BrSRK54*, *BoSRK60*, and *BoSRK04* that belong to a monophyletic group in our *SRK* phylogeny (Fig. 4, Fig. S2). However, *SRK* in this accession is non-functional, as only the four last exons could be found. Thus, this haplotype has been classified in the class I S-haplotypes and the *SRK* allele was considered as being a non-functional copy of *BoSRK28*. Furthermore, the *SCR*/*SP11* gene was not found in this genome. This individual was from the morphotype To1000 that is self-compatible (Parkin et al. 2014) and the non-functional state of the self-incompatibility system might open the way for the accumulation of multiple loss-of-function mutations at S-locus genes, making it difficult to infer the initial causal mutation. The *SRK* gene of the S-haplotype of the *B. oleracea* genome from Liu et al. (2014) showed a strong homology (99%) with the *BoSRK2* class II S-haplotype, thus this haplotype has been classified in the class II S-haplotypes.

The S-locus region is variable in size across haplotypes (Table 3), spanning from 48 kb in BrS-8 (where the three flanking genes SP1, AtPP and SIAH2/ORF-b were absent) and 193 kb in the S-haplotypes of the genome of *B. oleracea* from Liu et al. (2014), with an average size of 95 kb.

Table 3. Size and gene order within the different S-haplotypes.

Species	Haplotype	Phylogenetic class	Size of the S-locus (bp)	SI-related genes order
Brassica rapa	BrS8	I	48,581	SLG → SCR → SRK
	BrS46	I	65,377	SLG → SCR → SRK
	BrS60	II	85,523	SCR → SRK → SLG
Brassica oleracea	BoSRK2[a]	II	193,442	SCR → SRK → SLG
	BoSRK28[b]	I	125,200	SLG → SCR → SRK
Raphanus sativus	RsS19	I	50,234	SLG → SCR → SRK

[a] *B. oleracea* genome from Liu et al. (2014)
[b] *B. oleracea* genome from Parkin et al. (2014)

Synteny was highly conserved in the flanking regions on both sides of the S-locus (Fig. 6). Within the S-locus, the gene order was conserved among the four S-haplotypes of class I, as observed by Takuno et al. (2007) among three *B. rapa* S-haplotypes of class I, whereas it was different but conserved among the two S-haplotypes of class II (Fig 6). Hence, two major genomic organizations of the S-haplotypes can be described, each specific to each S-haplotypes class (Table 3, Fig. 6): (1) *SLG-SCR-SRK* (class I S-haplotypes genomic organization) and (2) *SCR-SRK-SLG* (class II S-haplotypes genomic organization).

The two S-haplotypes-linked small RNA, *SMI* and *SMI2*, were found within the S-locus upstream and downstream, respectively, in the class II S-haplotype of *B. rapa* whereas only *SMI2* was found in the class II S-haplotype of *B. oleracea* from Liu et al. (2014). The *B. rapa* class-I *SMI* sRNA suppresses expression of the recessive class-II *SCR/SP11* genes and does not have any known functional role in dominance hierarchy between class II S-alleles, which could explain its loss in this individual (Goring 2016). As expected, all the four class I S-haplotypes were carrying the *SMI* sRNA upstream within the S-locus (Fig 6).

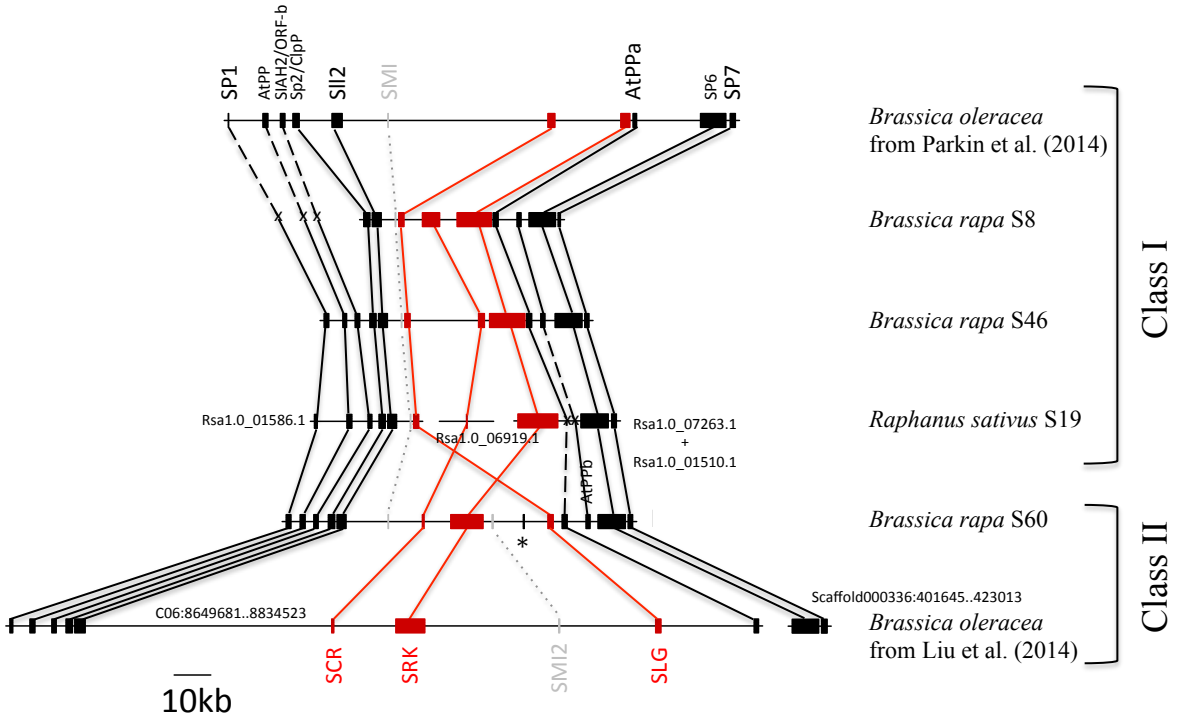


Fig. 6. Structural variation within the S-locus. The figure shows the synteny among the S-locus region for various S-haplotypes of three SI species of Brassiceae. The SI genes *SCR/SP11*, *SRK* and *SLG* are shown in red; flanking genes in black; the two S-haplotypes-linked small RNAs in grey. An annotated putative transposable element found in the *Brassica rapa* genome is indicated by an asterisk. Note that for the *SCR/SP11* gene of *Raphanus sativus*, the figure is not to scale. In the S-haplotype of the individual of *Brassica oleracea* sequenced by Parkin et al. (2014), only the four last exons have been found for the gene *SRK* and the gene *SCR* is lacking, in agreement with the self-compatibility reported by the authors. For the S-haplotypes that were distributed on several DNA fragments, the name of each fragment is indicated. Gene order is conserved among haplotypes of the same allelic class.

Discussion

Is there a translocation of the S-locus region in the ancestor of the tribe Brassiceae?

Although the Arabidopis and Brassica genera seem to share the same ancestrally derived *SCR/SP11 – SRK* based sporophytic SI system, their S-locus genomic locations are known to be distinct (Kusaba et al. 2001; Fobis-Loisy and Gaude 2004; Kim et al. 2016). Between these two, it is the "Arabidopsis" S-locus location (within the ancestral genomic block U) that is supposed to be ancestral, because species from several widely divergent Brassicaceae lineages share this genomic location (Chantha et al. 2013, Vekemans et al. 2014), including the close outgroup to the Brassiceae tribe, *Sisymbrium irio*. However, to the best of our knowledge, the derived S-locus genomic location found in *B. rapa* has never been investigated in other and more divergent Brassiceae species. Here, we compared the genomic location of the S-locus in both *B. rapa* and *B. oleracea* and we showed that they are shared between these two congeneric species. Moreover, noting that both species share an ancient whole genome triplication event (Liu et al. 2014), we showed that the S-locus is absent, possibly as a result of independent deletions and/or translocations, in the three homoeologous regions corresponding to the ancestral Arabidopsis S-locus location within the mesohexaploid Brassica genomes. Within the two Brassica genomes, the S-locus was found only within one subgenome at the "Brassica" location (within the ancestral genome block E), *i.e.* subgenome LF, which is known to be the least fragmented subgenome, suggesting, if we assume that the translocation occurred after the mesohexaploid event, a single translocation from the original Arabidopsis S-locus location from one of the three subgenomes to the final LF subgenome Brassica S-locus location.

Furthermore, we also analyzed the "Arabidopsis" S-locus region in the genome of several and more divergent Brassiceae species (*i.e. Brassica nigra*, *Raphanus sativus*, *Cakile maritima*, and *Psychine stylosa*) and found that the S-locus is absent at this genomic location in the studied Brassiceae species. However, our data do not allow us to determine whether the genomic location of the S-locus in these species (except for *R. sativus*) is the same as that of *B. rapa* and *B. oleracea*, which remains to be investigated. Within the *R. sativus* published genome sequence (Kitashiba et al. 2014), carrying the S-haplotype S-19, the S-locus was found to occupy the same genomic location than that of *B. rapa* and *B. oleracea* (Fig. 6 and Kim et al. 2016). *B. rapa* and *B. oleracea* belong to the same Brassiceae clade ("Oleracea"), whereas *B. nigra* and *Raphanus sativus* belong to another Brassiceae clade ("Nigra") and *Cakile maritima* and *Psychine stylosa* belong respectively to the clades "Cakile" and

"Savignya". These four clades were found to be monophyletic (see Chapter 1), with the clades "Oleracea" and "Nigra" as sister clades. Altogether, this strongly suggests that (1) the genomic translocation of the S-locus to the "Brassica" location probably occurred before divergence of the two clades "Oleracea" and "Nigra", and (2) this translocation could be more ancient and have occurred before the divergence of the four clades "Oleracea", "Nigra", "Cakile" and "Savignya" although this remains to be confirmed. If this latter hypothesis should be rejected, it means that two other independent genomic translocations of the S-locus have occurred, one in the "Cakile" lineage and the other in the "Savignya" lineage, which is not the most parsimonious scenario. The absence of conclusive evidence for the two Brassiceae species *Carrichtera annua* (Clade "Vella") and *Schouwia purpurea* (Clade "Zilla") could suggest either that the S-locus is still present at this ancestral genomic location in these two clades, or could be due to a lack of specificity of the primers used in amplifying the intergenic region between the two Arabidopsis S-locus flanking genes (generating a false negative result). The latter hypothesis is corroborated by the fact that we failed to obtained PCR fragments from two samples of *Brassica insularis*, whereas this species belong to the "Oleracea" clade (Warwick and Sauder 2005) and therefore, should probably share the genomic location of the S-locus with its relatives *B. rapa* and *B. oleracea*. Hence, we cannot conclude about the genomic location of the S-locus in the most basal clades of the Brassiceae tribe.

Testing for an association of whole genome triplication, genetic bottleneck and genomic translocation of the S-locus in the tribe Brassiceae

Previous phylogenies of the SI-related genes (*SRK*, *SLG*) in Brassica and Raphanus have suggested that there exist both class I and class II S-alleles in *B. rapa*, *B. oleracea* and *R. sativus* and that the diversification of the S-alleles occurred prior to their divergence (Lim et al. 2002; Sato et al. 2002). Our phylogenetic analysis of almost complete sets of extant *SRK* alleles from two pairs of SI Arabidopsis species, and from *B. rapa* and *B. oleracea*, strongly suggests that class I and class II *SRK* sequences evolved independently from ancestral S-allele lineages shared between the Arabidopsis and Brassica lineages. This observation supports the hypothesis that a strong genetic bottleneck was responsible for the S-alleles clustering in Brassica by reducing dramatically the number of S-alleles (Castric and Vekemans 2007, Edh et al. 2009). As the intensity of the negative frequency-dependent selection acting on the S-locus is inversely related to the number of S-alleles in the population (Wright 1939; Schierup et al. 1997), strong selection for allelic re-diversification would have followed the putative

bottleneck, and would have led to recent and independent allelic re-diversification within class I and class II allele lineages. This phenomenon has been suggested to be responsible for the higher estimates of the ratios of non-synonymous to synonymous substitutions (due to positive selection) in *B. rapa SRK* sequences than in *Arabidopsis halleri* and *A. lyrata* (Castric and Vekemans 2007). Nowadays, the number of S-alleles in Brassica is very high possibly because of an intense allelic diversification after the putative bottleneck. A similar observation was reported in the genus Biscutella where three S-allele clusters, different from both S-allele clusters of Brassica, have been identified (Leducq et al. 2014).

Phylogenetic analyses of *SRK* sequences from different species of Brassiceae in the present study suggest that the putative genetic bottleneck occurred prior to the divergence of all Brassiceae clades and after the divergence between the outgroup species *Sisymbrium irio* and *Orychophragmus violaceus*, and the Brassiceae tribe. Therefore, the genetic bottleneck at the S-locus occurred putatively in association with the allohexaploidy (Whole Genome Triplication, WGT) event experienced by the ancestor of the tribe Brassiceae (Lysak et al. 2007; Lysak et al. 2005; Wang et al. 2011; Chapter 2). Indeed, a bottleneck in population size associated with the founding of new polyploid lineages (Otto 2007), generating a strong genetic bottleneck at the genome scale, should dramatically reduce the number of S-alleles (Ramsey and Schemske 1998). The genomic translocation of the S-locus could have occurred shortly after the WGT event, reinforcing the S-locus genetic bottleneck caused by the WGT. In this hypothesis, two distantly related *S*-alleles (the ancestors of the current class I and class II S-alleles) would have been conserved in the allohexaploid ancestor of the tribe Brassiceae at the "Arabidopsis" S-locus location and then, the S-locus would have been translocated to a new location, the "Brassica" S-locus location. This implies that the two ancestral S-alleles would have been translocated to the same genomic location, which is far from being parsimonious. Furthermore, this is not supported by our results suggesting that the class II S-alleles, in contrast to class I S-alleles, may have diversified after the diversification of the Brassiceae clades and would thus have a more recent origin than the class I S-alleles (Fig. 5, Fig S3). An alternative hypothesis could be that the allohexaploid ancestor of the tribe Brassiceae had a paralogous copy close to the current class II S-alleles (called ClassII_P) at an unknown genomic location and that the initial genomic translocation of the S-locus did only concern the ancestral class I S-allele, shortly after the WGT and before the diversification of the tribe Brassiceae. Genomic translocations occurring shortly after an allopolyploidy event are known to be frequent, in association with the frequently induced transposable elements remobilization process (Parisod et al. 2010). Subsequently, strong

allelic re-diversification would have occurred at the "neo-S-locus" during a certain period of time, prior to the divergence of the Brassiceae clades, resulting in trans-specific polymorphisms among Brassiceae clades within class I S-alleles (Fig. 6). Later, recruitment of a copy of the paralog ClassII_P as a new S-allele could have occurred through ectopic recombination, creating a single class II S-haplotype, which would not have been subject to allelic re-diversification before separation of the Brassiceae clades. Knowing that class II S-alleles are recessive with respect to class I S-alleles, their more limited and slower allele re-diversification process would be in line with theoretical results suggesting much lower turnover rates and longer lifespan for recessive versus dominant alleles in a sporophytic SI system (Schierup et al. 1997). Such scenario could explain some of our observations, such as (1) that the *SRK* sequences of class II from Brassiceae species tended to form species-specific clades whereas those of class I were intermingled in our phylogenetic tree, and (2) that we found putative *SRK* paralogous sequences close to the class II S-alleles (referred to as ClassII_P) in most of the studied Brassiceae species (Fig 5, Fig S3). However, we do not have any information about the genomic location of our *SRK* sequences obtained from RNAseq data and we cannot distinguish between true class II S-alleles and paralogs (Fig. 5). Thus, no firm conclusion can be drawn concerning the status and origin of class II S-haplotypes within the Brassiceae tribe, but evidence for a shared S-locus genetic bottleneck among all Brassiceae is rather conclusive and would support the hypothesis that this bottleneck would have been caused by the shared mesohexaploid event.

Two distinct genomic organizations of the S-haplotypes in class I and class II alleles

Structural variation within the S-locus has been identified in several S-haplotypes of the species *Arabidopsis halleri* and *A. lyrata* where both the orientation of genes and the gene order was highly variable among S-haplotypes (Goubet et al. 2012). In Brassiceae, one previous study had reported heterogeneity in gene organization within the S-locus (Kim et al. 2016), with a difference in gene order within the S-locus region between two S-haplotypes, one of class I (*Raphanus sativus* S19) and the other of class II (*Brassica rapa* S60). In contrast, Takuno et al. (2007) have reported a conservation of the gene order among three *B. rapa* S-haplotypes of class I (BrS-8, BrS-46, BrS-54), whereas Shiba et al. (2003) have reported a similar conclusion for two other S-haplotypes of class I in *B. rapa* (BrS-12 and BrS-9). In contrast, in the class-II BrS60 haplotype, the order of SRK and SCR/SP11 was the reverse of that in class-I S haplotypes (Fukai et al. 2013). In our present study, the genomic sequences of six S-haplotypes from three Brassiceae species belonging to the two S-haplotype

classes were compared. Our results suggest the presence of two major genomic organizations of the S-haplotypes within Brassiceae, one specific to each S-haplotype class ('*SLG-SCR-SRK*' for class I, '*SCR-SRK-SLG*' for class II). Given the divergent phylogenetic positions of the two class II S-haplotypes analysed here (Fig. 4, Fig. S2), the observed genomic organization of the S-locus region is probably shared by all class II S-haplotypes, at least in Brassica. The same conclusion can be drawn for the class I S-haplotypes, in the light of our *SRK* phylogeny. It is tempting to speculate that the S-haplotype ancestor of all extant S-haplotypes of class I had the gene order *SLG-SCR-SRK* whereas those of all extant S-haplotypes of class II had the gene order '*SCR-SRK-SLG*', even if we should not exclude the scenario of an identical gene order in the two ancestral S-haplotypes followed by a gene reorganization within one or both S-haplotype classes. However, only few S-haplotype sequences were considered in our study and no firm conclusion should be drawn before analysing more sequences, especially for the class I S-haplotypes exhibiting very high sequence diversity. Moreover, future genomic resources for the basal clades of the tribe Brassiceae could be useful to clarify this hypothesis.

Surprisingly, the size of the *B. oleracea* S-haplotypes is, on average, at least two-fold higher than those of the *B. rapa* S-haplotypes, and three-fold higher than those of the *R. sativus* S-haplotype (Table 3). This could be explained by the greater accumulation of transposable elements in the genome of *B. oleracea* compared to those of *B. rapa* in association with a ~30% larger genome size in *B. oleracea* compared to *B. rapa*, although the two genomes share the same ploidy level and are largely collinear (Liu et al. 2014), whereas similar genome size and content in repetitive sequences are observed in *B. rapa* and *R. sativus* (Moghe et al. 2014). Furthermore, particularly high accumulation of transposable elements in the S-locus region has been documented in Arabidopsis and Brassica, and is thought to be consecutive to the suppression of recombination at the S-locus region, in analogy to Y chromosomes in dioecious species (Cui et al. 1999; Kimura et al. 2002; Shiba et al. 2003; Fujimoto et al. 2006; Goubet et al. 2012).

Supplementary figures.

Fig. S1. Overview of the S-locus region in Arabidopsis (on the chromosome 4). Thin arrows indicate the position and the orientation of the primers used for the amplification of the genomic region flanking by the *B80/U-Box* (*AT4G21350*) and *ARK3* (*AT4G21380*) genes (see Table 1).

Fig. S2. Phylogenetic relationships inferred from BI analysis between *SRK* sequences of *Brassica rapa* (Br, in purple), *B. oleracea* (Bo, in blue), *Arabidopsis halleri* (Ah, in light grey) and *A. lyrata* (Al, in dark grey) and *ARK3* sequences from multiple Brassicaceae species (*A. thaliana*-At, *A. lyrata*-Al, *B. rapa*-Br, *Capsella rubella*-Cr, in black). The *ARK3* paralog sequence of the species *Cleome hassleriana* (Cleomaceae) was used as outgroup. The two monophyletic groups of Brassica S-alleles sequences are indicated with red arrows. Posterior probability values of each node are indicated by a colour gradient (legend on the right) and the size of the circle on each node is proportional to the support. Accession numbers of all *SRK* sequences are detailed in Table S3.

Fig. S3. Phylogenetic relationships inferred from BI analysis between *SRK* and *SLG* sequences. The *SRK* and *SRK*-like sequences were obtained from the following species: *Brassica rapa* (Br), *B. oleracea* (Bo), *Raphanus sativus* (Rs), *Sinapis arvensis* (Sinapis), *Arabidopsis halleri* (Ah), *A. lyrata* (Al) (in black), *Schrenkiella parvula* (Sp*), Sisymbrium irio* (Si), *Eutrema salsugineum* (Es) (in black), *Carrichtera annua* (Can), *Zilla spinosa* (Zsp), *Schouwia purpurea* (Spu), *Psychine stylosa* (Pst), *Crambe maritime* (Crma), *Cakile maritima* (Cma) (see the colour legend) and *Orychophragmus violaceus* (Ory). *ARK3* sequences from multiple Brassicaceae species (*A. thaliana*-At, *A. lyrata*-Al, *B. rapa*-Br, *Capsella rubella*-Cr) are displayed in light grey as well as two Brassica paralogs of *SRK* (*SLR2* and *SUI*). The *ARK3* paralog sequence of the species *Cleome hassleriana* (Cleomaceae) was used as outgroup. The two monophyletic groups of Brassica S-alleles sequences are indicated with red arrows. Only the core clade of class II S-allele is indicated with a full line. Posterior probability values of each node are indicated by a colour gradient (legend on the right) and the circle size at each node is proportional to the support. Accession numbers of all SRK sequences from other studies are detailed in Table S3.

Supplementary tables.

Table S1. The selected genes flanking the two S-locus regions: « Arabidopsis » and « Brassica » locations. For each of the two genomic locations, flanking genes along with the genomic positions and the number of annotated genes are indicated for *A. thaliana* and the three subgenomes of *B. rapa* and *B. oleracea* (LF, MF1 and MF2). An asterisk indicates inversion of the region.

Table S2. Annotated genes from the S-locus region of the genome of *Brassica rapa* (Wang et al. 2011, Cheng et al. 2011). Genes in bold have been annotated manually by using blastn against the following reference sequences of *B. rapa*: AB097116.1 (SRK-60 and SLG-60) and AB067446.1 (SP11-60). TE: transposable element.

Table S3. Published *SRK*, *SLG* and *SRK* paralogous sequences (*SUI, SLR2*, *ARK3*) used in the phylogenetic analysis of the *SRK* sequences among the Brassiceae tribe. *B. oleracea*: *Brassica oleracea*; *B. rapa*: *Brassica rapa*; *B. napus*: *Brassica napus*; *R. sativus*: *Raphanus sativus*.

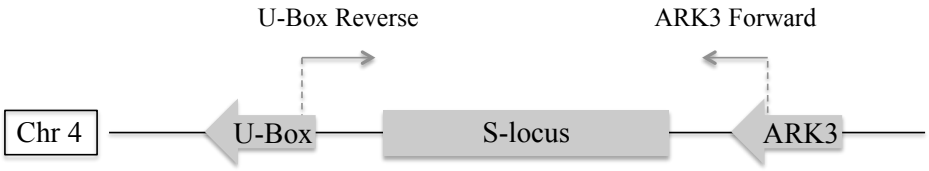


Fig. S1. Overview of the S-locus region in Arabidopsis (on the chromosome 4). Thin arrows indicate the position and the orientation of the primers used for the amplification of the genomic region flanking by the *B80/U-Box* (*AT4G21350*) and *ARK3* (*AT4G21380*) genes (see Table 1).

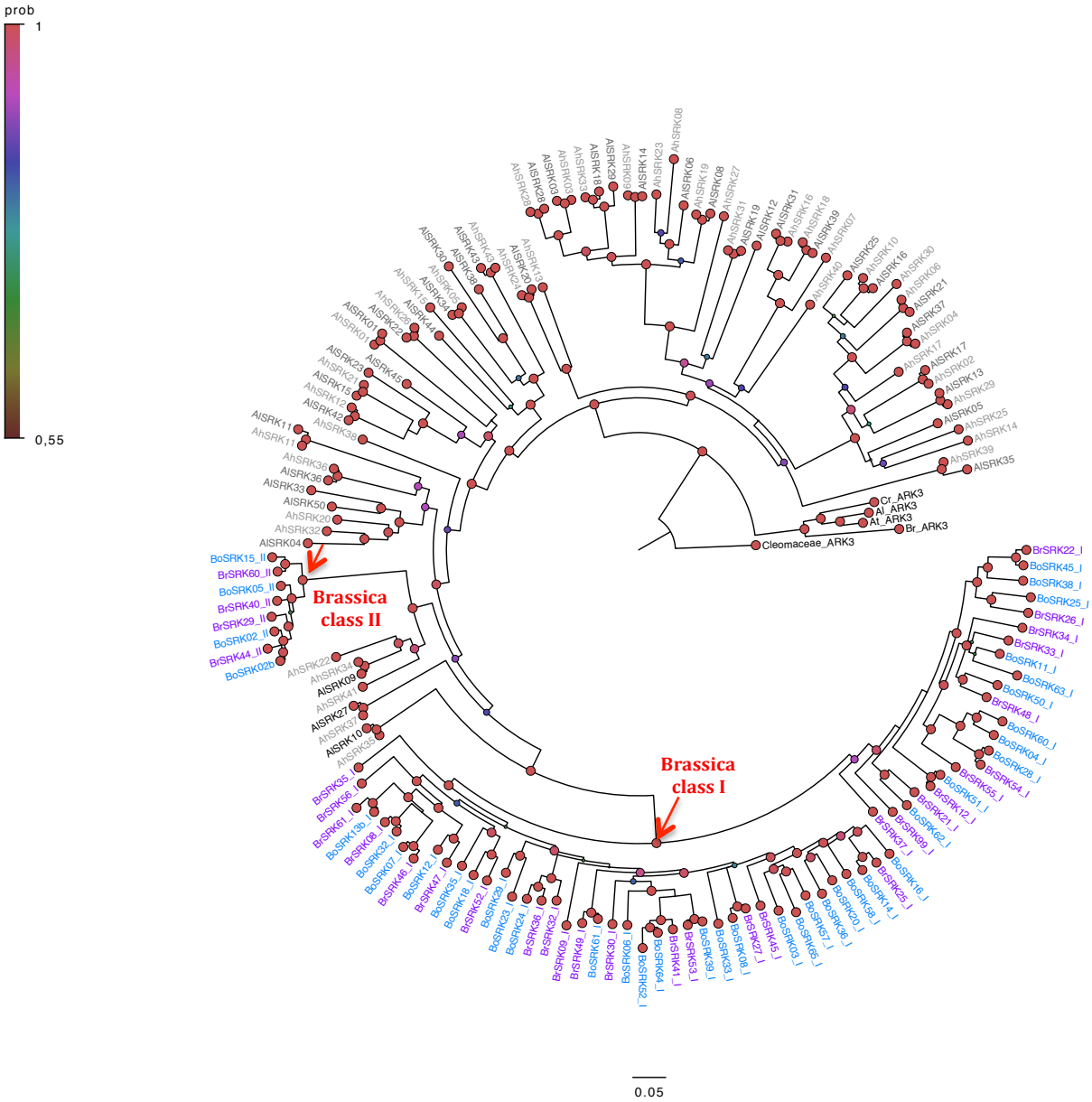


Fig. S2. Phylogenetic relationships inferred from BI analysis between *SRK* sequences of *Brassica rapa* (Br, in purple), *B. oleracea* (Bo, in blue), *Arabidopsis halleri* (Ah, in light grey) and *A. lyrata* (Al, in dark grey) and *ARK3* sequences from multiple Brassicaceae species (*A. thaliana*-At, *A. lyrata*-Al, *B. rapa*-Br, *Capsella rubella*-Cr, in black). The *ARK3* paralog sequence of the species *Cleome hassleriana* (Cleomaceae) was used as outgroup. The two monophyletic groups of Brassica S-alleles sequences are indicated with red arrows. Posterior probability values of each node are indicated by a colour gradient (legend on the right) and the size of the circle on each node is proportional to the support. Accession numbers of all *SRK* sequences are detailed in Table S3.

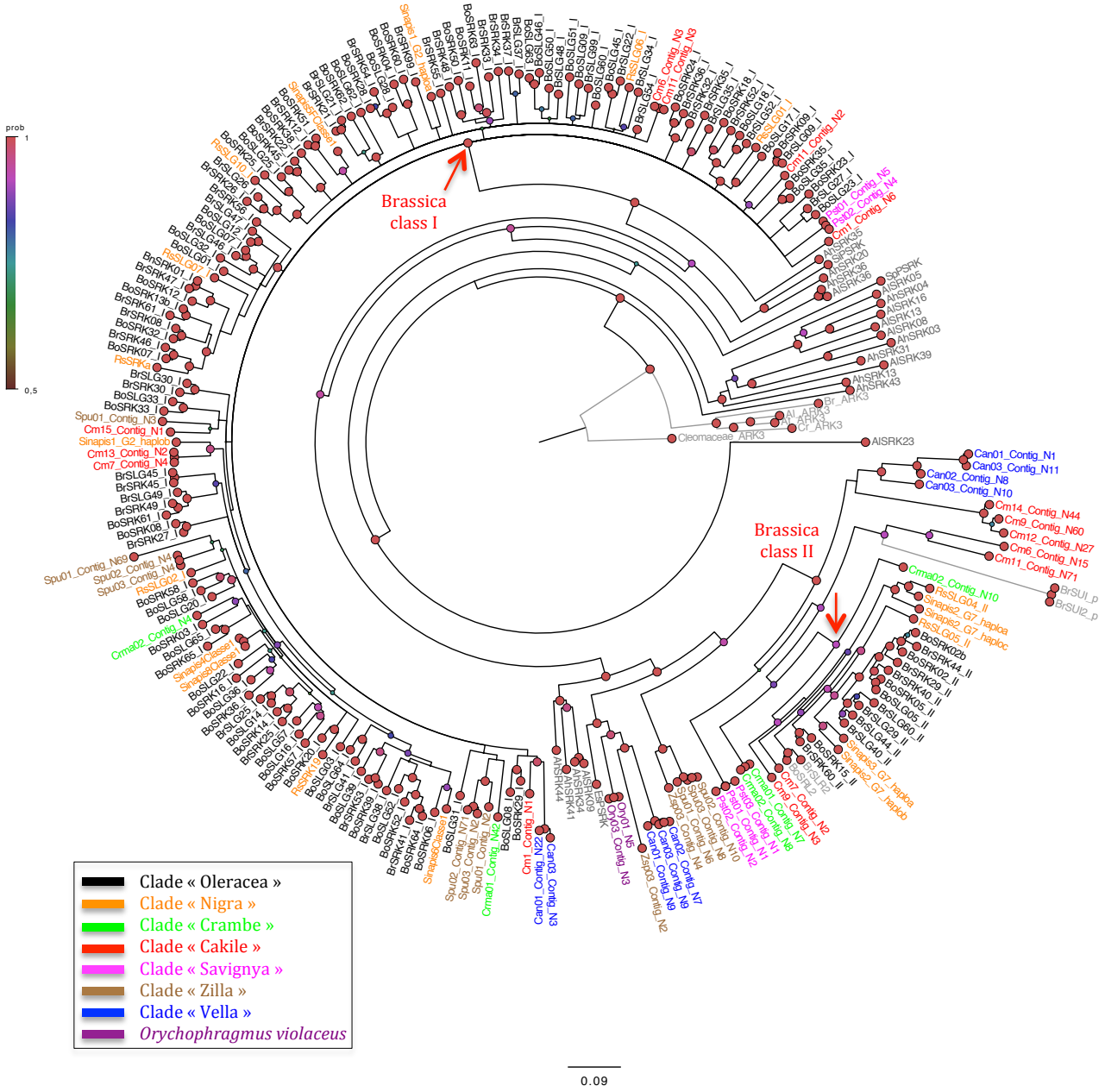


Fig. S3. Phylogenetic relationships inferred from BI analysis between *SRK* and *SLG* sequences. The *SRK* and *SRK*-like sequences were obtained from the following species: *Brassica rapa* (Br), *B. oleracea* (Bo), *Raphanus sativus* (Rs), *Sinapis arvensis* (Sinapis), *Arabidopsis halleri* (Ah), *A. lyrata* (Al) (in black), *Schrenkiella parvula* (Sp), *Sisymbrium irio* (Si), *Eutrema salsugineum* (Es) (in black), *Carrichtera annua* (Can), *Zilla spinosa* (Zsp), *Schouwia purpurea* (Spu), *Psychine stylosa* (Pst), *Crambe maritime* (Crma), *Cakile maritima* (Cma) (see the colour legend) and *Orychophragmus violaceus* (Ory). *ARK3* sequences from multiple Brassicaceae species (*A. thaliana*-At, *A. lyrata*-Al, *B. rapa*-Br, *Capsella rubella*-Cr) are displayed in light grey as well as two Brassica paralogs of *SRK* (*SLR2* and *SUI*). The *ARK3* paralog sequence of the species *Cleome hassleriana* (Cleomaceae) was used as outgroup. The two monophyletic groups of Brassica S-alleles sequences are indicated with red arrows. Only the core clade of class II S-allele is indicated with a full line. Posterior probability values of each node are indicated by a colour gradient (legend on the right) and the circle size at each node is proportional to the support. Accession numbers of all SRK sequences from other studies are detailed in Table S3.

Table S1. The selected genes flanking the two S-locus regions: « Arabidopsis » and « Brassica » locations. For each of the two genomic locations, flanking genes along with the genomic positions and the number of annotated genes are indicated for *A. thaliana* and the three subgenomes of *B. rapa* and *B. oleracea* (LF, MF1 and MF2). An asterisk indicates inversion of the region.

Species	***Arabidopsis thaliana* [genomic location] (#genes)**	**LF [genomic location] total length (total intergenic length - transposon length), #genes**	**MF1 [genomic location] total length (total intergenic length - transposon length), #genes**	**MF2 [genomic location] total length (total intergenic length - transposon length), #genes**
	« Arabidopsis S-locus location » (ancestral genomic Blockk *U*)			
B. rapa[a]	AT4G21110 - AT4G21440 [Ch4:11,267,185..11,419,760] (152,575 bp, 38)	Bra013502 - Bra013526 [A01:6,015,579..6,147,165] 131,586 bp (68,877bp - 34,712bp), 25 genes	Bra038786 – Bra038774 [A03:24,074,763..24,163,765] 89,002 bp (49,823bp - 23,796bp), 13 genes	Bra020886 – Bra020892 * [A08:10,581,365..10,652,660] 71,295 bp (58,527bp - 34,074bp), 7 genes
B. oleracea[b]		Bo1g023860 - Bo1g024170 [C1:8,633,599..8,843,225] 209,626bp (191,952bp - NA), 32 genes	Bo7g107310 - Bo7g107420 [C7:41,924,782..42,009,025] 84,243bp (68,724bp - NA), 12 genes	Bo3g162750 - Bo3g162580 * [C3:57,425,647..57,574,826] 149,179bp (119,916bp - NA), 18 genes
	« Brassica S-locus location » (ancestral genomic Blockk *E*)			
B. rapa	AT1G66330 – AT1G67035 [Ch1:24,729,629..25,019,057] (289,428 bp, 57)	Bra004159 - Bra004203 [A07:20,414,062..20,675,962] 261,900 bp (169,780bp – 106,964bp), 43 genes	Bra039760 - Bra034035 [A02:8,808,090..9,015,357] 207,267 bp (135,379bp – 97,950bp), 23 genes	absent
B. oleracea		Bo6g099840 – Bo6g103560 [C6:31,828,362..32,372,867] (544,505bp, 73 genes)	Bo2g052820 – Bo2g055270 [C2: 15,235,649..15,606,451] (370,802bp, 46 genes)	absent

Table S2. Annotated genes from the S-locus region of the genome of *Brassica rapa* (Wang et al. 2011, Cheng et al. 2011). Genes in bold have been annotated manually by using blastn against the following reference sequences of *B. rapa*: AB097116.1 (SRK-60 and SLG-60) and AB067446.1 (SP11-60). TE: transposable element.

Gene orientation	ID*	Gene name	Genomic location
-	*Bra004174*	*SP1*	[A07: 20,491,331..20,492,506]
+	*Bra004175*	*AtPP?*	[A07:20,494,781..20,496,028]
-	*Bra004176*	*SIAH2/ORF-b*	[A07:20,498,136..20,499,242]
-	*Bra004177*	*Sp2/ClpP*	[A07:20,501,733..20,503,313]
-	*Bra004178*	*Sll2*	[A07:20,503,896..20,506,112]
-	**not annotated**	***SCR/SP11-60***	**[A07:20,524,950..20,525,574]**
+	***Bra004179***	***SRK-60***	**[A07:20,531,941..20,534,324]**
+	***Bra004180***		**[A07:20,538,789..20,539,928]**
-	*Bra004181*	putative TE	[A07:20,549,864..20,550,121]
+	***Bra004182***	***SLG-60***	**[A07:20,555,934..20,557,262]**
-	*Bra004183*	*AtPPa*	[A07:20,559,448..20,560,703]
-	*Bra004184*	*AtPPb*	[A07:20,565,349..20,566,352]
+	*Bra004185*	*SP6*	[A07:20,568,308..20,575,068]
+	*Bra004186*	*SP7*	[A07:20,575,743..20,576,853]

*annotation from Wang et al. 2011

Table S3. Published *SRK*, *SLG* and *SRK* paralogous sequences (*SUI, SLR2*, *ARK3*) used in the phylogenetic analysis of the *SRK* sequences among the Brassiceae tribe. B. oleracea: *Brassica oleracea*; *B. rapa*: *Brassica rapa*; *B. napus*: *Brassica napus*; *R. sativus*: *Raphanus sativus*.

Sequence	Species	sequence ID	No. accession/reference
PSRK	*Sisymbrium irio*	SiPSRK	Vekemans et al. 2014
PSRK	*Schrenkiella parvula*	SpPSRK	Vekemans et al. 2014
PSRK	*Eutrema salsugineum*	EsPSRK	Vekemans et al. 2014
SRK	*B. oleracea*	BoSRK01_I	AB054706.1
SRK	*B. oleracea*	BoSRK02_II	AJ306590.1
SRK	*B. oleracea*	BoSRK03_I	X79432.1
SRK	*B. oleracea*	BoSRK04_I	AB298890.1
SRK	*B. oleracea*	BoSRK05_II	Y18259.1
SRK	*B. oleracea*	BoSRK06_I	M76647.1
SRK	*B. oleracea*	BoSRK07_I	AB180898.1
SRK	*B. oleracea*	BoSRK08_I	AB054708.1
SRK	*B. oleracea*	BoSRK11_I	AB054709.1
SRK	*B. oleracea*	BoSRK12_I	AB180901.1
SRK	*B. oleracea*	BoSRK13b_I	EU180597.1
SRK	*B. oleracea*	BoSRK14_I	AB298891.1
SRK	*B. oleracea*	BoSRK15_II	AB180903.1
SRK	*B. oleracea*	BoSRK16_I	AB054710.1
SRK	*B. oleracea*	BoSRK18_I	AB032473.1
SRK	*B. oleracea*	BoSRK20_I	AB054711.1
SRK	*B. oleracea*	BoSRK23_I	AB013720.1
SRK	*B. oleracea*	BoSRK24_I	AB054712.1
SRK	*B. oleracea*	BoSRK25_I	AB054713.1
SRK	*B. oleracea*	BoSRK28_I	AB190355.1
SRK	*B. oleracea*	BoSRK29_I	Z30211.1
SRK	*B. oleracea*	BoSRK32_I	AB050482.1
SRK	*B. oleracea*	BoSRK33_I	AB054714.1
SRK	*B. oleracea*	BoSRK35_I	AB054715.1
SRK	*B. oleracea*	BoSRK36_I	AB054716.1
SRK	*B. oleracea*	BoSRK38_I	AB054717.1
SRK	*B. oleracea*	BoSRK39_I	AB054718.1
SRK	*B. oleracea*	BoSRK45_I	AB054719.1
SRK	*B. oleracea*	BoSRK50_I	AB054720.1
SRK	*B. oleracea*	BoSRK51_I	AB054721.1
SRK	*B. oleracea*	BoSRK52_I	AB298901.1
SRK	*B. oleracea*	BoSRK57_I	AB054722.1
SRK	*B. oleracea*	BoSRK58_I	AB054723.1
SRK	*B. oleracea*	BoSRK60_I	AB032474.1
SRK	*B. oleracea*	BoSRK61_I	AB298902.1
SRK	*B. oleracea*	BoSRK62_I	AB054724.1
SRK	*B. oleracea*	BoSRK63_I	Z18921.1
SRK	*B. oleracea*	BoSRK64_I	AB054725.1
SRK	*B. oleracea*	BoSRK65_I	AB054726.1
SLG	*B. oleracea*	BoSLG01	D85198.1

SLG	*B. oleracea*	BoSLG03	X79431.1
SLG	*B. oleracea*	BoSLG05	X65814.1
SLG	*B. oleracea*	BoSLG07	D85199.1
SLG	*B. oleracea*	BoSLG08	AB054727.1
SLG	*B. oleracea*	BoSLG09	D85200.1
SLG	*B. oleracea*	BoSLG12	D85201.1
SLG	*B. oleracea*	BoSLG14	D85228.1
SLG	*B. oleracea*	BoSLG16	D85202.1
SLG	*B. oleracea*	BoSLG17	D85203.1
SLG	*B. oleracea*	BoSLG18	AB032471.1
SLG	*B. oleracea*	BoSLG20	AB054728.1
SLG	*B. oleracea*	BoSLG22	D85229.1
SLG	*B. oleracea*	BoSLG23	AB013719.1
SLG	*B. oleracea*	BoSLG25	D85204.1
SLG	*B. oleracea*	BoSLG28	D85205.1
SLG	*B. oleracea*	BoSLG31	AB054729.1
SLG	*B. oleracea*	BoSLG32	D88765.1
SLG	*B. oleracea*	BoSLG33	AB054730.1
SLG	*B. oleracea*	BoSLG35	D85206.1
SLG	*B. oleracea*	BoSLG36	AB054731.1
SLG	*B. oleracea*	BoSLG39	D85207.1
SLG	*B. oleracea*	BoSLG45	AB054732.1
SLG	*B. oleracea*	BoSLG46	D85208.1
SLG	*B. oleracea*	BoSLG50	AB054733.1
SLG	*B. oleracea*	BoSLG51_I	D85209.1
SLG	*B. oleracea*	BoSLG52	D85210.1
SLG	*B. oleracea*	BoSLG57	AB054734.1
SLG	*B. oleracea*	BoSLG58	AB054735.1
SLG	*B. oleracea*	BoSLG60	AB032472.1
SLG	*B. oleracea*	BoSLG62	AB054736.1
SLG	*B. oleracea*	BoSLG63	D85211.1
SLG	*B. oleracea*	BoSLG64	D85212.1
SLG	*B. oleracea*	BoSLG65	AB054737.1
SRK	*B. rapa*	BrSRK40_II	AB211197.1
SRK	*B. rapa*	BrSRK44_II	AB211198.1
SRK	*B. rapa*	BrSRK08_I	D38563.1
SRK	*B. rapa*	BrSRK09_I	D30049.1
SRK	*B. rapa*	BrSRK12_I	D38564.2
SRK	*B. rapa*	BrSRK21_I	AB270775.
SRK	*B. rapa*	BrSRK22_I	AB054061.1
SRK	*B. rapa*	BrSRK25_I	AB298875.1
SRK	*B. rapa*	BrSRK26_I	AB054691.1
SRK	*B. rapa*	BrSRK27_I	AB054692.1
SRK	*B. rapa*	BrSRK29_II	AB008191.1
SRK	*B. rapa*	BrSRK30_I	AB054693.1
SRK	*B. rapa*	BrSRK32_I	AB054694.1
SRK	*B. rapa*	BrSRK33_I	AB054695.1
SRK	*B. rapa*	BrSRK34_I	AB054696.1
SRK	*B. rapa*	BrSRK35_I	AB054697.1

SRK	*B. rapa*	BrSRK36_I	AB054698.1
SRK	*B. rapa*	BrSRK37_I	AB054699.1
SRK	*B. rapa*	BrSRK41_I	AB054700.1
SRK	*B. rapa*	BrSRK45_I	AB012106.1
SRK	*B. rapa*	BrSRK46_I	AB180897.1
SRK	*B. rapa*	BrSRK47_I	AB050483.1
SRK	*B. rapa*	BrSRK48_I	AB054701.1
SRK	*B. rapa*	BrSRK49_I	AB054702.1
SRK	*B. rapa*	BrSRK52_I	AB054703.1
SRK	*B. rapa*	BrSRK53_I	AB298884.1
SRK	*B. rapa*	BrSRK54_I	AB298592.1
SRK	*B. rapa*	BrSRK55_I	AB298885.1
SRK	*B. rapa*	BrSRK56_I	AB298886.1
SRK	*B. rapa*	BrSRK60_II	AB097116.1
SRK	*B. rapa*	BrSRK61_I	AB298887.1
SRK	*B. rapa*	BrSRK99_I	AB054704.1
SLG	*B. rapa*	BrSLG09_I	D30050.1
SLG	*B. rapa*	BrSLG21_I	D85213.1
SLG	*B. rapa*	BrSLG22_I	AB054060.1
SLG	*B. rapa*	BrSLG25_I	D85214.1
SLG	*B. rapa*	BrSLG26_I	D85215.1
SLG	*B. rapa*	BrSLG27_I	D85216.1
SLG	*B. rapa*	BrSLG29_II	AB008190.1
SLG	*B. rapa*	BrSLG30_I	D85217.1
SLG	*B. rapa*	BrSLG34_I	D85218.1
SLG	*B. rapa*	BrSLG35_I	D85219.1
SLG	*B. rapa*	BrSLG37_I	D85220.1
SLG	*B. rapa*	BrSLG38_I	D85221.1
SLG	*B. rapa*	BrSLG40_II	AB054058.1
SLG	*B. rapa*	BrSLG41_I	D85222.1
SLG	*B. rapa*	BrSLG44_II	AB054059.1
SLG	*B. rapa*	BrSLG45_I	AB012105.1
SLG	*B. rapa*	BrSLG46_I	D85224.1
SLG	*B. rapa*	BrSLG47_I	AB054705.1
SLG	*B. rapa*	BrSLG48_I	D85225.1
SLG	*B. rapa*	BrSLG49_I	D85226.1
SLG	*B. rapa*	BrSLG52_I	AB054815.1
SLG	*B. rapa*	BrSLG54_I	AB298593.1
SLG	*B. rapa*	BrSLG60_II	AB097116.1
SLG	*B. rapa*	BrSLG99_I	D85227.1
SRK	*B. napus*	BnSRK01_I	AB270767.1
SRK	*R. sativus*	RsSRK19	Kitashiba et al. 2014
SRK	*R. sativus*	RsSRKa	GQ121139.1
SLG	*R. sativus*	RsSLG01	AY052572.1
SLG	*R. sativus*	RsSLG02	AY052573.1
SLG	*R. sativus*	RsSLG04	AY052577.1
SLG	*R. sativus*	RsSLG05	AY052578.1
SLG	*R. sativus*	RsSLG06	AY052574.1
SLG	*R. sativus*	RsSLG07	AY052575.1
SLG	*R. sativus*	RsSLG10	AY052576.1

SRK paralog	*B. rapa*	BrSLR1_p	[A03:16373236..16374273]
SRK paralog	*B. rapa*	BrARK3_p	[A01:6114508..6118190]
SRK paralog	*B. rapa*	BrSUI_p	[A04:17228882..17234192]
SRK paralog	*B. rapa*	BrSUI2_p	Scaffold19561:62..2821
SRK paralog	*B. rapa*	BrUBOX_p	[A01:6112142..6113263]
SRK paralog	*B. rapa*	BrSLL2_p	[A07:20503896..20506112]
SRK paralog	*B. rapa*	BrSP6_p	[A07:20568308..20575068]

References

Altschul SF, Gish W, Miller W, Myers EW, Lipmanl DJ. 1990. Basic Local Alignment Search Tool. J. Mol. Evol. 215:403–410.

Arias T, Pires JC. 2012. A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae : Brassiceae): Novel clades and potential taxonomic implications. Taxon 61:980–988.

Boaz M, Plitmann U, Heyn CC. 1990. The ecogeographic distribution of breeding systems in the cruciferae (Brassicaceae) of Israel. Isr. J. Bot. 39:31–42.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Comparative N, Program S, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN : efficient tools for large-scale multiple alignment of genomic DNA outline of algorithms. Genome Res. 13:721–731.

Castric V, Vekemans X. 2007. Evolution under strong balancing selection: how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae? BMC Evol. Biol. 7.

Chantha S-C, Herman AC, Platts AE, Vekemans X, Schoen DJ. 2013. Secondary evolution of a self-incompatibility locus in the Brassicaceae genus Leavenworthia. PLoS Biol. 11:e1001560.

Cheng F, Liu S, Wu J, Fang L, Sun S, Liu B, Li P, Hua W, Wang X. 2011. BRAD , the genetics and genomics database for Brassica plants. Plant Biol. 11.

Cui Y, Brugière N, Jackman L, Bi Y, Rothstein SJ. 1999. Structural and transcriptional comparative analysis of the S locus regions in two self-incompatible Brassica napus lines. Plant Cell 11:2217–2231.

Dubchak I. 2007. Comparative analysis and visualization of genomic sequences using VISTA browser and associated computational tools. Methods Mol. Biol. 395:3–16.

Edgar RC. 2004. MUSCLE : multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Edh K, Widen B, Ceplitis A. 2009. Molecular population genetics of the SRK and SCR Self-Incompatibility genes in the wild plant species Brassica cretica (Brassicaceae). Genetics 181:985–995.

Fléchon L, Poux C, Vekemans X. 2012. Identification de l'origine d'un goulot d'étranglement ancien au locus d'auto- incompatibilité dans la tribu des Brassiceae (Brassicaceae). p.35.

Fobis-Loisy I, Gaude T. 2004. Molecular evolution of the S Locus controlling mating in the Brassicaceae. Plant Biol. 6:109–118.

Fujimoto R, Okazaki K, Fukai E, Kusaba M, Nishio T. 2006. Comparison of the genome structure of the self-incompatibility (S) locus in interspecific pairs of S haplotypes. Genetics 173:1157–1167.

Fukai E, Fujimoto R, Nishio T. 2003. Genomic organization of the S core region and the S flanking regions of a class-II S haplotype in Brassica rapa. Mol. Genet. Genomics 269:361–369.

Goubet P, Bergès H, Bellec A, Prat E, Helmstetter N, Mangenot S, Gallina S, Holl A-C, Fobis-Loisy I, Vekemans X, et al. 2012. Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. PLoS Genet. 8:e1002495.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies. Syst. Biol. 52:696–704.

Kim D, Jung J, Yeon-Ok C, Kim S. 2016. Development of a system for S locus haplotyping based on the polymorphic SLL2 gene tightly linked to the locus determining self-incompatibility in radish (Raphanus sativus L.). Euphytica 209:525–535.

Kimura R, Sato K, Fujimoto R, Nishio T. 2002. Recognition specificity of self-incompatibility maintained after the divergence of Brassica oleracea and Brassica rapa. Plant J. 29:215–223.

Kitashiba H, Li F, Hirakawa HI, Kawanabe T, Zou Z, Hasegawa YO, Tonosaki KA, Shirasawa SA, Fukushima AKI, Yokoi SH, et al. 2014. Draft sequences of the radish (Raphanus sativus L.) genome. DNA Res. 21:481–490.

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

Kusaba M, Nishio T, Satta Y, Hinata K, Ockendon D. 1997. Striking sequence similarity in inter- and intra-specific comparisons of class I SLG alleles from Brassica oleracea and Brassica campestris : Implications for the evolution and recognition mechanism. Proc. Natl. Acad. Sci. U. S. A. 94:7673–7678.

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14:1–14.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2 : New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772–773.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–360.

Leducq J, Gosset CC, Gries R, Calin K, Schmitt É, Castric V, Vekemans X. 2014. Self-Incompatibility in Brassicaceae : identification and characterization of SRK-Like Sequences linked to the S-Locus in the tribe Biscutelleae. G3(Bethesda) 4:983–992.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows – Wheeler transform. Bioinformatics 25:1754–1760.

Lim S, Cho H, Lee S, Cho Y, Kim B. 2002. Identification and classification of S haplotypes in Raphanus sativus by PCR-RFLP of the S locus glycoprotein (SLG) gene and the S locus receptor kinase (SRK) gene. Theor 104:1253–1262.

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5:1–11.

Lysak MA, Cheung K, Kitschke M, Bures P. 2007. Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. Plant Physiol. 145:402–410.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. Genome Res. 15:516–525.

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S. 2014. Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish Raphanus raphanistrum and Three Other Brassicaceae Species. Plant Cell 26:1925–1937.

Nasrallah JB. 2017. Plant mating systems : self-incompatibility and evolutionary transitions to self-fertility in the mustard family. Curr. Opin. Genet. Dev. 47:54–60.

Otto SP. 2007. The Evolutionary Consequences of Polyploidy. Cell 131:452–462.

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien M-A. 2010. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol. 186:37–45.

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien M-A. 2010. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol. 186:37–45.

Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, et al. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. Genome 15:R77.

R development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Ramsey J, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst. 29:467–501.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE : Multiple Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. PLoS One 6:e22594.

Ronquist FR, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2 : Efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Sato K, Nishio T, Kimura R, Kusaba M, Suzuki T, Hatakeyama K, Ockendon DJ, Satta Y. 2002. Coevolution of the S-Locus genes SRK, SLG and SP11/SCR in Brassica oleracea and B. rapa. Genetics 162:931–940.

Schierup MH, Mable BK, Awadalla P, Charlesworth D. 2001. Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of Arabidopsis lyrata. Genetics 158:387–399.

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

Schranz ME, Lysak MA, Mitchell-Olds T. 2006. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. 11:535–542.

Shiba H, Kenmochi M, Sugihara M, Iwano M, Kawasaki S, Suzuki G, Watanabe M, Isogai A, Takayama S. 2003. Genomic organization of the S-Locus region of Brassica. Biosci. Biotechnol. Biochem. 67:622–626.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57:758–771.

Stamatakis A. 2006. RAxML-VI-HPC : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stamatakis A. 2014. RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Suzuki T, Kusaba M, Matsushita M, Okazaki K, Nishio T. 2000. Characterization of Brassica S-haplotypes lacking S-locus glycoprotein. FEBS Lett. 482:102–108.

Takasaki T, Hatakeyama K, Suzuki G, Watanabe M, Isegai A, Hinata K. 2000. The S receptor kinase determines self-incompatibility in Brassica stigma. Nature 403:913–916.

Takayama S, Isogai A. 2005. Self-incompatibility in plants. Annu. Rev. Plant Biol. 56:467–489.

Takuno S, Fujimoto R, Sugimura T, Sato K, Okamoto S, Zhang S, Nishio T. 2007. Effects of recombination on hitchhiking diversity in the Brassica self-incompatibility locus complex. Genetics 177:949–958.

Tarutani Y, Shiba H, Iwano M, Kakizaki T, Suzuki G, Watanabe M, Isogai A, Takayama S. 2010. Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. Nature 466:983–987.

Vekemans X, Poux C, Goubet PM, Castric V. 2014. The evolution of selfing from outcrossing ancestors in Brassicaceae: what have we learned from variation at the S-locus? J. Evol. Biol. 27:1372–1385.

Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. Ann. Bot. 120:195–207.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43:1035–1040.

Warwick SI, Sauder CA. 2005. Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trn L intron sequences. Can. J. Bot. 83:467–483.

Wright S. 1939. The distribution of self-sterility alleles in populations. Genetics 24:538–552.

Yasuda S, Wada Y, Kakizaki T, Tarutani Y, Miura-uno E, Murase K, Fujii S, Hioki T, Shimoda T, Takada Y, et al. 2016. A complex dominance hierarchy is controlled by polymorphism of small RNAs and their targets. Nat. Plants 3:1–5.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Gil L, et al. 2018. Ensembl 2018. Nucleic Acids Res. 46:754–761.

DISCUSSION ET PERSPECTIVES

Développement d'une approche méthodologique pour inférer l'histoire évolutive de la tribu mésopolyploïde des Brassiceae

Dans le premier chapitre de cette thèse, nous avons présenté le développement d'une nouvelle méthodologie permettant d'assigner les homéologues à leur espèce parentale d'origine chez les Brassiceae, afin de construire une phylogénie d'espèces sur laquelle chacune des lignées parentales apparaît. Ce résultat représente une nouveauté par rapports aux méthodes actuellement disponibles qui analysent des données de séquençage à haut débit (Yang and Smith 2014). Notre méthode utilise une base de données contenant l'ensemble des gènes nucléaires retenus à l'état de triplets (une copie présente dans chacun des trois sous-génomes) chez *Brassica rapa* et *Brassica oleracea*, à partir de laquelle les copies homéologues des gènes sont assignées à leur espèce parentale d'origine. Il est donc ainsi possible de reconstruire les relations phylogénétiques entre les lignées parentales et l'ensemble des espèces actuelles étudiées. L'avantage de cette méthode réside aussi dans le fait qu'elle fonctionne avec des données aussi lacunaires que les génomes de faible couverture et les assemblages de transcriptomes qui sont par définition incomplets à cause i) de la variation temporelle et tissulaire du taux d'expression des gènes, ii) de phénomènes de « silencing » et iii) de problèmes d'assemblage des reads et/ou de couverture.

Cependant, notre méthode se base entièrement sur l'annotation d'un ou plusieurs génomes de référence (ici *Brassica rapa* et *Brassica oleracea*) dont les blocs génomiques ancestraux doivent avoir été correctement assignés à leur sous-génome d'origine (Liu et al. 2014). Nous avons relevé plusieurs discordances entre les différentes annotations indépendantes qui ont été proposées dans la littérature, en fonction des approches utilisées par les auteurs, ce qui suggère de rester prudent quant à nos interprétations. Notamment ces différences d'assignation des copies homéologues sont liées à la manière dont sont traités les éventuels évènements de conversion génique entre copies homéologues (Murat et al. 2015). Toutefois, nous avons trouvé des résultats similaires au niveau de la reconstruction phylogénétique des clades de la tribu des Brassiceae (bien qu'avec des supports de nœuds ancestraux plus faibles), en utilisant uniquement les triplets annotés de façon identique entre deux études distinctes (Liu et al. 2014; Murat et al. 2015). Il pourrait être intéressant de réaliser une étude par simulation numérique du processus d'évolution de sous-génomes après un évènement d'allopolyploïdie, en incluant ou non la possibilité de conversion génique, et

ensuite de tester différentes méthodes de reconstructions phylogénétiques, de manière à tester la robustesse de l'approche proposée.

Résolution de la topologie des clades de la tribu des Brassiceae

Grâce au développement de notre approche méthodologique basée sur la construction d'arbres de gènes homologues visant à inférer les relations d'orthologie, nous avons pu établir la toute première phylogénie nucléaire des différents clades de la tribu des Brassiceae à partir d'un grand nombre de gènes. L'obtention d'une telle phylogénie était particulièrement importante puisque de nombreuses études basées sur des marqueurs chloroplastiques, mitochondriaux, ou nucléaires (ITS) (Warwick and Black 1991; Pradhan et al. 1992; Warwick and Black 1993; Warwick and Black 1994; Warwick and Black 1997; Warwick and Sauder 2005; Hall et al. 2011; Arias and Pires 2012; Willis et al. 2014) ont donné des résultats souvent incongruents. Il est particulièrement rassurant de noter que les phylogénies, au sein des Brassiceae, obtenues pour les trois sous-génomes sont fortement concordantes.

Nous avons aussi dressé une phylogénie d'espèces à partir des régions codantes du génome chloroplastique, dans le but de confronter les topologies nucléaires et chloroplastiques qui se sont avérées être partiellement différentes. Par exemple, nos résultats confirment certains travaux précédents indiquant que la position phylogénétique du genre Raphanus (radis) est distincte entre le génome chloroplastique (co-ségrége avec les espèces du clade Oleracea), et le génome nucléaire (co-ségrége avec les espèces du clade Nigra). Cependant, bien qu'incluant six des huit clades reconnus de Brassiceae, notre analyse n'inclut qu'un faible nombre d'espèces (un à trois représentants par clade). Il serait donc intéressant de conforter nos résultats avec un jeu de données plus conséquent, incorporant divers genres au sein des différents clades de Brassiceae, notamment des genres rapportés comme étant polyphylétiques alors qu'on les pensait monophylétiques (Al-Shehbaz 2012). L'obtention récente de 3 transcriptomes de l'espèce *Crambe maritima* permettra de rajouter un représentant du clade « Crambe », et donc d'être quasiment exhaustif dans notre échantillonnage des clades de Brassiceae (sept sur les huit).

Evaluation de l'ordre des évènements de polyploïdisation à l'origine des Brassiceae

Selon le modèle d'allohexaploïdie en deux étapes ("two-step model of genome merging") proposé par Tang et al. (2012) pour la lignée Brassica, les génomes parentaux MF1 et MF2 auraient d'abord fusionné pour donner une lignée tétraploïde qui aurait elle aussi fusionné, après quelques réarrangements génomiques et pertes de gènes dans l'un ou l'autre de ces deux sous-génomes, avec une espèce diploïde apportant le génome LF. Cette hypothèse a

été proposée afin d'expliquer les pertes de gènes et l'expression génique biaisée entre sous-génomes, avec des niveaux d'expression et de rétention de gènes plus élevés retrouvés dans le sous-génome LF (Wang et al. 2011).

Les biais dans les pertes de gènes entre sous-génomes ont été rapportés chez de nombreuses espèces allopolyploïdes et il a été proposé qu'ils soient consécutifs au biais d'expression entre homéologues (« homoeolog expression bias »). Le biais d'expression entre les homéologues peut favoriser («unbalanced homoeolog expression bias ») ou non (« balanced homoeolog expression bias ») l'un des sous-génomes parentaux, et ces deux types de biais d'expression ont été rapportés chez de nombreux allopolyploïdes tels que chez Gossypium (Hu et al. 2015; Rapp et al. 2009; Renny-Byfield et al. 2015; Yang et al. 2006; Yoo et al. 2013; mais voir Zhang et al. 2015), Triticum (Leach et al. 2014; Pfeifer et al. 2014), Zea (Schnable et al. 2011), Tragopogon (Buggs et al. 2010; Koh et al. 2010), Arabidopsis (Chang et al. 2010), Brassica (Cheng et al. 2012, Liu et al. 2014, Parkin et al. 2014), Mimulus (Edger et al. 2017), Spartina (Chelaifa et al. 2010) et le récent allotétraploïde Capsella bursa-pastoris (Douglas et al. 2015). Le biais d'expression équilibré entre homéologues semble être marginalement moins fréquent que le biais déséquilibré chez les allopolyploïdes, mais aucune conclusion formelle ne peut être tirée. Cependant, une étude récente dans le genre Mimulus a montré que la dominance entre sous-génomes, décrite chez l'espèce néohexaploïde *M. peregrinus*, était associée à la densité et à la méthylation des éléments transposables dans les génomes parentaux et non pas à l'ordre des évènements de fusions de génomes (Edger et al. 2017). En effet, les interactions entre les génomes parentaux des allopolyploïdes peuvent induire des changements épigénétiques, comme la méthylation de l'ADN (Chen 2007), et de telles modifications ont été documentées chez de nombreuses plantes (Salmon et al. 2005; Parisod et al. 2009; Song and Chen 2015). De tels changements asymétriques dans les profils de méthylation de l'ADN pourraient expliquer, en tout ou partie, le biais d'expression entre homéologues systématiquement observé chez plusieurs espèces allopolyploïdes. En accord avec cette hypothèse, Woodhouse et al. (2014) ont montré, chez *B. rapa*, que le sous-génome dominant LF a une densité en éléments transposables plus faible que les sous-génomes MF1 et MF2, ce qui induit des différences de profils de méthylation de l'ADN entre les sous-génomes. Ces résultats suggèreraient donc que les patrons de dominance entre sous-génomes ne fournissent pas d'informations sur la chronologie de la fusion des génomes chez les espèces allopolyploïdes. Il a également été documenté que l'hybridation a un impact plus important sur l'expression des gènes que le doublement du génome (Chelaifa et al. 2010, Ainouche et al. 2012) et que c'est davantage la fusion des génomes que la polyploïdie en tant

que telle qui est responsable des profils non additifs de méthylation (chez Spartina) (Salmon et al. 2005; Parisod et al. 2009). Chez les autopolyploïdes, aucun marquage épigénétique nettement différent n'est attendu en raison de leur mode de formation (impliquant des génomes quasi identiques), ce qui pourrait expliquer l'absence de biais d'expression entre homéologues chez les autopolyploïdes (Garsmeur et al. 2014). Ainsi, la mesure du biais d'expression entre homéologues (par exemple en utilisant des données de transcriptomique) pourrait être un indicateur utile pour déterminer si une espèce polyploïde provient d'auto- ou d'allopolyploïdisation, même si cette approche ne pourrait s'appliquer qu'aux néo- et mésopolyploïdes dont les segments dupliqués peuvent être assignés à des génomes ancestraux spécifiques.

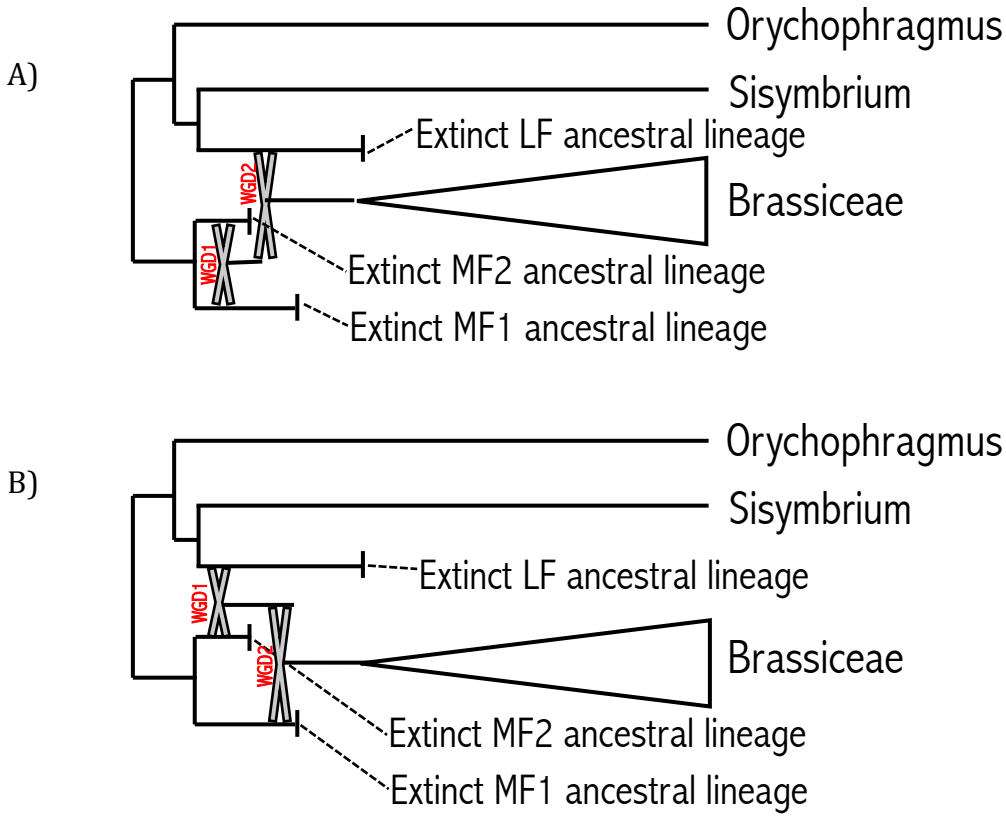


FIGURE 16. Deux scénarios alternatifs d'allohexaploïdie en deux étapes à l'origine de la tribu des Brassiceae. Le premier scenario (A) correspond au célèbre modèle d'allohexaploïdie en deux étapes (Wang et al. 2011, Tang et al. 2012) : une première hybridation allopolyploïde entre les espèces parentales diploïdes MF1 et MF2 suivie d'une seconde hybridation allopolyploïde entre l'hybride tétraploïde nouvellement formé et l'espèce parentale diploïde LF. Le deuxième scenario (B) correspond à une première hybridation allopolyploïde entre les espèces parentales diploïdes LF et MF2 (ou MF1, scénario non représenté) suivie d'une seconde hybridation allopolyploïde entre l'hybride tétraploïde nouvellement formé et l'espèce parentale diploïde MF1 (ou MF2, scénario non représenté). Dans les deux scénarios ici présentés, les positions phylogénétiques de Sisymbrium et du genre Orychophragmus comme clades sœurs de l'espèce parentale LF correspondent aux résultats originaux obtenus dans la présente étude.

Notre analyse phylogénomique montre que les espèces parentales MF1 et MF2 sont phylogénétiquement proches l'une de l'autre, et que l'ancêtre LF provient d'une lignée plus divergente. Ce patron est plutôt en accord avec l'hypothèse de fusion des génomes en deux étapes (« two-step model of genome merging ») (FIGURE 16A). Cependant nous ne pouvons pas exclure un scénario dans lequel la lignée LF et la lignée MF1 (ou MF2) seraient à l'origine de la première hybridation allopolyploïde (FIGURE 16B). Dans ce cas, la dominance du sous génome LF serait associée à différents niveaux de méthylation des sous-génomes, et non pas à l'ordre chronologique des hybridations (Lee and Chen 2001; Chen 2007; Phillips 2008; Schnable et al. 2011; Edger et al. 2017). Cependant il reste difficile d'estimer l'importance respective de la méthylation différentielle et du temps écoulé depuis l'évènement de polyploïdisation dans l'évolution du fractionnement du génome des Brassiceae.

Dans l'hypothèse où l'événement de WGT des Brassiceae se serait produit selon le scénario classique d'allohexaploïdie en deux étapes (Figure 16A) impliquant une première fusion des génomes MF1 et MF2, la question reste ouverte de savoir si ce premier événement de fusion de génomes correspond réellement à une allopolyploïdie ou à une autopolyploïdie. En effet, à l'exception de l'étude de Cheng et al. (2012) qui rapporte que le nombre de gènes pour lesquels la copie MF1 s'exprime plus que la copie MF2 est significativement plus élevé que le nombre de gènes pour lesquels la copie MF2 s'exprime plus, il n'a jamais été fait mention d'une différence significative dans l'expression et la rétention de gènes entre les sous-génomes MF1 et MF2.

L'hypothèse d'un premier événement d'autopolyploïdie chez l'ancêtre des « lignées » actuelles MF1 et MF2 suivi d'une hybridation allopolyploïde entre l'individu autotétraploïde nouvellement formé et l'espèce parentale LF reste donc entièrement plausible. Si tel a été le cas, l'autotétraploïde MF1-MF2 a probablement été sujet à des phénomènes de recombinaison entre chromosomes homéologues, puisque les autopolyploïdes – qui contiennent des jeux de chromosomes identiques – présentent souvent des appariements de plus de deux chromosomes (multivalents) durant la méiose. Ce qui n'est pas le cas pour les allopolyploïdies – contenant des jeux de chromosomes ayant divergé bien avant l'événement d'hybridation allopolyploïde – qui ont tendance à former des bivalents durant la méiose, un type d'appariement chromosomique semblable à celui observé chez les diploïdes (Glover et al. 2016). Murat et al. (2016) ont effectivement inféré des évènements de recombinaison entre les sous-génomes MF1 et MF2 qu'ils expliquent par le fait que les sous-génomes MF1 et

MF2 auraient fusionnés dans un premier temps et auraient évolués séparément du génome LF pendant un certain temps. Les données disponibles à ce jour ne nous permettent pas de discerner clairement les sous génomes impliqués dans le premier événement de polyploïdisation.

Évènement de WGD indépendant chez Orychophragmus violaceus

Nous avons vu dans le second chapitre de cette thèse que l'espèce tétraploïde *Orychophragmus violaceus* semble avoir subi un événement de WGD indépendant de ceux subis par la lignée des Brassiceae, contrairement à ce qu'avait supposé Lysak et al. (2007). En d'autres termes, la lignée Orychophragmus n'est pas le représentant actuel de l'hybride tétraploïde (ou l'autotétraploïde) MF1-MF2. Bien que nos données comportaient certainement quelques biais qu'il faudra estimer, notamment la présence de transcrits alternatifs dans les arbres phylogénétiques, parmi l'ensemble des arbres phylogénétiques d'homéologues qui pouvaient nous apporter une information sur l'événement de WGD de *O. violaceus*, seuls 5% présentaient une topologie en faveur d'un partage d'un événement de WGD entre la lignée Orychophragmus et la tribu des Brassiceae (10,8% si on élimine les arbres dans lesquels les copies d'*O. violaceus* sont monophylétiques et qui comportent potentiellement un seul des deux homéologues). Ce résultat suggère fortement que l'événement de WGD de la lignée Orychophragmus est indépendant, et par conséquent il serait intéressant de déterminer la nature de cet événement (auto- versus allopolyploïdie) et les relations phylogénétiques entre les lignées parentales, les lignées LF, MF1 et MF2 et les espèces actuelles dans le cas où il s'agit d'un événement d'allopolyploïdie. Pour cette étude, il serait nécessaire d'élargir le jeu de données en utilisant, en plus des triplets d'homéologues, les doublets d'homéologues présents chez *Brassica rapa* et *Brassica oleracea* (Liu et al. 2014) et de n'utiliser que ces deux espèces comme représentants des Brassiceae. Ainsi, nous pourrions obtenir un plus grand nombre de gènes pour lesquels nous disposons de deux copies par individus de *O. violaceus*, et l'analyse des arbres de gènes correspondant pourrait nous permettre de décrire plus clairement l'événement de WGD de la lignée Orychophragmus. L'obtention d'un génome assemblé chez Orychophragmus constituerait néanmoins la meilleure solution pour affiner l'étude de l'évènement de polyploidisation, car il n'est pas certain que le degré de conservation des paires d'homéologues soit corrélé entre évènements indépendants de polyploïdie, bien que certains travaux dans le domaine semblent le suggérer (Moghe et al. 2014, Mandakova et al. 2017).

Datation de l'évènement d'allohexaploïdie de la tribu des Brassiceae

A ce jour, les estimations de l'âge de l'évènement de WGT de la tribu des Brassiceae sont comprises entre 15,3 et 29 millions d'années et ont été obtenues à partir des distributions des valeurs de *Ks* entre paires de gènes paralogues au sein du génome de *Brassica rapa* (Kagale et al. 2014), à partir de la divergence entre les doublets d'homéologues dans les génomes de *B. rapa* et de *B. oleracea* (Liu et al. 2014) ou dans les génomes de *Raphanus raphanistrum* et de *B. rapa* (Moghe et al. 2014). Town et al. (2006) ont estimé chez *B. rapa* la divergence de trois segments génomiques paralogues à 29 millions d'années, ce qui correspond à proprement parler à l'âge de la divergence entre les espèces parentales et non pas à l'âge de l'événement de WGT en lui-même. Notre analyse détaillée des relations phylogénétiques entre les trois sous-génomes parentaux et notre large échantillonnage des différents sous-clades de la tribu des Brassiceae pourrait permettre d'obtenir une datation plus précise de l'événement de WGT.

En effet, l'évènement de WGT qui a donné naissance à l'ancêtre commun de l'ensemble des espèces de Brassiceae s'est produit indubitablement après la divergence des trois lignées parentales LF, MF1 et MF2 mais nécessairement avant la diversification de la tribu des Brassiceae. Tous ces évènements peuvent être datés dans notre arbre grâce à des méthodes de datations moléculaires de façon à inférer une fenêtre temporelle dans laquelle l'événement de WGT se serait produit. Pour cela l'arbre phylogénétique utilisé doit être calibré grâce à des informations paléontologiques sur les âges de divergence de certaines lignées dans l'arbre (calibration primaire). Cependant aucun des nœuds de notre arbre ne peut être calibré avec de telles informations et nous devrions donc utiliser des calibrations secondaires, c'est à dire des calibrations non issues de fossiles mais d'estimations moléculaires. Ces estimations sont de fait moins précises que les calibrations primaires puisque elles accumulent les incertitudes liées (i) à l'estimation de l'âge des fossiles utilisés pour obtenir ces calibrations secondaires et (ii) au calcul même de ces datations. Plusieurs points de calibration pourraient être utilisés : les estimations pour la divergence Arabidopsis-Brassiceae (26,9-28 Mya), la divergence Schrenkiella-Brassiceae (18,6-19,4 Mya) et l'âge de l'ancêtre commun le plus récent des clades « Nigra », « Oleracea » et « Cakile » (12,6-13,1 Mya) (Huang et al. 2016). L'événement de WGT se serait donc produit, comme discuté plus haut, après la divergence entre les espèces parentales MF1 et MF2 et avant la diversification des Brassiceae. On s'attend à ce que les âges obtenus soient relativement proches, étant donné la faible longueur des branches qui relient l'ancêtre commun le plus récent des lignées MF1 et MF2 aux Brassiceae. On s'attend également à ce que la divergence entre les trois lignées

parentales se soit produite dans un laps de temps très court, comme indiqué par la longueur des branches de notre arbre phylogénétique mais aussi par le fait qu'on ne retrouve qu'un seul pic dans les distributions de valeurs de *Ks*, là où on devrait en observer deux.

Translocation génomique du locus S chez les membres de la tribu des Brassiceae

Dans le troisième chapitre de cette thèse, nous avons cherché à savoir si la translocation génomique du locus S observée chez *Brassica rapa* (Fobis-Loisy and Gaude 2004; Chantha et al. 2013; Kim et al. 2016) est spécifique à cette espèce, au clade Oleracea ou bien encore à l'ensemble des membres de la tribu des Brassiceae, de façon à pouvoir éventuellement établir un lien entre la translocation génomique du locus S et l'événement de triplication de génome qui a touché l'ancêtre commun de la tribu. Pour cela, nous avons décrit l'organisation génomique de la région étendue du locus S en position « Arabidopsis » (entre les gènes *B80/U-box* et *ARK3*) et en position « Brassica » (entre les gènes *Sll2* et *AtPPa*) dans chacun des trois sous-génomes de *Brassica rapa* et *B. oleracea*. Nous avons ainsi montré que le locus S chez ces deux espèces se trouve dans le sous-génome parental LF, sur le bloc génomique ancestral E, alors qu'il se trouve sur le bloc génomique ancestral U chez Arabidopsis. Ainsi, la translocation génomique du locus S en position « Brassica » précède la divergence de ces deux espèces. Nous avons aussi pu confirmer que la localisation génomique du locus S chez *Raphanus sativus* (clade Nigra) est identique à celle de *B. rapa* et *B. oleracea*. De plus, nous avons également montré que le locus S n'est pas présent en position « Arabidopsis » chez *Brassica nigra* (clade Nigra), ce qui nous permet d'inférer que la translocation génomique du locus S précède la divergence des clades Oleracea et Nigra. Enfin, nous sommes également parvenus à amplifier par PCR la région intergénique comprise entre les gènes *U-Box* et *ARK3* chez les deux espèces *Cakile maritima* (clade Cakile) et *Pscyhine stylosa* (clade Savignya), et les alignements confirment que la région amplifiée est bien homologue à la région intergénique comprise entre *U-Box* et *ARK3* des espèces *B. rapa*, *B. oleracea*, *R. sativus* et *B. nigra*, suggérant que le locus S a aussi été transloqué chez *C. maritima* et *P. stylosa*. Bien que nos données ne nous permettent pas de confirmer la présence du locus S en position « Brassica » chez ces deux espèces, il est fort probable qu'un seul événement de translocation précédant la divergence des quatre clades Savignya, Cakile, Oleracea et Nigra ait eu lieu. L'absence d'amplification chez les clades basaux de Brassiceae (clade Zilla et clade Vella) pourrait ne pas être lié à la présence du locus S en position Arabidopsis, comme le suggère l'absence d'amplification chez deux individus de *Brassica insularis* (clade Oleracea) qui partage *a priori* la translocation du locus S avec les

membres des clades Nigra et Oleracea. Le design de nouvelles amorces plus spécifiques à des positions adaptées pourrait peut-être permettre de résoudre cette question chez les clades basaux des Brassiceae. Néanmoins, et là encore, l'obtention d'un génome assemblé chez un membre de chacun des clades de Brassiceae constituerait la meilleure solution pour affiner l'étude de la localisation génomique du locus S chez les Brassiceae.

L'absence du locus S en position "Arabidopsis" et en position "Brassica" dans les sous-génomes MF1 et MF2 de *B. rapa* et *B. oleracea* suggère que la translocation génomique du locus S, au sein du sous-génome LF, aurait eu lieu après l'événement de WGT, et donc ces deux sous-génomes pourraient avoir perdu leur locus S (en position Arabidopsis) par délétion après le premier évènement d'allopolyploïdie, ou éventuellement ultérieurement. L'observation que les deux gènes flanquants du locus S en position Arabidopsis soient côte à côte dans le sous-génome LF suggère que la translocation ait bien eu lieu à partir du sous-génome LF, car dans les sous-génomes MF1 et MF2 la région délétée est bien plus importante.

Diversité allélique au locus S chez les membres de la tribu des Brassiceae

Une perte drastique de diversité phylogénétique (on parle ici de phylogénie d'allèles à un locus multiallélique) au locus S a été observée dans les genres Brassica et Raphanus (Lim et al. 2002; Sato et al. 2002; Edh et al. 2009). La richesse allélique à ce locus reste forte mais tous les allèles se regroupent dans deux clades (les allèles de classe I et les allèles de classe II, voir section 3.3.2). Par opposition, la diversité phylogénétique au locus S chez Arabidopsis est très élevée, ce qui suggère qu'un goulot d'étranglement majeur aurait eu lieu chez l'ancêtre commun de ces deux genres, suivi d'une re-diversification allélique sous l'effet de la sélection fréquence-dépendante négative agissant sur le système SI. L'objet du troisième chapitre de cette thèse a également été de vérifier si les différents clades de Brassiceae partageaient les mêmes signatures de goulot d'étranglement au locus-S que celles observées chez *Brassica rapa*. Ceci afin de pouvoir éventuellement mettre en relation ce goulot d'étranglement génomique avec l'événement de triplication de génome partagé par l'ensemble des Brassiceae. A cette fin, nous avons caractérisé la diversité allélique du gène *SRK* chez différents membres de la tribu des Brassiceae ainsi que les relations phylogénétiques entre les allèles S. Nos résultats suggèrent que le goulot d'étranglement au locus S précède la diversification des différentes lignées de Brassiceae, puisque seuls des allèles de classe I et de classe II ont été retrouvés chez les différents membres étudiés de la tribu des Brassiceae, tandis qu'un allèle proche de la séquence *SRK* de *Eutrema salsugineum* a été retrouvé chez l'espèce proche *Orychophragmus violaceus*, et par ailleurs un allèle

proche d'autres allèles Arabidopsis a été retrouvé chez une autre espèce proche des Brassiceae, *Sisymbrium irio* (Vekemans et al. 2014). Ainsi, nous avons été amené à confirmer la possible association entre le goulot d'étranglement génomique partagé par l'ensemble des Brassiceae et l'événement de triplication de génome lui aussi partagé par l'ensemble des Brassiceae.

Tandis que le scénario évolutif semble plutôt clair en ce qui concerne les allèles de classe I (un seul allèle chez l'ancêtre commun à partir duquel a eu lieu la re-diversification allélique, avant la divergence des clades de Brassiceae), le scénario évolutif concernant les allèles de classe II ne peut être encore parfaitement dressé. En effet, nous avons retrouvé des séquences (excepté chez *Raphanus sativus*, *Sinapis arvensis*, *Brassica rapa* et *B. oleracea*) qui semblent correspondre à un ou plusieurs gènes paralogues de *SRK* proches du clade des allèles de classe II chez Brassica, qui auraient pu donner naissance à ce clade de classe II par recombinaison ectopique ou conversion génique entre l'un de ces paralogues (de localisation génomique inconnue) et le locus S en position "Brassica" occupé initialement par des allèles de classe I. Dans le genre Brassica, ce nouvel allèle putatif de classe II aurait ensuite subi une diversification allélique, néanmoins bien plus modérée que les allèles de classe I. Sur le plan théorique, il est en effet attendu que le processus de diversification allélique soit plus faible pour des allèles récessifs, tels que les allèles de classe II (Schierup et al. 1997). Toutefois, l'absence d'information concernant la localisation génomique des séquences reconstruites ne permettent pas d'étayer ces hypothèses.

Enfin, nous avons retrouvé un allèle S de classe I différent entre les différentes espèces de Brassiceae auto-compatibles étudiées (*Carrichtera annua*, *Schouwia purpurea*, *Psychine stylosa*). Ces différents allèles représentent sans doute autant de pertes indépendantes du système d'auto-incompatibilité dans ces trois lignées, comme observé auparavant dans d'autres clades de la famille des Brassicaceae (Vekemans et al. 2014).

CONCLUSION

Au cours de cette thèse, nous avons cherché à déterminer si la perte de diversité phylogénétique au locus d'auto-incompatibilité et le changement de localisation génomique du locus S sont liés à l'évènement de triplication du génome qui a touché l'ancêtre commun de la tribu des Brassiceae. Pour répondre à une telle question, nous avons échantillonné différents clades de Brassiceae, de façon à identifier puis mettre en relation la triplication de génome, l'événement de goulot d'étranglement et le changement de localisation génomique du locus-S. Nous avons également cherché à déterminer si le genre mésotétraploïde Orychophragmus partage ou non un WGD avec la tribu des Brassiceae, dont il est proche phylogénétiquement.

Pour cela, nous avons (1) établi une phylogénie robuste des lignées de Brassiceae à partir de gènes nucléaires d'une part et de gènes chloroplastiques d'autre part, (2) trouvé la trace d'un événement de WGT partagé chez les espèces de Brassiceae étudiées et d'un événement de WGD indépendant chez *Orychophragmus violaceus*, (3) appréhendé le système de reproduction des espèces étudiées (SI versus SC), (4) validé la localisation génomique du locus-S chez *Brassica rapa* et *Brassica oleracea* et (5) établi des phylogénies d'allèles *SRK* qui ont permis de mettre en évidence l'événement de goulot d'étranglement à l'échelle de la tribu des Brassiceae. Le premier aspect de la thèse, à savoir l'obtention d'une phylogénie robuste des Brassiceae, était capital pour parvenir à positionner sur l'arbre phylogénétique des espèces les évènements de duplication et de triplication de génome, de translocation et de goulot d'étranglement au locus-S. Nous avons ainsi développé une méthodologie originale pour inférer de façon fiable les relations d'orthologie chez les espèces appartenant à la tribu allo-hexaploïde des Brassiceae puis retracer l'histoire évolutive de cette tribu. Le deuxième chapitre de la thèse nous a permis de valider l'existence d'un événement de triplication de génome commun à l'ensemble des Brassiceae et indépendant de l'événement de duplication de génome décrit chez Orychophragmus. Ces premiers résultats nous ont permis de dessiner quelques attendus théoriques sur la diversité phylogénétique des haplotypes au locus d'auto-incompatibilité. En effet, les Brassiceae devraient partager l'événement de translocation et de goulot d'étranglement au locus-S, alors qu'aucun allèle de classe I ou de classe II ne devrait être retrouvé chez Orychophragmus. Nos premières investigations à partir des transcriptomes ont confirmé les attendus concernant le goulot d'étranglement au locus S qui semble avoir précédé la diversification des différentes lignées de Brassiceae. En revanche, le changement de localisation génomique du locus S n'a pu être décrit que chez les espèces *B. rapa*, *B.*

oleracea, et *R. sativus*. Nous avons cependant pu montrer que le locus S n'est pas présent en position « Arabidopsis » (entre *UBOX* et *ARK3*) chez *C. maritima*, *P. stylosa*, et *B. nigra*, sans pouvoir affirmer qu'il soit présent en position « Brassica ». D'après ces résultats, la translocation génomique du locus S n'est pas restreinte au clade « Oleracea » et aurait eu lieu au moins avant la divergence des clades « Savignya », « Cakile », « Nigra » et « Oleracea ».

Il conviendrait d'obtenir davantage de séquences de *SRK* chez différents clades de Brassiceae ainsi que de plus amples informations quant à la localisation génomique du locus S chez les clades basaux des Brassiceae afin de pouvoir clairement affirmer l'existence d'une association entre l'événement de triplication de génome, la translocation et le goulot d'étranglement au locus S.

Il semblerait que la polyploïdie et les transitions de système de reproduction ne soient pas indépendantes chez les Brassicaceae. En effet, on observe souvent qu'un évènement récent (<100 000 ans) d'allopolyploïdie entre deux espèces parentales –dont au moins une est auto-incompatible– génère un hybride allopolyploïde auto-compatible tel que *Arabidopsis kamchatica* (Tsuchimatsu et al. 2012), *Arabidopsis suecica* (Novikova et al. 2017), *Capsella bursa-pastoris* (Douglas et al. 2015) ou encore *Cardamine flexuosa* (Mandáková et al. 2014). Il semblerait ainsi que l'allopolyploïdie provoque une instabilité des systèmes SI, et cela pourrait s'expliquer de plusieurs manières. Premièrement, le nombre de chromosomes multiplié chez les individus polyploïdes par rapport aux individus parentaux diploïdes et les changements morphologiques et éco-physiologiques pouvant altérer la biologie des polyploïdes sont à l'origine d'un fort isolement reproducteur entre les nouveaux hybrides allopolyploïdes et les espèces parentales, ce qui conduit souvent à une spéciation allopolyploïde (Otto 2007). Cet isolement reproducteur fort est responsable d'un effet de fondation drastique, qui peut avoir des conséquences sur l'ensemble du génome mais plus particulièrement sur les locus très polymorphes tels que les locus d'auto-incompatibilité. L'effet de fondation est à l'origine d'une forte réduction de l'effectif efficace qui implique une très forte limitation en disponibilité de partenaires sexuels et une perte drastique du nombre d'allèles S chez les nouveaux individus fondateurs, ce qui renforce encore davantage la limitation en disponibilité de partenaires sexuels. La sélection pour l'assurance reproductive en condition de faible densité dans la nouvelle lignée polyploïde pourrait ainsi conférer un avantage sélectif aux individus auto-compatibles et conduire, à terme, à la perte totale du système d'auto-incompatibilité dans la nouvelle lignée polyploïde. Deuxièmement, si l'une des espèces parentales impliquée dans l'événement d'allopolyploïdie est auto-compatible, un allèle mutant SC se retrouve mélangé aux quelques allèles S fonctionnels

restants, ce qui peut fortement déstabiliser le SI et plus rapidement conduire la nouvelle lignée polyploïde vers un système auto-compatible. L'évolution vers des taux d'auto-fécondation accrus pourrait totalement isoler reproductivement l'hybride polyploïde de ses parents diploïdes et ainsi promouvoir l'établissement du nouvel hybride polyploïde. Chez les Angiospermes, il a en effet été montré que les taxa polyploïdes présentent en moyenne des taux d'autofécondation plus élevés que les taxa diploïdes, et ce quelque soit la catégorie de traits d'histoire de vie considérée (herbacées annuelles, pérennes ou ligneuses pérennes) (Barringer 2007). La capacité des polyploïdes d'augmenter les niveaux d'autofécondation sans subir les conséquences néfastes de la dépression de consanguinité devrait favoriser l'augmentation des taux d'autofécondation chez les polyploïdes.

En définitive, il semblerait que suite à une hybridation allopolyploïde entre deux espèces parentales diploïdes (SI x SI ou SI x SC), on observe 1) un isolement reproducteur entre l'espèce allopolyploïde nouvellement formée et les espèces parentales, 2) une mutation cassant le SI de façon partielle ou complète, 3) une transition rapide vers l'auto-compatibilité par sélection pour l'assurance reproductive et 4) à terme, un recouvrement possible du SI ou une perte irréversible du SI. Chez les espèces mésopolyploïdes au sein de la famille des Brassicaceae, on observe que le SI a été restauré à plusieurs reprises. Chez les Biscutelleae par exemple, l'espèce mésopolyploïde *B. laevigata subsp laevigata* présente des traces d'un goulot d'étranglement passé au locus d'auto-incompatibilité (Leducq et al. 2014) puisque trois classes d'allèles S ont été retrouvées tandis que nous avons retrouvé deux classes d'allèles S chez les Brassiceae, signature d'un goulot d'étranglement possiblement associé à l'événement de WGT partagé par la tribu. Enfin, chez Leavenworthia qui a subi un événement de WGT, un système homologue au système ancestral des Brassicaceae a évolué à partir de gènes paralogues des gènes *SCR* et *SRK* (Chantha et al. 2013), ce qui suppose que le SI ancestral a été perdu. La polyploïdie peut ainsi conduire à une perte totale du SI ou à une diversification des SI. Le système Leavenworthia est différent et non orthologue au système Arabidopsis tandis que le système Brassica présente des différences fonctionnelles avec le système Arabidopsis puisque des gènes impliqués dans la cascade moléculaire de la réponse SI chez Brassica ne sont pas retrouvés chez Arabidopsis, et que *SLG* semble très important dans le système SI de Brassica bien qu'il soit absent chez Arabidopsis. Chez les Brassicaceae, les évènements de mésopolyploïdie et de néopolyploïdie sont très répandus (Mandáková et al. 2017) et il serait particulièrement pertinent d'analyser la diversité allélique au locus S des espèces polyploïdes afin de pouvoir valider un scénario d'évolution du système d'auto-incompatibilité dans un contexte de polyploïdie.

RÉFÉRENCES

Al-shehbaz IA. 2012. A generic and tribal synopsis of the Brassicaceae (Cruciferae). Taxon 61:931–954.

Ainouche ML, Chelaifa H, Ferreira J, Bellot S, Ainouche A, Salmon A. 2012. Polyploid evolution in Spartina: Dealing with highly redundant hybrid genomes. In P. S. Soltis and D. E. Soltis [eds.], Polyploidy and genome evolution, 225–243. Springer, New York, New York, USA.

Arias T, Pires JC. 2012. A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae : Brassiceae): Novel clades and potential taxonomic implications. Taxon 61:980–988.

Barringer BC. 2007. Polyploidy and self-fertilization in flowering plants. Am. J. Bot. 94:1527–1533.

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu C, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2 : A software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. 10:e1003537.

Buggs RJA, Chamala S, Wu W, Gao L, May GD, Schnable PS, Soltis DE, Soltis PS, Barbazuk WB. 2010. Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. Mol. Ecol. 19:132–146.

Chang PL, Dilkes BP, Mcmahon M, Comai L, Nuzhdin S V. 2010. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol. 11:1–17.

Chantha S-C, Herman AC, Platts AE, Vekemans X, Schoen DJ. 2013. Secondary evolution of a self-incompatibility locus in the Brassicaceae genus Leavenworthia. PLoS Biol. 11:e1001560.

Chelaifa H, Monnier A, Ainouche M. 2010. Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina × townsendii and Spartina anglica (Poaceae). New Phytol. 186:161–174.

Chen ZJ. 2007. Genetic and epigentic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu. Rev. Plant Biol. 58:377–406.

Cheng F, Sun C, Wu J, Schnable J, Woodhouse MR, Liang J, Cai C, Freeling M, Wang X. 2016. Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. New Phytol. 211:288–299.

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One 7:1–9.

Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Hazzouri KM, Wang W, Platts AE, et al. 2015. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. PNAS 112:2806–2811.

Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-W, Bewick AJ, Ji L, Platts AE, Bowman MJ, et al. 2017. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell 29:2150–2167.

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. 2014. Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. 31:448–454.

Glover NM, Redestig H, Dessimoz C. 2016. Homoeologs : What are they and how do we infer them ? Trends Plant Sci. 21:609–621.

Hall JC, Tisdale TE, Donohue K, Wheeler A, Al-yahya MA, Kramer EM. 2011. Convergent evolution of a complex fruit structure in the tribe Brassiceae (Brassicaceae). Am. J. Bot. 98:1989–2003.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements : A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19:1419–1428.

Hu G, Koh J, Yoo M, Chen S, Wendel JF. 2015. Gene-expression novelty in allopolyploid Cotton : A proteomic perspective. Genetics 200:91–104.

Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-shehbaz I, Edger PP, et al. 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol. Biol. Evol. 33:394–412.

Koh J, Soltis PS, Soltis DE. 2010. Homeolog loss and expression changes in natural populations of the recently and repeatedly formed allotetraploid Tragopogon mirus (Asteraceae). BMC Genomics 11:1–16.

Leach LJ, Belfield EJ, Jiang C, Brown C, Mithani A, Harberd NP. 2014. Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. BMC Genomics 15:1–19.

Leducq J, Gosset CC, Gries R, Calin K, Schmitt É, Castric V, Vekemans X. 2014. Self-Incompatibility in Brassicaceae : identification and characterization of SRK-Like Sequences linked to the S-Locus in the tribe Biscutelleae. G3(Bethesda) 4:983–992.

Lee H-S, Chen ZJ. 2001. Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. PNAS 98:6753–6758.

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5:1–11.

Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. 91:3–21.

Mandáková T, Marhold K, Lysak MA. 2014. The widespread crucifer species Cardamine flexuosa is an allotetraploid with a conserved subgenomic structure. New Phytol. 201:982–992.

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S. 2014. Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish Raphanus raphanistrum and three other Brassicaceae species. Plant Cell 26:1925–1937.

Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, Crollius HR, Salse J. 2015. Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol. 16:1–17.

Novikova PY, Tsuchimatsu T, Simon S, Nizhynska V, Voronin V, Burns R, Fedorenko OM, Holm S, Prat E, Marande W, et al. 2017. Genome sequencing reveals the origin of the allotetraploid Arabidopsis suecica. Mol. Biol. Evol. 34:957–968.

Otto SP. 2007. The evolutionary consequences of polyploidy. Cell 131:452–462.

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M-A, Ainouche M. 2009. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol. 184:1003–1015.

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Mayer KFX, Olsen O-A. 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. Science (80-.). 345.

Phillips T. 2008. The role of methylation in gene expression. Nat. Educ. 1:116.

Pradhan AK, Prakash S, Mukhopadhyay A, Pental D. 1992. Phylogeny of Brassica and allied genera based on variation in chloroplast and mitochondrial DNA patterns : molecular and taxonomic classifications are incongruous. Theor. Appl. Genet. 85:331–340.

Rapp RA, Udall JA, Wendel JF. 2009. Genomic expression dominance in allopolyploids. BMC Biol. 7.

Renny-Byfield S, Gong L, Gallagher JP, Wendel JF. 2015. Persistence of subgenomes in paleopolyploid cotton after 60 My of evolution. Mol. Biol. Evol. 32:1063–1071.

Salmon A, Ainouche M, Wendel JF. 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). Mol. Ecol. 14:1163–1175.

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

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. PNAS 108:4069–4074.

Song Q, Chen ZJ. 2015. Epigenetic and developmental regulation in plant polyploids. Curr. Opin. Plant Biol. 24:101–109.

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC. 2012. Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics 190:1563–1574.

Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ, et al. 2006. Comparative genomics of Brassica oleracea and Arabidopsis thaliana reveal gene Loss, fragmentation, and dispersal after polyploidy. Plant Cell 18:1348–1359.

Tsuchimatsu T, Kaiser P, Yew C, Bachelier JB, Shimizu KK. 2012. Recent loss of self-incompatibility by degradation of the male component in allotetraploid Arabidopsis kamchatica. PLoS Genet. 8:e1002838.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43:1035–1040.

Warwick SI, Black LD. 1991. Molecular systematics of Brassica and allied genera (Subtribe Brassicinae , Brassiceae) - chloroplast genome and cytodeme congruence. Theor. Appl. Genet. 82:81–92.

Warwick SI, Black LD. 1993. Molecular relationships in subtribe Brassicinae (Cruciferae, tribe Brassiceae). Can. J. Bot. 71:906–918.

Warwick SI, Black LD. 1994. Evaluation of the subtribes Moricandiinae, Savignyinae, Vellinae, and Zillinae (Brassicaceae, tribe Brassiceae) using chloroplast DNA restriction site variation. Can. J. Bot. 72:1692–1701.

Warwick SI, Black LD. 1997. Phylogenetic implications of chloroplast DNA restriction site variation in subtribes Raphaninae and Cakilinae (Brassicaceae, tribe Brassiceae). Can. J. Bot. 75:960–973.

Warwick SI, Sauder CA. 2005. Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trn L intron sequences. Can. J. Bot. 83:467–483.

Willis CG, Hall JC, Casas RR De, Wang TY, Donohue K. 2014. Diversification and the evolution of dispersal ability in the tribe Brassiceae (Brassicaceae). Ann. Bot. 114:1675–1686.

Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. PNAS 111:5283–5288.

Yang SS, Cheung F, Lee JJ, Ha M, Wei NE, Sze S, Stelly DM, Thaxton P, Triplett B, Town CD, et al. 2006. Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. Plant J. 47:761–775.

Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes : improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31:3081–3092.

Yoo M-J, Szadkowski E, Wendel JF. 2013. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity 110:171–180.

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, et al. 2015. Sequencing of allotetraploid cotton (Gossypium hirsutum L . acc . TM-1) provides a resource for fiber improvement. Nat. Biotechnol. 33:531–540.

ANNEXE

Analyses préliminaires

Hénocq L.*, Collard C.*, Vekemans X., Poux C.

*Les auteurs ont contribué de manière égale dans les travaux de détermination du système de reproduction des espèces.

I. Détermination du système de reproduction des espèces étudiées

La première année de thèse a permis de répondre à un certain nombre de questions concernant la biologie de la reproduction des espèces étudiées, qui ne sont pas des espèces modèles et pour lesquelles peu d'information existe dans la littérature (*Schouwia purpurea*, *Psychine stylosa*, *Carrichtera annua*, *Orychophragmus violaceus* et *Cakile maritima*). Des pollinisations contrôlées avec de l'auto-pollen ont été réalisées manuellement afin d'être comparées à des pollinisations contrôlées avec de l'allo-pollen. Afin d'optimiser notre protocole de pollinisations manuelles et de corroborer les résultats des croisements, nous avons également observé la présence ou l'absence de tubes polliniques après dépôt d'auto- ou d'allo-pollen dans différentes conditions (analyses non rapportées dans ce document). Enfin, le nombre de grains de pollen produits par fleur rapporté au nombre d'ovules contenus dans l'ovaire a été estimé, et ce afin de déterminer les valeurs du rapport pollen-ovule et tenter d'interpréter ces résultats en termes d'allogamie versus autogamie, au regard des données publiées par Preston (1986).

1.1 Espèces étudiées

Les espèces de Brassiceae sélectionnées pour notre étude sont les suivantes : *Schouwia purpurea* (clade Zilla), *Carrichtera annua* (clade Vella), *Psychine stylosa* (clade Savignya) et *Cakile maritima* (clade Cakile). L'espèce *Orychophragmus violaceus* (Brassicaceae), proche phylogénétiquement des Brassiceae, a également été sélectionnée comme groupe externe des Brassiceae. La détermination du système de reproduction a été conduite sur les cinq espèces mais des données exploitables ont été obtenues seulement pour les trois premières espèces. Par conséquent, seules les analyses correspondant à ces trois espèces sont rapportées ici.

Schouwia purpurea est une espèce annuelle. Il s'agit d'une plante désertique (Naggar and Soliman 1999) qui se retrouve en Afrique subtropicale (du sud de l'Algérie jusqu'en Somalie). La couleur de ses pétales peut être violette ou blanche (FIGURE 17A); les étamines sont libres ; le stigmate forme deux lobes ; le fruit est une silique ovale (FIGURE 17B), déhiscente contenant plusieurs graines de petites tailles (moins de 1.4 mm3) et de forme globulaire (Crespo et al. 2000). *Schouwia purpurea* est une espèce diploïde et qui possède n=18 chromosomes (Warwick and Al-Shehbaz 2006).

Des graines provenant du jardin botanique de Madrid ont été semées en serre en 2013, les individus résultant de ces semis ont produit les graines que nous avons utilisées pour nos manipulations. Sur la quarantaine de graines semées, 43 ont germé et fleuri, toutes appartenant à une seule et même population (TABLEAU 1).

(A) (B)

FIGURE 17 – (A) Fleurs et (B) fruit de *Schouwia purpurea*

Psychine stylosa est plante provenant de l'Afrique du Nord-Ouest. La couleur de ses pétales est soit violette soit blanche (la couleur des pétales pour les individus que nous avons étudiés est blanche) (FIGURE 18A) ; les étamines sont libres ; le stigmate est arrondi en son sommet ; le fruit est une silicule (FIGURE 18B) déhiscente contenant plusieurs graines de petites tailles (moins de 1.4 mm3) et de forme lenticulaire (Crespo et al. 2000). *Psychine stylosa* est une espèce diploïde, dont le nombre de chromosomes est de n=15 (Warwick and Al-Shehbaz 2006).

Les graines que nous avons semées proviennent du jardin botanique de Barcelone (population 3515) et de deux localités Marocaines (population 8184 et 8182). Le nombre de graines semées ainsi que le nombre d'individus ayant germé et fleuri pour chaque population sont représentés dans le TABLEAU 1.

(A) (B)

FIGURE 18 – (A) Fleurs et (B) fruit de *Psychine stylosa*

Carrichtera annua est une espèce herbacée annuelle adaptée au climat semi-aride de la région méditerranéenne et qui pousse en hiver (Gutterman 1990, 1993, Cooke et al. 2012). La température optimale de germination pour cette espèce se situe entre 15 et 20°C (Gutterman 1990). La couleur de ses pétales est jaune avec des veines de couleur pourpre (Figure 19A); les étamines sont libres ; le stigmate est arrondi au niveau de la tête ; le fruit est une silicule (Figure 19B) déhiscente contenant plusieurs graines de petites tailles (moins e 1.4 mm3) et de forme globulaire (Crespo et al. 2000). La dispersion des graines se fait environ à 25 cm de la plante mère par la pluie (Gutterman 1993), mais peuvent être dispersées plus loin par les fourmis (Gutterman and Shem-tov 1997), le vent et l'eau (Loria and Noy-Meir 1979-80). Le nombre de chromosomes est de n=8 (Warwick and Al-Shehbaz 2006).

Les graines qui ont été semées proviennent de Dijon (population 8168), de Jérusalem (population 3612), du Mainz (population 3166), de Munich (population 3037), d'Espagne (population 8191) et du jardin botanique de Leipzig en Allemagne (population 3548). Les graines utilisées pour représenter la population allemande sont plus précisément issues d'une première génération en serre au cours du printemps 2013. Le nombre de graines semées et le nombre d'individus ayant germé et fleuri pour chaque population sont représentés dans le TABLEAU 1.

(A) (B)

FIGURE 19 – (A) Fleurs et (B) fruit de *Carrichtera annua*

TABLEAU 1 – Nombre de graines semées et nombres d'individus ayant germé et fleuri chez les trois espèces étudiées.

Espèce	Population	Nombre de graines semées	Nombre d'individus ayant germé et fleuri
Schouwia purpurea	8087	~40	43
	3515	~40	37
Psychine stylosa	8184	27	3
	8182	20	6
	8168	13	0
	3612	19	0
Carrichtera annua	3166	10	0
	3037	19	0
	8191	24	0
	3548	~30	13

1.2 Pollinisations contrôlées

Pour *Schouwia purpurea* et pour *Carrichtera annua*, nous avons pollinisé les 10 premiers individus en fleurs. 5 individus ont été utilisés pour réaliser les autofécondations manuelles (AUF) et les 5 autres pour les allofécondations manuelles (ALF). Pour *Psychine stylosa*, les ALF ont été réalisées sur 5 individus de la population 3515 (les premiers individus en fleur). Les AUF ont quant à elles été réalisées sur 3 individus de la population 3515, 3 individus de la population 8184 et 3 individus de la population 8182. Afin de protéger les fleurs de toute contamination inopinée en pollen, un sachet micro-perforé a été déposé sur plantes. De plus, les individus soumis à l'étude ont été déplacés dans une autre cellule de la serre et éloignés d'environ 30 centimètres les uns des autres afin, là encore, de limiter les contaminations. La présence de thrips, insectes mangeurs de pollen, peut toutefois générer un certain nombre de contaminations, mais nous avons considéré les mouvements de cet insecte d'une plante à l'autre négligeable, compte tenu de l'éloignement des individus.

Au moins 10 réplicas par individu ont été réalisé à la fois pour les expériences d'allofécondation et pour celles d'autofécondation. Une étape de mise au point (pollinisations en serre et en milieu gélosé de stigmate puis coloration au bleu d'aniline, non rapportées ici) a été nécessaire pour connaître la période de réceptivité du stigmate de chaque espèce, de façon à réaliser les pollinisations au bon moment. Par conséquent le protocole n'est pas exactement le même pour toutes les espèces. Les AUF ont été réalisés sur des fleurs du jour (ouverture le

jour J). Les fleurs ont été émasculées (retrait des étamines) puis chaque pistil a été fécondé directement en utilisant le pollen provenant de la même fleur ou d'une autre fleur du même individu. Des étiquettes ont ensuite été placées autour de chaque fleur fécondée, indiquant la date, le numéro du croisement ainsi que le type de croisement (AUF). Les ALF ont été réalisées sur des fleurs sur le point de s'ouvrir (stade bouton floral mature, avant que les étamines ne déhiscent). Pour *Schouwia purpurea* et *Psychine stylosa*, les fleurs sélectionnées comme tel ont été émasculées. Les pollinisations ont été effectuées 48 à 72h après l'émasculation par de l'auto ou de l'allo-pollen. Cette méthode de pollinisation a été utilisée par Grundt et al. (2005) sur l'espèce *Draba palanderiana* (Brassicaceae), en pollinisant le pistil 1 à 2 jours après l'émasculation. Pour *Carrichtera annua*, les fleurs ont été émasculées puis fécondées directement par du pollen provenant d'un autre individu de la même population.

Des mesures ont été effectuées sur les fruits pour chaque espèce. La longueur et la largeur du pistil ont été mesurées le jour de la pollinisation et 7 jours après la pollinisation pour pouvoir (1) estimer la croissance du fruit et (2) vérifier si la valeur de ces deux variables 7 jours après la pollinisation est un bon indicateur de la réussite du croisement (*i.e* germination des grains de pollen, fécondation des ovules, fructification et maturation des graines). La distribution des valeurs de (longueur à J+7 – longueur à J_0) et de (largeur à J+7 – largeur à J_0) nous a permis d'établir un seuil de croissance en deçà duquel la pollinisation est considérée comme n'ayant pas fonctionné. Les valeurs au dessus de ce seuil sont donc associées à des succès de pollinisation.

Pour *Schouwia purpurea*, les graines étaient visibles par transparence et nous avons considéré qu'une pollinisation avait fonctionnée dès que l'on observait des graines. Dans certains cas, le fruit était très bien développé mais ne contenait pas de graines, nous avons quand même considéré que la pollinisation avait fonctionné. Pour *Carrichtera annua* et *Psychine stylosa*, l'observation des graines sans passer par une dissection du fruit n'était pas possible, nous avons utilisé la présence de fruits bien développés comme indicateur de succès de la pollinisation. Nous avons confronté ces données empiriques à notre estimateur de succès de pollinisation (voir ci-dessus) afin d'estimer un taux d'erreur et de déterminer si notre mesure de croissance de fruit est un bon estimateur du succès de la pollinisation (données non rapportées ici). Seuls les estimations empiriques (observation de fruit) seront considérées pour la suite des analyses. Une régression logistique a été réalisée afin de comparer les taux de fructification entre les ALF et les AUF, après avoir effectué un test de Shapiro pour vérifier la

normalité des données. En raison d'un faible nombre d'individus allo- et autofécondés issus d'une seule et même population, ce test n'a pas été réalisé chez *P. stylosa*.

En moyenne, chez *Schouwia purpurea* et *Carrichtera annua* 73.3% et 96.7% des autopollinisations ont fonctionné contre 77.6% et 30.9% des allopollinisations, respectivement. D'après les résultats de l'analyse, le taux de fructification des individus pollinisés avec de l'auto-pollen n'est pas significativement différent de celui des individus pollinés avec de l'allo-polen (P-value = 0.6782) chez *Schouwia purpurea. Schouwia purpurea* peut donc être considérée comme une espèce auto-compatible. Chez *Carrichtera annua*, cette différence est significative (P<2.2e-16), et ce sont les individus autofécondés qui présentent un taux de fructification plus élevé que celui des individus allofécondés. L'écart observé entre les deux types de pollinisation s'explique très probablement par le fait que les allofécondations ont été réalisées en toute fin de cycle de floraison chez cette espèce tandis que les autofécondations ont été réalisées en plein cœur du cycle de floraison. *Carrichtera annua* est donc très probablement auto-compatible, comme cela a déjà été rapporté (Boaz et al. 1990). Pour *Psychine stylosa*, aucune analyse statistique n'a pu être effectuée mais on constate que le taux de mise à fruit était assez faible chez la population 3515, que se soit pour les AUF ou pour les ALF (24% et 52%, respectivement) mais les résultats suggèrent malgré tout que l'espèce est auto-compatible. Un réajustement du protocole de pollinisation par la suite (pollinisation plus tardive) a confirmé que notre protocole de pollinisation initial n'était pas parfaitement adapté, expliquant certainement les faibles taux de mise à fruit chez cette espèce dans cette étude. Enfin, aucune pollinisation avec de l'auto-pollen n'a fonctionné chez la population 8184. Le fait qu'aucun croisement n'ait fonctionné peut être expliqué par : (1) l'existence d'un système d'auto-incompatibilité fonctionnel dans cette population, (2) un protocole de pollinisation non adapté (or il l'est dans un cas sur quatre pour les autopollinisations et dans un cas sur deux pour les allopollinisations chez la population 3515), ou (3) un problème d'adaptation aux conditions en serre pour cette population qui provient du Maroc (les plantes avaient en effet plus de mal à pousser et se sont vite desséchées). Cependant, nous avons pu observer deux ou trois fruits se former spontanément en dehors de nos réplicas sur les individus utilisés pour les tests d'autofécondation. Sachant que ces individus étaient sous sachets micro-perforés, (1) soit ces fruits se sont formés suite à une fécondation par de l'auto- pollen provenant d'autres fleurs de la plante, (2) soit il y a eu des contaminations via les thrips qui ont apporté passivement de l'allo-pollen. Pour vérifier ces hypothèses, il faudrait refaire les manipulations d'auto et d'allopollinisation en pollinisant les stigmates lorsqu'ils sont le plus réceptifs (pollinisation 48 heures après l'émasculation).

Aucune comparaison avec des allopollinisations n'étant possible pour cette population, nous ne pouvons pas déterminer si cette population est auto-incompatible ou pas. Malheureusement, nous ne disposions plus de graines provenant de cette population pour poursuivre davantage les investigations. La détermination du système de reproduction chez *O. violaceus* et *C. maritima* ont indiqué que la première est auto-compatible et la seconde auto-incompatible.

1.3 Le rapport P/O pour estimer le régime de reproduction des espèces

Le rapport du nombre de grains de pollen sur le nombre d'ovules par fleur (communément appelé rapport P/O) est largement utilisé dans la littérature pour estimer le régime de reproduction des espèces végétales (Preston 1986). Nous avons voulu savoir si ce rapport calculé pour les trois espèces de notre étude nous permet de tirer des conclusions sur leur régime de reproduction.

Pour chaque espèce, 13 individus ont été sélectionnés. Sur chaque individu, les 6 étamines et les gynécées ont été prélevés sur deux fleurs en ouverture, avant que les étamines ne soient déhiscentes, et placées dans des tubes Eppendorf. Les tubes contenant les étamines ont été placés à l'étuve pendant 48 heures à 55°C afin de rendre les étamines déhiscentes et ainsi permettre de mieux les vider et d'en extraire tout le pollen. Un compteur à particules a été utilisé afin de déterminer le nombre de grains de pollen total par fleur analysée (protocole non rapporté ici). Cette estimation du nombre total de grains de pollen ne permet pas de discriminer les grains de pollen viables des grains de pollen non viables. Cependant, une précédente estimation par montage des étamines sous lame et lamelle et utilisation de la coloration de Peterson (Peterson 2010) a montré que la proportion de grains de pollen non viables est faible chez les trois espèces, nous l'avons donc supposer négligeable et avons considéré que notre estimation obtenue par le compteur de pollen est une bonne estimation du nombre de grains de pollen viables.

Pour chaque fleur, l'ovaire a été disséqué sous une loupe binoculaire et les ovules ont été comptés. Le rapport pollen-ovule a été calculé en divisant le nombre de grains de pollen présents dans les six étamines par le nombre d'ovules déterminé. Ce rapport est ensuite recalculé en log (rapport moyen) pour chaque espèce afin de le confronter aux données publiées par Preston (1986) dans le but d'interpréter ces résultats en terme d'allogamie versus autogamie.

Le nombre de grains de pollen et d'ovules est très variable entre les espèces (TABLEAU 2). Au regard des données publiées par Preston (1986), on remarque que *Schouwia purpurea*

aurait un rapport correspondant aux espèces allogames. En revanche, pour *Psychine stylosa* et *Carrichtera annua*, les rapports calculés ne permettent pas de les classer dans un des régimes. Ces deux espèces ont un rapport P/O qui se situent sur la zone de chevauchement entre les valeurs des espèces autogames (le rapport P/O le plus grand est 3220, chez *Thysanocarpus curvipes* var. *elegans*) et allogames (le plus petit rapport P/O est 1100, chez *Barbarea orthoceras*) (Preston, 1986). Cependant, pour calculer le rapport P/O, Preston (1986) se base sur le pollen présent dans une seule étamine (il ne précise pas s'il s'agit d'une grande ou d'une petite) puis multiplie le nombre de grains de pollen présent dans cette étamine par le nombre d'étamines présentes sur la fleur (grandes et petites confondues). Il a donc fait l'hypothèse très forte qu'il n'existe pas de différence en termes de nombre de grains de pollen entre les grandes et les petites étamines (différence qui existe chez *Psychine stylosa*, données non rapportées ici), et qu'il n'existe pas de variation (1) entre les fleurs d'un même individu et (2) entre les étamines d'un même bouton. Contrairement à lui, nous avons pris toutes les étamines pour s'affranchir de ces biais, il est donc délicat de confronter nos résultats aux siens qui contiennent tous ces biais. Il faudrait ainsi refaire la manipulation, pour chaque espèce, en utilisant la même méthode que lui (une fois avec une grande et une autre fois avec une petite étamine) et voir si cela change substantiellement le rapport que l'on a trouvé et la position des espèces dans la distribution bimodale autogame/allogame.

Espèce	Nombre moyen d'ovules	Nombre moyen de grains de pollen	Rapport P/O moyen	Log(P/O)
Schouwia purpurea	12±1.83	173823±48963.07	14723.2	4.17
Psychine stylosa	22±5.08	35616±9572.49	1668.1	3.22
Carrichtera annua	4±0.9	6018±1139.86	1405.2	3.15

TABLEAU 2 – Nombre moyen d'ovules, de grains de pollen par fleurs, ainsi que le rapport moyen et sa valeur transformée en log pour chaque espèce.

1.4 Conclusion

Nous avons testé l'allofécondation et l'autofécondation pour chacune des espèces en réalisant des pollinisations manuelles en serre après émasculation des boutons floraux. Ces investigations ont fait l'objet d'un stage de master 1 (Chloé Collard). Un minimum de dix réplicas par individu a permis de comparer le taux de fructification entre les individus autofécondés et les individus allofécondés. Sur les trois espèces investiguées, toutes se sint révélées auto-compatible. Au cours d'analyses plus exploratoires n'ayant pas fait l'objet

d'analyses statistiques, *Cakile maritima* (clade Cakile) s'est avérée être auto-incompatible, ce qui est en accord avec les études précédemment faites sur cette espèce (Willis 2013). Nous avons toutefois noté chez cette espèce la présence d'individus totalement incompatibles et, en moindre fréquence, d'individus partiellement ou totalement auto-compatibles, ce qui avait aussi déjà été rapporté dans plusieurs études (Willis 2013 et références associées). *Orychophragmus violaceus*, quant à lui, s'est avéré être auto-compatible, ce qui est conforme au résultat d'une étude précédente (Zhang et al. 2007).

Les résultats obtenus nous permettent d'affirmer que les espèces *S. purpurea*, *P. stylosa, O. violaceus* et *C. annua* sont toutes auto-compatibles. Selon le système de reproduction des espèces (auto-compatible *vs.* auto-incompatible), notre attendu sur la diversité allélique au locus-S est très différent. Chez les espèces auto-compatibles, le locus-S n'a plus de rôle fonctionnel, nous nous attendons donc à trouver un ou seulement quelques allèles potentiellement non fonctionnels fixés dans la population voire aucun locus-S. En revanche, chez les individus auto-incompatibles de l'espèce *Cakile maritima*, nous devrions retrouver un nombre important d'allèles, tous fonctionnels.

II. Croisements pour la localisation génomique du locus S

Afin de tester la localisation génomique du locus-S, nous avons initialement pensé à faire des cartographies génétiques en étudiant la ségrégation des allèles du locus S avec ceux d'un couple de gènes flanquants le locus S de part et d'autre. Nous souhaitions étudier la ségrégation des allèles (1) au locus-S et (2) pour les gènes flanquants le locus-S en position « Arabidopsis » (ancestrale) et en position « Brassica ». Si le locus S est en déséquilibre de liaison avec des gènes flanquants représentant une des positions, nous pourrons en déduire sa localisation. Un minimum de deux couples par espèce a été sélectionné afin d'obtenir une vingtaine à une cinquantaine de graines par famille. En revanche, pour *Cakile maritima*, un seul couple a été choisi en raison de la difficulté des croisements chez cette espèce. De même, en raison de la difficulté des croisements et de l'efficacité très médiocre des pollinisations manuelles, nous n'avons pas effectué de tels croisements pour l'espèce *Carrichtera annua*. Les graines récupérées à l'issue de cette manipulation ont été semées au printemps 2016 et nous avons prélevé les feuilles de chacun des descendants (comme nous l'avons fait pour les parents) pour en extraire l'ADN et séquencer le locus S ainsi que les gènes flanquants afin de tester l'association entre les allèles de ces gènes et du locus-S. Les individus parentaux de chacun des couples et leurs descendants sur lesquels les analyses ont été conduites et ont fonctionné sont répertoriés dans le TABLEAU 3 ci-dessous.

Couple	Nombre de graines obtenues	Nombre de graines semées	Nombre de descendants obtenus
S. purpurea 1	~60	59	22
S. purpurea 2	86	86	26
P. stylosa	>200	215	51
O. violaceus	~40	43	42
C. maritima	41	-	-

TABLEAU 3 – Etat des croisements effectués pour déterminer la localisation génomique du locus-S pour chacune des espèces.

Les gènes flanquants choisis pour la position « Arabidopsis » sont le gène *B80/U-box* (pour lequel on dispose des amorces Forward et Reverse) et le gène *ARK3*. Pour la position « Brassica », après avoir cherché à amplifier différents gènes flanquants, notre choix s'est arrêté sur deux gènes flanquants (*Bra004178* en amont et *Bra004185* en aval du locus-S) que l'on retrouve *a priori* en une seule copie chez les espèces étudiées (selon nos données RNAseq). Toute la difficulté est que pour ces quatre gènes flanquants, nous avons besoin de définir des amorces généralistes (dans des régions bien conservées chez les différentes espèces étudiées) qui encadrent des régions polymorphes au sein d'une espèce. Nous maximisons nos chances d'obtenir du polymorphisme en définissant nos amorces de telle sorte qu'on amplifie un ou plusieurs introns, qui contiennent par définition davantage de polymorphisme que les exons.
Les résultats de séquençage obtenus n'ont pas été satisfaisants pour les raisons suivantes : (1) la portion amplifiée des gènes *Bra004185* et *B80/U-Box* ne montraient aucun polymorphisme chez les différentes espèces ; (2) il existe probablement un ou plusieurs paralogues du gène *Bra004178* puisque plusieurs bandes d'ADN étaient obtenues après migration des produits de PCR ; (3) les séquences du gène *ARK3* obtenues chez *Schouwia purpurea* et *Orychophragmus violaceus* présentaient de très nombreux sites polymorphes dont la plupart n'étaient pas partagés entre individus, ce qui ne nous permettait pas d'utiliser ces séquences pour l'analyse de ségrégation des allèles.

La difficulté supplémentaire de la méthode d'analyse de co-ségrégation des allèles, pourtant simple en apparence, est qu'elle nécessite d'obtenir un grand nombre de descendants, ce qui n'est pas toujours chose aisée. Notamment, pour l'espèce *Cakile maritima*, il est difficilement concevable d'obtenir plus de 50 graines par couple, étant donné le nombre réduit de fleurs produites sur un cycle de floraison et la présence d'une graine seulement par fruit (2 au maximum dans de très rares cas). C'est ainsi que nous avons été amenés à envisager le séquençage de la région du locus-S dans son ensemble pour ce qui est de la position ancestrale « Arabidopsis » (chez *Brassica rapa*, il y a environ 1200bp entre la *B80/U-box* et *ARK3*). Cette approche a pu au moins nous permettre de savoir si le locus S est présent ou absent en position « Arabidopsis », même si elle ne nous a pas permis pas de savoir s'il se trouve en position « Brassica », sans qu'il ne soit nécessaire de faire des croisements.
Le séquençage du locus S pour les analyses de co-ségragation des allèles avait été effectué dans un premier temps à partir des couples d'amorces utilisés dans Fléchon et al. (2012) (définis sur des allèles de classe I et de classe II des espèces *Cakile maritima* et *Sinapis arvensis*) mais ces amorces se sont avérées peu efficaces sur les espèces étudiées dans notre étude, malgré les différentes mises au point expérimentales. D'autres couples d'amorces ont donc été utilisé, les couples KD5/KD8 et KD4/KD7 développés par Edh et al. (2009) dans le domaine kinase du gène *SRK* et qui permettent l'amplification des allèles de classe I et de classe II, respectivement. Les premiers essais étaient plutôt encourageants, les amplifications fonctionnaient dans la plupart des cas et permettaient d'obtenir de belles séquences, bien que pour une partie seulement des espèces. En revanche, aucun polymorphisme n'a été trouvé dans les séquences des espèces *Carrichtera annua* et *Schouwia purpurea*, et seuls quelques sites polymorphes ont été trouvés chez *Cakile maritima* uniquement dans les séquences correspondant aux allèles de classe I.

Toutes ces difficultés de manipulations ne nous ont pas permis de poursuivre ces analyses. En revanche, nous disposons toujours des ADN de chacun des couples de parents et de leurs descendants (exceptés pour *Cakile maritima* dont les graines obtenues n'ont pas encore été semées), et il serait intéressant à l'avenir de réitérer l'analyse.

III. Association génotype et phénotype au locus-S

De façon à analyser l'association entre le génotype et le phénotype au locus S et ainsi nous assurer que deux individus qui présentent le même phénotype au locus S (information que l'on obtient *via* le séquençage du locus S) ne peuvent effectivement pas se reproduire entre

eux, nous voulions suivre le couple de l'espèce *C. maritima* (espèce auto-incompatible) sur deux générations. Autrement dit, nous souhaitions nous assurer que nous avions bien séquencé le locus S, et non l'un des multiples autres gènes appartenant à cette vaste famille de gènes. Cette manipulation n'était envisageable que chez les espèces auto-incompatibles (dans notre étude, seule *Cakile maritima* l'est). Nous n'étions pas en mesure de faire une telle vérification chez les quatre autres espèces investiguées. Etant donné que nous avions abandonné l'analyse de ségrégation des allèles et donc le séquençage du locus S chez les différentes espèces, pour les raisons exposées plus haut, nous n'avons pas poursuivi cette vérification, mais les graines sont toujours disponibles et une telle vérification pourra se faire à l'avenir dans le cas où nous souhaiterions retenter les amplifications du gène *SRK* à partir de nouvelles amorces.

IV. Séquençage du locus-S

Une à plusieurs populations de chaque espèce ont été maintenues en serre de février 2015 jusqu'au dernier trimestre de 2016. L'ADN de chaque individu a été extrait et l'exon 1 du gène SRK a été amplifié en vue du séquençage. Nous avons utilisé pour cela des amorces SLG-F/SLG-R barcodées (un code-barres différent pour chaque individu), ce qui permet de pooler les produits PCR de plusieurs individus avant le séquençage. Nous avons envisagé de faire du séquençage d'amplicons en utilisant la technologie PacBio. Nous espérions obtenir le génotype au locus S pour au minimum dix individus par espèce, soit au moins 50 génotypes. Avec ces séquences, nous voulions établir des phylogénies d'allèles *SRK* à l'instar des analyses présentées dans le chapitre III. Nous avons dû passer par une phase de mise au point des protocoles de biologie moléculaire afin d'optimiser l'amplification du gène *SRK* en terme de concentration (minimum 100 ng/µl) et de pureté (rapport A_{260}/A_{280}). En effet, durant de long mois, nous ne parvenions pas toujours à obtenir des concentrations suffisantes (estimées par le BioAnalyzer, kit ADN 1000) et la purification des produits PCR ne semblait pas fonctionner puisqu'elle n'éliminait pas les amorces utilisées pour l'amplification (kit PCR Clean-up de MACHEREY-NAGEL®). En théorie, la dilution de moitié d'un des produits de la purification devait permettre de ne pas retenir les fragments d'ADN de petites tailles, ce qui aurait dû permettre en toute logique d'éliminer les amorces récalcitrantes. Or ce ne fut pas le cas.

Dans un premier temps, et en guise de test, nous avons envoyé un échantillon de 10 individus pour séquençage, pour lesquels nous n'avions pas eu de problèmes particulier pour la purification des échantillons. Nous avons commandé au préalable 10 couples d'amorces

barcodées que nous avons testés afin de déterminer les conditions PCR optimales pour chacun d'eux. 6 individus de l'espèce *Psychine stylosa* et 4 individus de deux espèces de Biscutelles (pour répondre à des questions similaires chez le genre Biscutella) ont ainsi été séquencés en PacBio. Nous avons obtenu 150 000 subreads par échantillon, or les analyses informatiques nous permettent à ce jour d'exploiter uniquement 20 000 subreads par échantillon (soit uniquement 13% des données produites). Nous essayons d'optimiser l'exploitation des résultats, et cela s'avère d'autant plus nécessaire que nous ne retrouvons pas toujours le gène *SRK* dans nos échantillons. Nous pensons que les séquences du gène *SRK* se trouvent dans les données (les subreads) non exploitées. Le séquençage a révélé une contamination par *Arabidopsis thaliana* pour 5 échantillons sur 10. Nous ignorons à quelle étape des manipulations celle-ci s'est effectuée. Le gène *SRK* de l'espèce *A. thaliana* est préférentiellement séquencé au détriment de celui des espèces étudiées, ce qui pose un sérieux problème. Nous devons identifier l'origine de la contamination et prendre des précautions drastiques lors des prochaines manipulations. Les premiers résultats pour les 6 individus de *Psychine stylosa* (dont un est contaminé) laissent supposer l'existence d'un paralogue proche des allèles *SRK* de classe II, ce qui s'avère congruent avec les données obtenues dans le chapitre III.

A l'avenir, nous pourrons opter pour un protocole de purification alternatif au moyen de billes magnétiques, qui fournit une qualité supérieure d'amplicons, sans rétention de sels ni d'autres produits indésirables (AMPure Purification kit). Il en résulte un ADN d'une pureté particulièrement élevée.

RÉFÉRENCES

Cooke J, Ash J, Groves R. 2012. Population dynamics of the invasive, annual species, Carrichtera annua, in Australia. Rangel. J. 34:375–387.

Crespo M, Lledó DM, Fay MF, Chase MW. 2000. Subtribe Vellinae (Brassiceae, Brassicaceae): a combined analysis of ITS nrDNA sequences and morphological data. Ann. Bot. 86:53–62.

Edh K, Widen B, Ceplitis A. 2009. Molecular population genetics of the SRK and SCR Self-Incompatibility genes in the wild plant species Brassica cretica (Brassicaceae). Genetics 181:985–995.

Fléchon L, Poux C, Vekemans X. 2012. Identification de l'origine d'un goulot d'étranglement ancien au locus d'auto- incompatibilité dans la tribu des Brassiceae (Brassicaceae). :35.

Grundt HH, Elven R, Brochmann C. 2005. A rare case of self-incompatibility in arctic plants: Draba palanderiana (Brassicaceae). Flora: Morphology, Distribution, Functional Ecology of Plants. 200:321–325.

Gutterman Y, Shem-Tov S. 1997. The efficiency of the strategy of the mucilaginous seeds of some common annuals of the Negev adhering to the soil crust to delay collection by ants. Isr. J. Plant Sci. 45:317–327.

Gutterman Y. 1990. Do germination mechanisms differ in plants originating in deserts receiving winter or summer rain? Isr. J. Plant Sci. 39:355–372.

Gutterman Y. 1993. Seed Germination in Desert Plants. Springer-Verlag: New York.

Loria M, Noy-Meir I. 1979. Dynamics of some annual populations in a desert loess plain. Isr. J. Bot. 28:211–255.

Naggar SME, Soliman MA. (1999). Biosystematic studies on Schouwia DC. (Brassicaceae) in Egypt. Flora Mediterranea. 9:175–183.

Peterson R, Slovin JP, Chen, C. 2010. A simplified method for differential staining of aborted and non-aborted pollen grains. International Journal of Plant Biology. 1:66– 69.

Preston RE. 1986. Pollen-Ovule Ratios in the Cruciferae. American Journal of Botany. 73: 1732.

Warwick SI, Al-Shehbaz IA. 2006. Brassicaceae: Chromosome number index and database on CD-Rom. Plant Syst. Evol. 259:237–248.

Willis CG. 2013. The role of dispersal and adaptive divergence in the diversification and speciation of the tribe Brassiceae and Genus Cakile. :205.

Zhang L-J, Han Y, Dai S-I. 2007. The pollination biology of Orychophragmus violaceus. North. Hortic.

« Si la piste est trop facile et que tu crois tenir le jaguar, c'est qu'il est derrière toi, les yeux fixés sur ta nuque. »

Luis Sepúlveda ; Le Vieux qui lisait des romans d'amour (1992)

« Dans les larmes qui me nettoient les yeux purulents, je vois mille petits cristaux de toutes les couleurs et bêtement je pense : on dirait les vitraux d'une église. Dieu est avec toi aujourd'hui, Papi. C'est au milieu des éléments monstrueux de la nature, le vent, l'immensité de la mer, la profondeur des vagues, le toit vert imposant de la brousse, qu'on se sent infiniment petit relativement à tout ce qui vous entoure et que peut-être, sans le chercher, on rencontre Dieu, on le touche du doigt. »

Henri Charrière ; Papillon (1969)

RÉSUMÉ

La plupart des plantes à fleurs ont connu un ou plusieurs évènements de duplication de génome entier (polyploïdisation) au cours de leur histoire évolutive et particulièrement les membres de la famille des Brassicaceae. A titre d'exemple, l'ancêtre commun des membres de la tribu des Brassiceae aurait subi deux évènements successifs d'allopolyploïdie, générant une triplication de génome (WGT pour Whole Genome Triplication). Les évènements de polyploïdisation sont généralement suivis d'une diploïdisation impliquant des modifications génétiques et épi-génétiques, ainsi que des changements transcriptionnels aboutissant à la formation d'un génome diploïde. Par ailleurs, lors d'une duplication, la dynamique des éléments transposables est perturbée, ce qui peut conduire à une augmentation des évènements de translocation. Dans une lignée au sein de la tribu des Brassiceae, une perte drastique de diversité allélique, une réduction de la divergence moléculaire entre allèles ainsi qu'une translocation génomique ont été observées au locus responsable de l'auto-incompatibilité (locus S). Nous suspectons ces patrons d'être associés aux évènements d'allopolyploïdie. A partir d'approches phylogénomiques et d'analyse de la diversité du locus S dans la tribu des Brassiceae, cette thèse a pour but de déterminer si le goulot d'étranglement observé au locus S chez les Brassiceae est contemporain à l'événement de triplication de génome et s'il est, le cas échéant, associé à une translocation du locus S. Nos analyses suggèrent que toutes les espèces de Brassiceae partagent à la fois un même événement de triplication de génome mais aussi une perte de diversité phylogénétique au locus S qui semble précéder la diversification des Brassiceae. Néanmoins, nos données ne nous permettent pas de conclure fermement quant au lien entre translocation génomique du locus S et événement de triplication de génome à l'échelle des Brassiceae, bien qu'elles indiquent que la translocation observée chez Brassica est partagée par au moins plusieurs clades de Brassiceae.

ABSTRACT

Whole genome duplication events are common in flowering plants and especially within the Brassicaceae family. For example, the common ancestor of the Brassiceae tribe has experienced two successive events of allopolyploidy, generating a whole genome triplication (WGT). Polyploidy events are generally followed by a diploidization process involving genetic, epigenetic and structural changes leading to a diploid genome. Furthermore, after such an event, the dynamic of transposable elements is disturbed, which can lead to an increase in translocation events. In one lineage of the Brassiceae tribe, a drastic loss of allelic diversity, a decrease of molecular divergence among alleles and a genomic translocation have been observed at the self-incompatibility locus (S locus). We suspect that these patterns are associated with the allopolyploidy events. Using phylogenomic approaches combined with S-locus diversity analyses, we aim at determining whether the bottleneck observed at the S-locus in the Brassiceae tribe is contemporaneous with the inferred whole genome triplication and whether these events are also associated with the translocation of the S-locus. Our analyses suggest that all Brassiceae species share the same whole genome triplication event as well as a loss of phylogenetic diversity at the S-locus predating the divergence of Brassiceae lineages. Nevertheless, our data do not allow us to conclude about the association between the genomic translocation of the S locus and the whole genome triplication event, although they indicate that the translocation found in Brassica is shared by at least several Brassiceae clades.